MODELLING MUTUAL INTERACTION OF FINANCE AND HUMAN FACTOR VIA VARIOUS SORTS OF INDICES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF APPLIED MATHEMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BETÜL KALAYCI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
FINANCIAL MATHEMATICS

JUNE 2023

Approval of the thesis:

MODELLING MUTUAL INTERACTION OF FINANCE AND HUMAN FACTOR VIA VARIOUS SORTS OF INDICES

submitted by **BETÜL KALAYCI** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Financial Mathematics Department, Middle East Technical University** by,

Prof. Dr. A. Sevtap Selçuk Kestel
Dean, Graduate School of **Applied Mathematics** ______________

Prof. Dr. A. Sevtap Selçuk Kestel
Head of Department, **Financial Mathematics** ______________

Prof. Dr. Vilda Purutçuoğlu
Supervisor, **Statistics, METU** ______________

Prof. Dr. Gerhard-Wilhelm Weber
Co-supervisor, **Faculty of Engineering Management, Poznan University of Technology** ______________

Examining Committee Members:

Prof. Dr. A. Sevtap Selçuk Kestel
Financial Mathematics, METU ______________

Prof. Dr. Vilda Purutçuoğlu
Statistics, METU ______________

Prof. Dr. Ömür Uğur
Scientific Computing, METU ______________

Assoc. Prof. Dr. Ahmet Şensoy
Business Administration, Bilkent University ______________

Assoc. Prof. Dr. Özlem Defterli
Mathematics, Çankaya University ______________

Date: ______________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: BETÜL KALAYCI

Signature :

ABSTRACT

MODELLING MUTUAL INTERACTION OF FINANCE AND HUMAN FACTOR
VIA VARIOUS SORTS OF INDICES

KALAYCI, BETÜL
Ph.D., Department of Financial Mathematics
Supervisor : Prof. Dr. Vilda Purutçuoğlu
Co-Supervisor : Prof. Dr. Gerhard-Wilhelm Weber

June 2023, 154 pages

This thesis represents the mutual effects between some financial processes and sentiment indices by using various models from machine learning approaches and non-parametric models to parametric volatility models. In the analyses, we compare the gain in accuracy and computational time. We also evaluate the forecasting performance of sentiment index, consumer confidence index, consumer price index, unemployment rate and currency rate. Hereby, initially, we use sole multivariate adaptive regression splines (MARS), neural network (NN) and random forest (RF) models. Then, we apply two-stage hybrid models, namely, MARS-NN, MARS-RF, RF-MARS, RF-NN, NN-MARS, and NN-RF. Finally, we implement volatility models for sentiment index and consumer confidence index, and investigate plausible relationships with the selected macroeconomic data to improve the performance of forecast. In the interpretation of the findings, as the underlying datasets are prone to exhibit significant structural breaks, we apply the Markov switching model, define the location of breaks and lastly, we perform distinct time series volatility models. The results indicate better accuracy under Markov switching generalized autoregressive conditional heteroskedastic model, shortly, MSGARCH, among alternatives.

Keywords: Investor Sentiment, Consumer Confidence Index, Sentiment Index, Machine Learning, Volatility Model, Markov Switching Model.

ÖZ

FİNANS VE İNSAN FAKTÖRÜNÜN KARŞILIKLI ETKİLEŞİMİNİN ÇEŞİTLİ ENDEKS TÜRLERİYLE MODELLENMESİ

KALAYCI, BETÜL
Doktora, Finansal Matematik Bölümü
Tez Yöneticisi : Prof. Dr. Vilda Purutçuoğlu
Ortak Tez Yöneticisi : Prof. Dr. Gerhard-Wilhelm Weber

Haziran 2023, 154 sayfa

Bu tez, makine öğrenimi yaklaşımları ve parametrik olmayan modellerden parametrik volatilite modellerine kadar çeşitli modeller kullanılarak bazı finansal süreçler ve duyarlılık endeksleri arasındaki karşılıklı etkileri temsil etmektedir. Analizlerde, doğruluk ve hesaplama zamanındaki kazancı karşılaştırıyoruz. Ayrıca duyarlılık endeksi, tüketici güven endeksi, tüketici fiyat endeksi, işsizlik oranı ve döviz kurunun tahmin performansını da değerlendiriyoruz. Buna dayanarak, başlangıçta, tek çok değişkenli uyarlanabilir regresyon splinleri (MARS), sinir ağı (NN) ve rastgele orman (RF) modellerini kullanıyoruz. Ardından MARS-NN, MARS-RF, RF-MARS, RF-NN, NN-MARS ve NN-RF olmak üzere iki aşamalı hibrit modeller uyguluyoruz. Son olarak, duyarlılık endeksi ve tüketici güven endeksi için oynaklık modelleri uyguluyoruz ve tahmin performansını iyileştirmek için seçilen makroekonomik verilerle makul ilişkileri araştırıyoruz. Bulguların yorumlanmasında, altta yatan veri kümeleri önemli yapısal kırılmalar sergilemeye eğilimli olduğundan, Markov anahtarlama modelini uyguluyoruz, kırılmaların yerini tanımlıyoruz ve son olarak farklı zaman serisi oynaklık modelleri gerçekleştiriyoruz. Sonuçlar, alternatifler arasında Markov anahtarlama genelleştirilmiş otoregresif koşullu heterosketastik model, kısaca MSGARCH, altında daha iyi doğruluğu göstermektedir.

Anahtar Kelimeler: Yatırımcı Duyarlılığı, Tüketici Güven Endeksi, Duyarlılık Endeksi, Makine Öğrenmesi, Volatilite Model, Markov Geçiş Modeli

To my father

ACKNOWLEDGMENTS

I would like to sincerely thank my thesis supervisor, Prof. Dr. Vilda Purutçuoğlu. It is a great honor for me to be her PhD. student. Prof. Dr. Purutçuoğlu guided, encouraged and supported me during the development and preparation of this thesis from beginning to end. I am very grateful for all of her contributions and motivations during the whole period of my PhD. studies. She has become my role model from our first day of studying to the end of my life.

I would like to express my very great appreciation to my thesis co-supervisor Prof. Dr. Gerhard Wilhelm Weber for his endless help and support. His willingness to give his time and to share his experiences has brightened my path.

I would like to thank to my dear jury members; Prof. Dr. A. Sevtap Kestel, Assoc. Prof. Dr. Özlem Defterli, Prof. Dr. Ömür Uğur, and Assoc. Prof. Dr. Ahmet Şensoy for their valuable comments, suggestions, support and motivation.

I owe my deepest thanks to Dr. Ayşe Özmen for her share of experiences and expertise.

I deeply thank all members of the Institute of Applied Mathematics (IAM) at Middle East Technical University for their friendship, understanding and help during my studies.

I am very thankful to my dear friends Nurdan Kalaycı, Murat Turan, Aytaç Kara, Özenç Murat Mert and Pelin Çiloğlu for their warmly support, friendship, endless help and motivation during this process.

Finally, and most significantly, I would like to thank to my family for their patience and love. I owe my grateful thanks to my father Sami Kalaycı who has enlightened my vision, guided me with his invaluable experiences and thoughts about every area of life since I was a little girl. I devote this thesis to him.

I would like to express my gratitude to my mother Aynur who has always supported and believed in me at my both the best and the worst days, from the first day of my life to now. To thank her will never be enough to express my gratitude to her.

I would like to thank my brother Haluk and my aunt Esra for their motivation, help and suggestions. Lastly, I thank to my dear and lovely nephew Kerem Tuna who makes me smile with his presence.

TABLE OF CONTENTS

CHAPTERS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
ANFIS	Adaptive Neuro-Fuzzy Inference Systems
ANN	Artificial Neural Network
AR	Autoregressive
ARCH	Autoregressive Conditional Heteroskedasticity
ARIMA	Autoregressive Integrated Moving Average
BF	Basis Function
BIC	Bayesian Information Criterion
BIST	Borsa Istanbul
BPN	Backpropagation Neural Network
CART	Classification and Regression Trees
CCI	Consumer Confidence Index
CMARS	Conic Multivariate Adaptive Regression Splines
CPI	Consumer Price Index
CPS	Child Protective Services
DG ECFIN	Directorate-General for Economic and Financial Affairs
EGARCH	Exponential Generalized Autoregressive Conditional Heteroskedasticity
EVDS	Elektronik Veri Dağıtım Sistemi
ES	Expected Shortfall
FARIMA	Fractional Autoregressive Integrated Moving Average
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GCV	Generalized Cross Validation
HMM	Hidden Markov Model
KYC	Know You Customer
MARS	Multivariate Adaptive Regression Splines
MCCI	University of Michigan's Consumer Confidence Index
MGHMM	Mixture Gaussian Hidden Markov Model
MIDAS	Mix Data Sampling

MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
MOPSO	Multi-Objective Particle Swarm Optimization
MS	Markov Switching
MSE	Mean Square Error
NBER	National Bureau of Economic Research
NN	Neural Network
NSGA	Non-dominated Sorting Genetic Algorithm
NYSE	New York Stock Exchange
OOB	Out of Bag
RCMARS	Robust Conic Multivariate Adaptive Regression Splines
RF	Random Forest
RFA	Random Forest Algorithm
RMSE	Root Mean Square Error
RSM	Regime Switching Model
RSS	Residual Sum of Square
SDE	Stochastic Differential Equation
SVR	Support Vector Regression
TCMB	Türkiye Cumhuriyet Merkez Bankası
TURKSTAT	Turkish Statistical Institute
TÜİK	Türkiye İstatistik Kurumu
UN	Unemployment
VaR	Value at Risk
YSMI	Yale School of Management Index

CHAPTER 1

INTRODUCTION

1.1 Motivation and Literature Review

The neurological activities of humans influence financial decisions. What goes on in investors' minds while they make financial decisions like purchasing and selling? In a dangerous or risk-averse setting, which hormones are released? These are some examples of interesting questions in human side. Hence, in an area known as "the neurological underpinnings of the decision-making based on one's emotional condition", studies on these subjects are described. In order to create the multidisciplinary field of neurofinance, which includes every occurrence and interactions and explains emotions and ideas that are beyond people's grasp, neuroscience seeks to connect with finance [88, 93]. On the other hand, apart from investors' hormonal or brain acitivities, emotions, beliefs, and ideas all play a role in how people behave, and economists investigate these influences in many empirical findings. This is what the emerging field of *Behavioral Finance* entails [28, 39, 47]. *Sentiment* is a general term for all of these emotional states. The influences of investor and consumer sentiment on the stock market and the overall economy are examined by behavioral finance. Dealing with worldviews as assumptions about the circumstances of variables upon present data or present knowledge reflects this [39, 47].
As Keynes associated that the responsiveness to changes in the economy (sentiment) with "expectations for a long-term situation" and "situation of reliance", the responsiveness of producers and consumers to economic changes are crucial in explaining fluctuations in the economy. The projections of economic agents play a substantial role in determining the trajectory of macroeconomic and financial indicators. Ob-

serving these indicators beforehand can give advance insight into their anticipated path [14].

The behavior of the consumers is of utmost importance in macroeconomic modeling, as it is essential to examine how reliance of consumers impacts their economical behavior. Numerous investigators and analysts have attempted to research the correlation between macroeconomic indicators and Consumer Confidence Index (CCI). CCI primarily based on consumers' answers as “positive”, “negative”, or “neutral” to specific questions regarding present and outlook economic circumstances, both on an individual and national level . This index is a helpful and valuable indicator for investors, policymakers and entrepreneurs within a country [45].

In our study, we use the term “human factor” for both investors' and consumers' sentiment. We aim to see the forecasting results of these indexes.

Since these kinds of human factor and financial data might subject to high fluctuation and high correlation, the traditional statistical methods usually are not able to cope with these complex problems [99]. In the literature of these fields, some kinds of data mining techniques have been studied. Bahrammirzaee (2010) studied a comparative research review of three well-known artificial intelligence methods in finance sector; which are, expert systems, hybrid intelligence systems and artificial neural networks (ANN). In his study, financial market has been classified and explained on three aspects, namely, assessing creditworthiness, managing portfolios, and making financial projections and plans. As a result, artificial intelligent techniques have obtained more accurate results than the conventional statistical approaches to financial issues [10]. Lu et al. (2010) compared the forecasting performance of Multivariate Adaptive Regression Splines (MARS), support vector regression (SVR), multiple linear regression (MLR) and backpropagation neural network (BPN) models in Shanghai B-Share stock index in order to predict stock index prices.According to experimental findings, MARS performs better than SVR, MLR and BPN in terms of both estimating error and accuracy [64]. In order to increase prediction accuracy, Kao et al. (2012) introduced a forecasting model for novel stock price called Wavelet-MARS-SVR that combines multivariate adaptive regression splines (MARS), support vector regression (SVR) and wavelet transform. By contrasting the prediction outcomes produced by Wavelet-MARS-SVR with those produced by the other five competing approaches listed as (Wavelet-MARS, Wavelet-SVR, single SVR, single Adaptive-Network-based Fuzzy Inference Systems (ANFIS) and single Autoregressive Integrated Moving Average (ARIMA)), the effectiveness of the suggested strategy is utilized. The research's conclusions demonstrate that the suggested strategy outperforms other competing models [49]. Jadhav et al. (2016) aimed to study on data mining methods used in financial institutions between 2010 and 2015. When analogized to time series prediction, money laundering, and loan prediction, reviews show that academics are particularly interested in stock prediction and credit rating. Recently, it has been described by the MARS approach as an alternative of the underlying method [90]. Syah et al. (2020) investigated the Know Your consumer (KYC) System and MARS optimization model for consumer behavior. One of the technologies that has lately been used to regulate consumer behavior and verify correct data for security and user pleasure is KYC. Their study aimed to obtain optimal models by using MARS.The authors estimate growth based on the findings of their study and the use of the model for optimization for minimizing data from the KYC System, therefore the significance of this research is to foresee and establish sustainable business decisions. The MARS technique works really well for finding the ideal model [90]. Kalaycı et al. (2020) studied the interactions between various financial processes and investor sentiment with building a linked system of non-autonomous SDEs that evolved over time. It is challenging to analyze and solve these equations. So, by using discretization and MARS model, we streamlined these equations' expression [48].

As it is seen in literature, since these kinds of human factor and financial data expose to high variation, the general and conventional statistical methods usually are not able to handle these problems [99]. At this point, these traditional models have bring about a growing interest in machine learning techniques [10, 48]. Nonlinear mapping techniques have been worked out more than linear techniques because of the dynamics, uncertainty and variety of data. Many investors' decisions and opinions in the fields of financial sector are studied by machine learning models. This provides help in the financial industry to comprehend the data and to achieve a edge from the data [46]. In addition, recently, Hybrid models which combine the several machine learning techniques tend to improve gradually [30, 46]. The association derived by machine learning has proved succeeding in hybrid methods. Furthermore, it has been also shown in the study that hybrid methods have more precise forecast, nearly proceeded by Neural Network model [46]. Li (2010) sought to build a financial distress warning system for banking operations in emerging nations between 1998 and 2006 using a novel two-stage hybrid model of logistic regression-ANN.

The advantages of logistic regression and ANN were combined in this suggested two-stage hybrid model, which minimized computing complexity. The proposed strategy outperformed existing models after implementing certain novel treatments. According to the findings, there is a strong correlation between the determinants of liquidity, capital, and asset quality and banks' financial challenges in emerging markets. In terms of predicting financially troubled banks, the suggested a two-phase mixed design, which outperformed more traditional ones in terms of prediction power, gave best fit measures on the grounds that the $RMSE$ and R^2 [62].

Ravi et al. (2017) proposed two 3-stage hybrid prediction models in which Multi-Layer Perceptron (MLP) (Stage-1) and Multi-Objective Particle Swarm Optimization (MOPSO) (Stage-2) and elitist Non-dominated Sorting Genetic Algorithm (NSGA-II) (Stage-3) are used concurrently. Stage-1 uses Chaos theory to build phase space. Stage 3 in each of these hybrid models advances the conclusion reached in stage 2's forecast. On financial datasets that include data on the US Dollar (USD) exchange rates vs British Pound (GBP), Euro (EUR), the Japanese Yen (JPY) and Gold price with regard to USD, the suggested models' effects are evaluated [79]. As seen in literature, even though combining several machine learning techniques can lead to observe more accurate results, the process should always take into consideration the significance of qualification of data, as the ambiguity of data need the machine learning technology's resilience all the while. [46].

In this study, we concentrate on behavior of financial and economical problems which is based on the investors' and consumers' behavior introduced as Sentiment. Furthermore, we compare the forecasting performance of sentiment indexes by using single MARS, RF (Random Forest), NN (Neural Network) models, and two-stage MARS-NN, MARS-RF, RF-MARS, RF-NN, NN-MARS, and NN-RF hybrid models. Here, MARS denotes Multivariate Adaptive Regression Splines, NN implies the Neural Network model and RF indicates the Random Forest approach. In this thesis, we discuss both the financial and psychological areas in a way that is collaborative. The decisions of investors and consumers, as well as some features of their investments and consumptions, are of interest to certain scholars and seasoned traders. Due to the high possibility of fluctuation, by using MARS algorithm, our goal is to offer a more accurate and consistent approximation for the data while minimizing this dispersion [48]. Initially, we consider to extend this model by using distinct data mining techniques as presented in Section 2.

In the following parts; firstly, the machine learning methods which we use in this study are introduced. Secondly, we give information about investors' behavior and consumers' confidence and the term "Sentiment". Because, in the literature, there are a lot of studies for machine learning techniques applied to finance, however, we add "human factor" part into these machine learning approaches. In this part, we briefly mention those and afterwards, we see the application part and discussion of the results of the application both for single and two-stage machine learning techniques.

The field of behavioral finance studies the act of both rational and noisy traders. According to the general studies in the literature, influence of noise traders has a greater impact on the market. The market's presence of aggressive trader leads to changes in both market returns and volatility, as a result of their cognitive mistakes and emotional enthusiasm [74]. On the other hand, in the literature that the demands of consumers for consumption are substantially influenced by their reactions to economic factors. Moreover, it was mentioned that consumers base their consumption decisions on their expectations about their future financial situation, as well as their needs and desires [14]. These expectations and reactions also bring about volatility.

Higher levels of volatility in markets and in the economy can cause alterations in the way how risks are spread among financial assets and macroeconomic variables [81]. These alterations have an impact on the sentiment of investors. Additionally, not only investors but also consumers are extremely affected by these changes. Therefore, we can make an evaluation that decisions regarding finance are frequently made based on a interchange between return and risk. It is essential to analyze risk using econometric techniques in various financial aspects such as risk management, asset pricing, portfolio optimization, and option pricing. To achieve this, different time series analyses such as the Autoregressive Conditional Heteroskedasticity (ARCH) and Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) models have been employed, and their use in economics and finance has been particularly effective [81].

From this point of view, in our study, we aim to use volatility models to our sentiment and financial datasets as an alternative of MARS and other data mining techniques under certain assumption. We apply volatility models to our sentiment indexes for both univariate and multivariate case as well.

There are numerous causes for significant changes in financial patterns, such as economic downturns, business slowdowns, insolvencies, market alarms, and variations in governmental regulations or investor assumptions that arise from changes in leadership [15]. All these lead to regime switching in a model. Each regime has its own unique pattern of instability. The study of the tendency of prices to fluctuate over time, known as volatility analysis, is extremely important in many financial contexts. The GARCH model has been extensively used by researchers and practitioners in various fields [3]. Standard GARCH models may produce inaccurate results if the series has constitutional defect [22]. Therefore, at this stage, it would be better to consider a more suitable model. This is because each regime has a distinct level of volatility, resulting in different GARCH behaviors for each state of the regime chain. To avoid any bias, combining GARCH models with a Markov switching chain broadens the dynamic structure of the model and makes it possible to generate more accurate predictions of volatility [3, 22]. When facing such scenarios, a technique called *Markov-Switching GARCH* (MS-GARCH) models can be used. This method allows the model's parameters to change over time based on a discrete hidden variable [22]. The MS-GARCH model is used to create an approach that takes parameter changes into account, known as regime switching. The GARCH model is expanded upon by this one and enables different degrees of persistence in the conditional variance for each regime [3, 13, 15, 22].

The general aim of this thesis is to analyze both the machine learning models and as an alternative to them volatility models for the effect of sentiment and confidence level of investors and consumers to financial and macroeconomic variables. While we are analyzing, we prefer to classify these investors and consumers according to the optimism or pessimism level so that we can clarify which emotional state people are tend to be more or is there any probability for them to switch from one state to another state, if there are, what is the probability of it. We examine all these by employing Hidden Markov Model and Markov-Switching Model. After observing the probability of states, we also search for the volatility of each state by employing GARCH model (MS-GARCH) to see at which state people's emotional situation is fluctuant.

1.2 Scope of the Thesis

The main content and aim of this thesis are as follows.

- We discuss the interplay between the "human factor" and financial aspects.
- By using MARS model, we have purposed to decrease random variation, which comes from finance, as well as probable and to ensure a more seamless as well as more consistent rough estimate of the data.
- We consider to extend this model by using distinct data mining techniques. By adapting different clustering approaches into our financial and sentiment datasets in advance of MARS such as random forest (RF), neural network (NN) etc., we aim to investigate the subgroups in the main list of variables.
- We employ the underlying three major approaches as one-stage and two-stage modelling to both Sentiment Index and Consumer Confidence Index (CCI). For that reason, before we see the application results, we briefly introduce what these indexes are and what they define. All of these methods, definitions of indexes and results of applications are mentioned in Chapter 3.
- Afterwards, we introduce volatility models (ARCH, GARCH and EGARCH) and implement them to each of the sentiment indexes. At first, we apply these volatility models to the Sentiment Index and obtain statistical results for it, following, we carry out for each of the macroeconomic variables (Consumer

Confidence Index, Consumer Price Index, Unemployment rate and USD/TRY currency index) first for the univariate case, then, for the multivariate case.

- Then, present Markov Regime Switching model and combine this model with GARCH model. As a result, we obtain three regimes and we apply GARCH model to each of these regimes in order to observe different volatility structures of all of these regimes.

Accordingly, the next section will provide a thorough explanation of the mathematical concepts in machine learning techniques. Subsequently, we will examine how these techniques are implemented and which outcome they produce in Chapter 3. We represent the volatility models and their implementations in Chapter 4. Chapter 5 is dedicated to the Markov Switching model and its application together with volatility model. Finally, we conclude our findings and discuss the future work on Chapter 6. In addition to these chapters, computations, statistical results and equations together with graphs and tables of all the models are comprehensively indicated in Appendix parts A,B,C and D.

CHAPTER 2

MACHINE LEARNING TECHNIQUES

2.1 Machine Learning Techniques

In this thesis, because of their powerful properties which can cope with the extreme fluctuations, provide more accurate results, take into account the efficieny and robustness, we introduce and apply for three machine learning techniques, which are MARS, RF and NN.

2.1.1 Multivariate Adaptive Regression Splines (MARS)

MARS is comparable to stepwise linear regression [37]. The input variables are not handled independently by MARS; rather, it considers their interactions, in contrast to additive models. With more details, MARS is a method for fitting relationships between a limited number of regressive variables and the outcome variable by employing smoothing splines. Piecewise, a very seamless path or layer that may capture “alterations” in the relationship between these variables is created. These transitions take place at designated “knots” and provide a seamless changeover between “regimes”.

In addition to looking into all knot situations, the model also looks into all positions where variables can interact with one another. This is accomplished by integrating variables known as *basis functions*, which are then referred to as splines. After MARS determines the ideal number of basis functions and knot placements, the fitted value is estimated using the selected basis functions in a least-squares regression at the end [86].

As a result, the multivariate additive model that is produced is decided using a two-step procedure known as the *forward stage* and the *backward stage*. MARS develops a potentially huge model that typically overfits the dataset by quickly identifying the basis functions (BFs) that are associated to the model at the *forward stage*. the model's greatest amount of basis functions, which is a fixed number decided by consumers, is reached, the procedure is repeated. In fact, BFs in this model both contribute the most and the least to the total performance. As a result, the forward stage of the model is more intricate and filled with erroneous words. The overfit model is cut at the *backward stage* to reduce the model's complexity. Nevertheless, the model incorporates the data fit and improves the general efficiency. Each iteration at this backward stage involves the removal from the model of the BFs that contribute the least to a rise in the *residual sum of squares (RSS)*. Eventually, an idealized estimation model is constructed [37, 58, 69, 68]. MARS employs enlargements the following individually linear a single-dimensional basis functions, which have the form: $(x-t)_+$ and $(x-t)_-$:

$$(x-t)_+ = \begin{cases} x-t, & \text{if} \quad x>t, \\ 0, & \text{otherwise}, \end{cases} \quad \text{and} \quad (x-t)_- = \begin{cases} t-x, & \text{if} \quad x<t, \\ 0, & \text{otherwise}. \end{cases}$$

Each of these maps is a *truncated linear function* with a dataset-derived a single variable knot at value t, determined employing the dataset. As an illustration, Figure 2.1 displays the BF combinations for $t = 0.5$. Each of these maps is a compressed linear function with a dataset-derived a single variable knot at value t.

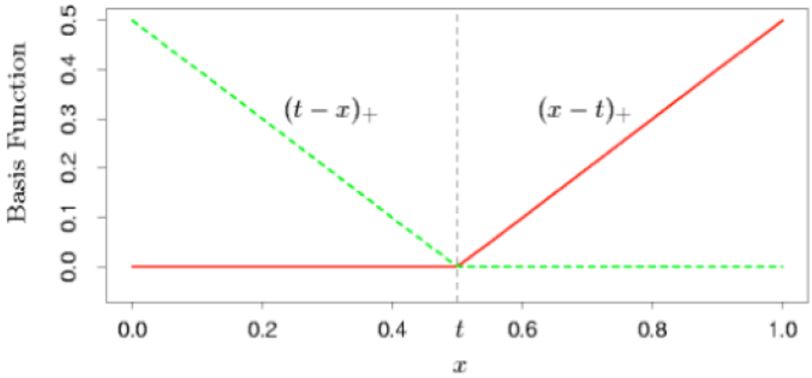


Figure 2.1: The MARS' BFs employed for $t = 0.5$ [68].

A *expressed combination* is the name given to those two functions. The objective is to display an expressed combination with loops at every single one of the input's measured values, x_{ij}, for each input, X_j. Consequently, the sum of the basis functions is

$$C = \{(x_j - t)_+, (x_j - t)_- : \quad t \in \{x_{1j}, x_{2j}, ..., x_{Nj}\}, j = 1, 2, ..., p\}, \tag{2.1}$$

If all of the input values are distinct, there are $2Np$ one-dimensional basis functions where N is the quantity of measurements and p is the size of the input space. Whereas only one X_j is required for each of the aforementioned fundamental functions, like in the case of $h(\boldsymbol{x}) = (x_j - t)_+$, they are all considered functions across the full input space $\mathbb{R}^p$.

As was already said, the model-building technique initially employs "forward stepwise linear regression", however, it is allowed to substitute functions from the set C and their products for the original inputs. Consequently, The initial form of the model is as stated below:

$$f(\boldsymbol{x}) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(\boldsymbol{x}). \tag{2.2}$$

This occurred while in the MARS advance phase, where M denotes the group of basis functions utilized in the present framework and $h_m(\boldsymbol{x})$ denotes a product of two or more multidimensional basis functions from C or a mixture of at least two of these functions. Additionally, $\boldsymbol{x} = (x_1, x_2, ..., x_p)^T$ denotes the undetermined components at either the mth basis function or the constant 1 $(m = 0)$. The mth basis function can be illustrated as follows:

$$h_m(\boldsymbol{x}) = \prod_{k=1}^{K_m} \left(s_{km} \cdot (x_{v(k,m)} - t_{km})\right)_+, \tag{2.3}$$

where K_m is the total number of multiplied trimmed linear functions in the mth basis function and $x_{v(k,m)}$ is the input variable that corresponds to the kth trimmed linear function in the mth basis function. Furthermore, t_{km} represents the knot value appropriate for the variables $x_{v(k,m)}$ and $s_{km} = \pm 1$. Herein, an evaluation of the likely basis functions is done using the *lack-of-fit approach.*

MARS moving forward in steps approach starts by estimating β_0 using the constant function $h_m(\boldsymbol{x}) = 1$, and the remaining elements in the set C are all candidate functions. The basis functions $h_m(\boldsymbol{x})$ come in a variety of forms, some of which are described below.

- 1,
- x_k,
- $(x_k - t_i)$,
- $x_k x_l$,
- $(x_k - t_i)_+ x_l$,
- $(x_k - t_i)_+ (x_l - t_j)_+$.

For any of these fundamental functions, the MARS method cannot accept identical input variables. Because of this, the basic functions mentioned above employ two different sets of parameters for input, x_k and x_l, together with the respective knots, t_i and t_j. Each time, a new basis function pair is considered by combining a function h_m from the model set M with one of the reflected pairings in C. The addition of the form term to the model set results in the greatest magnitude of reduction in training error [37, 47, 69, 68]:

$$\hat{\beta}_{M+1} h_l(x) \cdot (x_j - t)_+ + \hat{\beta}_{M+2} h_l(x) \cdot (t - x_j)_+, \qquad h_l \in \mathcal{M}. \tag{2.4}$$

The coefficients $\hat{\beta}_{M+1}$ and $\hat{\beta}_{M+2}$ in the equation (2.4) were estimated using the least square method (LS), along with all the other $M + 1$ parameters in the method.

When the model set $\mathbb{M}$ has the greatest amount of preset terms, the procedure is complete and the final products are produced by the model. For instance, the subsequent fundamental functions could be contenders for inclusion in the model [37, 68]:

- 1,
- x_k,
- $(x_k - t_i)$, if the model presently includes x_k,
- $x_k x_l$, if the model presently includes x_k and x_l,

- $(x_k - t_i)_+ x_l$, if the model presently includes $x_k x_l$ and $(x_k - t_i)$,
- $(x_k - t_i)_+ (x_l - t_j)_+$, if the model presently includes $(x_k - t_i)_+ x_l$ and $(x_k - t_i)_+ x_k$.

At the end of each phase of this advance phase, we have a sizable model that frequently matches the data. Consequently, a reverse removal is also used. Using a backward method, we exclude from the model the element that improves the residual squared error at every phase by the least amount. When the final model achieves the ideal number of effective terms, the process ends. The result is an estimated best model $\hat{f}_\lambda$ of any size, or λ (i.e., number of terms). In the MARS model, to determine the ideal value of λ, the *Generalized Cross-Validation (GCV)* is employed. This need is designated as

$$GCV(\lambda) = \frac{\sum_{i=1}^{N} (y_i - \hat{f}_\lambda(x_i))^2}{(1 - M(\lambda)/N)^2}. \tag{2.5}$$

In Equation (2.5), the number $M(\lambda)$ represents the model's influential parameters and N represents the sample's number of observations.
Regression approaches and other traditional statistical methods are sometimes effective at analyzing interaction terms. These techniques must test numerous combinations of the dataset's variables. As a result, they are unable to offer a computationally effective solution. In contrast, MARS automatically looks for suitable interactions between independent variables, which is often preferable when there are a lot of interactive factors. Consequently, it can spot interactions as well as a very limited number of sophisticated starting variable changes known as regressor variables. Along with these benefits, it also offers a chance to identify nonlinearities that may be present in the correlation between the two variables that are dependent and independent. Additionally, it creates graphs that make connections easier to see and understand [37, 68].

2.1.2 Random Forest Algorithm

The random forest algorithm (RF) is one of the supervised learning methods when it comes to data mining techniques. In this algorithm, the classification and the regression are considered. The objective of the categorization is to place each observation in the appropriate subgroups whose components are already recognized [20, 87]. This calculation has several steps, namely, the boosting, the calculation for the overfitting problem and the bagging. Below, we describe each step with more details.

(i)Boosting

The early predictors have a major role in the weighting process of the subsequent trees. In this process, the classification problems can occur. To remove these problems, a method called *Boosting* is used. In this computation, initially, among all variables, the weighted vote is chosen for the prediction.

The remaining weak classifiers are then combined into a strong ensemble to form a strong classifier committee. A weak classifier means that it is not ensuring a reduced mistake rate compared to guessing at random ensures. Accordingly, each time these weak classifiers are applied, the data are changed successively. Then, it finds an inadequate predictor link which generates the boosting. *Adaboost* is another phrase that exists which is an adaptive version of boosting. This boosting type creates a committee of trees by recalculating the weights of the previous ensembles without using any random elements [87].

(ii) Overfitting Problem

The branching process in conventional classification trees continues by dividing every single node according to the optimal combinations of each of the parameters. Controlling the changes in the *generalization error, strength* and *correlation* allows the branching to continue in the process of choosing the best groups. However, sometimes, it also examines measurement noises, which might cause an overfitting issue while performing calculations [87].

(iii) Bagging

Selecting the forest's individual trees separately from the previously chosen trees is one way to solve the overfitting problem. The bagging is one of these algorithms which provide such a solution. On this wise, it builds every tree individually by choosing a sample to serve as a bootstrap from each dataset of the datasets and then, for the prediction, it selects the tree with the most votes. By this way, the variance can be impressively decreased by bagging. However, because it causes bias during the variance reduction, it is insufficient to completely eliminate the overfitting issue. For that reason, the *adaptive bagging* is recommended as the bagging method that productively decreases both variation and skew. Furthermore, it improves estimates of the primary metrics by contrasting the *correlation, generalization error* and *strength* of mixed groups of trees. Thus, it avoids having the bagging error affect estimations of the measures [87].

Mathematical Details of the Random Forest Algorithm

The random forest algorithm chooses the class that has received the most votes from among those classes made up of numerous created trees and their outputs, where the various trees' styles of class presentation [87, 99]. Another way to interpret this technique is as a synthesis of the forest's tree-structured classifiers. Since it uses its unique categorization as the tree building algorithm and builds each tree using different bootstrap samples, the random forest has an advantage over bagging. In opposition to the traditional trees, RF generates constrained groups of predictors, and under stochastic choice, each link is divided by employing the best predictor among these predictors. An upper barrier for the generalization error $(PE*)$ is established as the forest grows larger, preventing the overfitting issue without the need for huge datasets in RFA. Therefore, The RF algorithm (RFA) does not have this issue, however other algorithms might lead to an overfitting dilemma. The primary principle of RFA is to maximize the strength between nodes with the lowest correlation. Adaptive bagging is a useful foundation because it provides accurate estimates of the key metrics of strength, correlation, and generalization error. Additionally, by setting a higher limit, it reduces generalization error, which improves the precision of the generated networks, motifs, and modules. The generalization error of a random forest is determined by the potency of each distinct tree as well as the relationships between clusters of such trees. Besides, every single node can be divided by stochastic choices of features with a similar error rate as the other nodes.

In addition to this, the internal estimates, which show the correlation, strength and generalization error are used to demonstrate the reaction to the more features are being employed during the dividing stage. The relevance of the variable can also be determined using the internal estimates. Moreover, small communities (ensembles) are formed in random forests, and these ensembles decide which class is the most popular. The most popular method for producing these tiny communities is to produce random vectors, which are indicated by the symbol "$\boldsymbol{\theta}$". The random vector $\boldsymbol{\theta}_k$ that represents the kth tree in the forest is used to refer to that tree. In a forest, random vectors are $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_{k-1}$ resulting in $\boldsymbol{\theta}_k$ and a classifier $h(\boldsymbol{X}, \boldsymbol{\theta}_k)$, where $\mathbf{X}$ is an entry vector and k additionally denotes the forest's tree convergence. Each tree's growth is limited by these vectors in the neighborhood. Several techniques, such as bagging, stochastic break determination, and selecting the training set from an arbitrary number of weights, may be employed to generate these random vectors. The generalization error, the strength of the individual tree classifiers, and the reliance evaluation, which is the interaction between these classification algorithms, on the other hand, can all be used to describe the reliability of a random forest. Thus, the generalization error PE^* of RFA is checked as shown in Equation (2.6):

$$PE^* \leqslant \frac{\bar{\rho}(1 - s^2)}{s^2}. \tag{2.6}$$

Within this disparity, $\bar{\rho}$ represents an average correlation value between the random vectors θ and $\theta^{'}$, while $\theta^{'}$ represents the offering tree in the subsequent iteration. At this place, s stands for the resilience of tiny populations using the formula $s = E_{\boldsymbol{X},\boldsymbol{Y}} mr(\boldsymbol{X}, \boldsymbol{Y})$. E defines the expected value among the vectors of randomness $\boldsymbol{X}$ and $\boldsymbol{Y}$, and, the margin function $(mr(.))$. To accurately construct the trees, it is necessary to understand several fundamental aspects of the random forest method. To illustrate, in the random forest method, the strength and the correlation are also as important as the generalization error. The representation of this committee is taken as $h_1(x), h_2(x), ..., h_k(x)$ for creating an ensemble of classifiers. Equation (2.7) can be used to define the margin function utilizing the aforementioned characteristics.

$$mg(\boldsymbol{X}, \boldsymbol{Y}) = av_k I(h_k(\boldsymbol{X}) = \boldsymbol{Y}) - \max_{j \neq Y} av_k I(h_k(\boldsymbol{X}) = j). \tag{2.7}$$

In Equation (2.7), $\boldsymbol{Y}$ and $\boldsymbol{X}$ are the stochastic vectors $(Y = [1...j]), I(h_k(X) = ...)$ denotes the indicator function and av_k is the kth tree's average.

On the other side, It is mentioned previously that the generalization error's maximum threshold. It is now specified in terms of how the margin function determines it in order to enlarge it. Equation (2.8) illustrates the generalization error represented by the margin function by extending it:

$$PE^* = P_{X,Y}(mg(\boldsymbol{X}, \boldsymbol{Y}) < 0). \tag{2.8}$$

In the above expression, the generalization error is represented by PE^*.

The number of trees should be raised in accordance with the random forest algorithm's guideline. In this case, the generalization error converges to the following formula for the entire ensemble ($\boldsymbol{\theta}_1, ...$):

$$P_{X,Y}(P_\theta(h(\boldsymbol{X}, \boldsymbol{\theta}) = Y) - \max_{j \neq Y} P_{\boldsymbol{\theta}}(h(\boldsymbol{X}, \boldsymbol{\theta}) = j) < 0), \tag{2.9}$$

where $\max_{j \neq Y} P_{\boldsymbol{\theta}}(h(\boldsymbol{X}, \boldsymbol{\theta}) = j)$ represents the highest probability value across all classifier values, with the exception of its value at the location $\boldsymbol{Y}$ and $h(\boldsymbol{X}, \boldsymbol{\theta})$ .The random vector's classifier is represented by the number $\boldsymbol{X}$. Because the guideline in the random forest is to increase strength and decrease correlation as much as possible, resilience is one of the most crucial characteristics of the random forest method for the operation of building trees. Thus, by taking a look at the margin function, Equation (2.10) can be utilized to determine the effectiveness of the combination of classifiers:

$$s = E_{X,Y} mr(\boldsymbol{X}, \boldsymbol{Y}). \tag{2.10}$$

Currently, as in the subsequent inequality, it is suitable to link the strength and generalization error:

$$PE^* \leqslant var(mr)/s^2. \tag{2.11}$$

In Equation (2.11), PE^* is the generalization error and s^2 state the square of the strength. By inserting the ccorrelation to modify those formulas, we see the represantation below:

$$var(mr) = \bar{\rho}(E_{\boldsymbol{\theta}} sd(\boldsymbol{\theta}))^2 \leqslant \bar{\rho} E_{\boldsymbol{\theta}} var(\boldsymbol{\theta}). \tag{2.12}$$

In Equation (2.12), $\bar{\rho}$ refers to the average value of the $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{'}$ correlation, E presents the expectation, sd implies the standard deviation and var shows the variance.

Thus, all the aforementioned formulas are combined for the aim of defining an upper bound for the generalization error, as shown in Equation (2.6). The number of characteristics that will be chosen for each node is calculated based on the internal estimates. These internal estimates—also known as *out-of-bag* (OOB) estimates—belong to the dependence, classifier power, and generalization error. The following is a list of the various applications for out-of-bag estimates [87]:

- The predictions of the generalization error include OOB estimations as a component.

- For the arbitrary classifiers, the generalization error is inferred using OOB estimates of the variance.

- By employing a training set and test set with equal length, it is demonstrated that the OOB estimate is accurate.

- The estimate eliminates the requirement to set aside the test set by utilizing the OOB error.

2.1.3 Neural Network

A neural network model simulates the human nervous system. Neurons are cells that make up the human neurological system [5]. These neurons function as computational units. They receive information from other neurons, process it, and then transmit it to further neurons. The computing function which is placed at a neuron is identified by the weights on the input links to that neuron. One can compare the strength of a synaptic link to this weight. The computation function can be obtained by switching these weights conveniently, which is comparable to how organic neural networks acquire and develop synaptic strength. The training data in artificial neural networks ensures the "external stimulus" for determining these weights. Hereby, the main goal is to gradually alter the weights whenever inaccurate predictions are made by the current set of weights [5]. An unrivaled statistical method known as an artificial neural network (ANN) uses extremely powerful processing, vast amounts of memory, learning, and error tolerance. This means A computer system known as ANN uses complicated information processing to mimic the interconnection of neurons in organisms . Neural network is an adjustable system which has potential to learn. ANN may be taught with a variety of algorithms to ensure requested result via varied algorithms [58].

A *Neural Network* system is made up of processing units that interact and are closely coupled that support neuroscience-based algorithms. Neural networks process information through the interactions of a wide range of processor and their links to extrinsic inputs [58]. The impact of the neural network is the arrangement of the links among nodes. Here, Neural Network is an architecture that is responsible for this arrangement. Architectures come in a broad variety, consisting of “simple single-layer perceptron” and “complex multilayer networks” [5].

2.1.3.1 Single-Layer Neural Network: The Perceptron

The perceptron is the name of a neural network's most basic design. The two tiers of nodes that make up the perceptron are the input points and a single output point. Here, the dimensionality, d of the underlying data is equal to the number of input nodes. Each of these input nodes accepts one numerical property and sends it to the output node. For that reason, the input nodes merely transfer input values without doing any computation on them. The output node is the sole node in the primary perceptron model to apply a mathematical function to its inputs. The training data's specific attributes are expected to be numerical. It shall be assumed for the sake of simplicity that every input variable is an integer in the further discussion [5]. The function obtained by the perceptron is called the *activation function*, which is a signed linear function [5].

For a multidimensional data record d, let $\bar{W} = (w_1, ..., w_d)$ be the weights for the links of d different inputs to the output neuron. Furthermore, the function of activation is managed by bias b [5]. The output $z_i \in \{-1, +1\}$ for the feature vector $({x_1}^d, ..., {x_i}^d)$ of the ith data record $\bar{\boldsymbol{X}}_i$, is as follows [5]:

$$\begin{aligned} z_i &= \text{sign}\{\sum_{j=1}^{d} w_j {x_i}^j + b\}, \\ &= \text{sign}\{\bar{W}\bar{X}_i + b\}. \end{aligned} \tag{2.13}$$

Here, z_i is the perceptron's estimated value for the $\bar{X}_i$ class variable. For that reason, it is preferable to determine the weights so that z_i and y_i have the same value for as many training cases as probable. The error in estimation $(z_i - y_i)$ could receive on either of the values of -2, 0, or $+2$. When the estimated class is true, a value of zero is reached. The main purpose in the neural network algorithms is to learn the vector of weights $\mathbf{W}_t$ and bias b, in this way the difference between z_i and the true class variable y_i becomes as smallest as possible [5].

A stochastic vector of components is used as the initial state in the basic perceptron method. Afterwards, in order to generate the estimation z_i, the algorithm provides the input data elements $\bar{X}_i$ into the neural network one at a time. Then, based on the error value $(z_i - y_i)$, the weights are renewed. In particular, the weight vector $\mathbf{W}_t$ is renewed, when the tth iteration's input contains the data point $\bar{\mathbf{X}}_i$ [5]. It is seen as below [5]:

$$\bar{W}^{t+1} = \bar{W}^t + \eta(y_i - z_i)\bar{X}_i. \tag{2.14}$$

Here, η represents the neural network's pace of learning. The perceptron technique repeatedly sets the weights up to the convergence process is attained by recursively cycling through all of the training instances in the data [5]. One training data piece may be repeated numerous times. Here, each cycle is called an *epoch* [5]. Lower values of η lead to the convergence of better solutions, however the Convergence occurs slowly [5]. In operation, in beginning, the amount of η is selected to be high and step by step decreased, as the weights approximate to their ideal levels [5].

2.1.3.2 Multilayer Neural Networks

The input and output layers of multilayer neural networks are joined by a hidden layer. Theoretically, various topologies can be used to connect the nodes of the hidden layer [5]. To illustrate, The concealed layer could include further layers. Additionally, the nodes in one layer can connect to the nodes of the following layer. In fact, it is assumed that the nodes in one layer are completely linked to the nodes in the subsequent layer. This is called the *multilayer feed-forward network* [5]. Because of this assumption, once the analyst has calculated the number of layers and the number of nodes in each layer and the topology of the multilayer feed-forward network is automatically specified [5]. One layer of feed-forward networking can be compared to the fundamental perceptron. A multilevel feed-forward network with just one hidden layer is a well-known preferred model. A network such as this could be thought a two-layer feed-forward network [5]. The multilayer feed-forward network has additional qualities, such as the feature that it is not limited to using inputs' linear signed functions. Any arbitrary functions, like the logistic, sigmoid, or hyperbolic tangents, can be applied using the variable nodes of the hidden layer and output layer [5].
To give an example to this kind of function, the training tuple $\bar{\mathbf{X}}_i = ({x_i}^1, ..., {x_i}^d)$, yield, an output value of z_i, as below [5]:

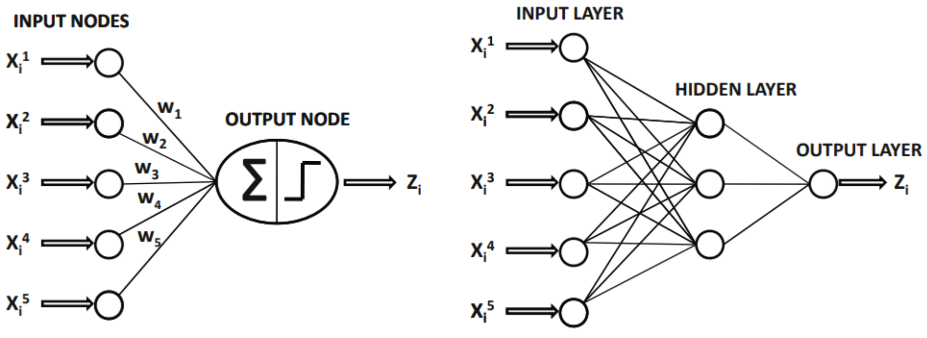
$$z_i = \sum_{j=1}^{d} w_j \frac{1}{1 + e^{{x_i}^j}} + b. \tag{2.15}$$

If here is a function which is calculated at the nodes of hidden layer, then the value of z_i can not be an estimated ultimate class label output $\{-1, +1\}$ any more. Afterwards, this output is expansed forward to the subsequent layer [5]. Since the training label value is known to be identical to the predicted output of the output node, the training process of the single-layer neural network is rather simple. The "gradient-descent method" is used to update the weights after using this "ground truth". which is utilized to construct an optimization problem in the least squares form. Since in the case of a single-layer network, the sole neuron with weights is the output node, the renewed period is simple to apply. The problem arises for multilayer networks because the hidden layer nodes' ground-truth output is unknown because the outputs of these systems lack training labels. For that reason, when a training example is applied inaccurately, the weights of these nodes should be calculated. Obviously, when an error is obtained, different types of "feedback" on expected outputs and associated mistakes are needed from the forward layers to the nodes in prior layers. With the use of the reverse propagation algorithm, this process is man-aged [5]. The backpropagation algorithm includes two main phases [5]:

Forward phase: Here, the neural network is provided the inputs needed for training. As a result, in each layer's forward step of computations, the present array of weights is used. The ultimate estimated output might be contrasted to the class label of the training process, to control the estimated label is an error or not [5].

Backward phase: By providing an estimate of a node's output error in the early layers based on errors in the later layers, this phase's basic objective is to establish weights in the reverse direction. In the hidden layer, the weights and error estimates of the nodes in the layer's foreground are used to determine a node's error estimate. Following that, employing this, the weights of the node are revised and an error gradient is generated. On a cognitive level, the original apprised equation is comparable to the fundamental perceptron. There are a couple of differences that occur because of the nonlinear functions, which are generally applied in hidden layer nodes and errors at these nodes are mostly calculated via "backpropagation", instead of being derived directly by comparing the output to a learning sign [5]. This whole process is outspread backward to renew the network's nodes' average weights [5].

An illustration of a perceptron and a two-layer feed-forward network is illustrated as follows [5]:



(a) Perceptron (b) Multilayer

Figure 2.2: Single and Multilayer Neural Networks [5].

2.2 Two-stage Machine Learning Approaches

We consider the application of the two-stage machine learning models in two cases. **At first case**, we apply the clustering method to our dataset, and then we use one of the regression methods to each one of these sub groups that we obtained in the clustering part. As an example to these clustering methods, we can employ Hidden Markov Model (HMM), Support Vector Machine (SVM), Fuzzy Clustering, etc. On the other hand, for the regression methods, we consider to implement MARS. However, during our researches, we also think to discuss the application of the MARS alternatives such as Conic MARS (CMARS), and Robust CMARS (RCMARS) [69, 72, 96, 100, 101, 68].

At the second case, we focus on data mining methods. Initially, we apply one of these approaches. According to the results, we discard meaningless and less important variables from out dataset. Afterwards, we use the remaining dataset for the MARS model. To give an example, we implement the Random Forest (RF) to isolate less important variables and then use MARS to discuss remaining dataset or vice versa. We can also work with other data mining methods such as the neural network, CART as SVM. Indeed, the advantage of such a mixing approach is also cited in the literature. Lee et al. (2005) a sought to determine how well credit scoring performed utilizing a two-stage hybrid modeling process combining MARS and artificial neural networks (ANNs). Their investigation follows a methodology that first creates a scoring model using a MARS credit rating approach, employing MARS as a supporting tool for neural networks, then uses the gleaned important variables as the model's input nodes for envisioned neural networks. The hybrid credit scoring model, compared to logistic regression, linear discriminant analysis, back propagation neural network (BPN) and MARS, has the best credit scoring capabilities, according to analytical data [57]. Furthermore, when compared to logistic regression techniques, Sledjeski et al. (2008) investigated whether Classification and Regression Trees (CART) (pattern-centered) analysis would more accurately identify families with a high likelihood of a repeat incident and ensure a more informative risk story. By ensuring that families receive the right resources, child protective services (CPS) aims to decrease the possibility of child abuse and its iteration. According to the CART analysis, the biggest predictor of future abuse was prior CPS involvement [89].

The hybrid system that Andres et al. (2011) presented combines MARS with fuzzy clustering. They evaluate the efficacy of their methods in a real environment using a database made up of 138 troubled enterprises that filed for bankruptcy during 2007 and 59,336 Spanish companies that are still in business. They also performed the discriminant analysis, a feed-forward neural network and MARS. The study showed that the author's proposed hybrid model has better results with respect to the other systems, not only in regards to both in terms of the profitability produced by lending decisions and the number of accurate classifications [30]. Yao et al. (2013) studied the hybrid of MARS and RF to predict the affliction. At first, they used an initial evaluation of variables and determined important rankings using the RF algorithm. According to these ranks, the new dataset is generated by selected critical predictors and input into the MARS process, which creates understandable models to estimate the disease survival. To compare RF, MARS and RF with MARS; the classification accuracy of RF and MARS is marginally better than the solitary RF model, but marginally worse than the MARS model. At the final part of this study, the outcomes of the RF/MARS hybrid algorithm is evaluated in comparison to the C4.5 algorithm (a decision tree classifier used in data mining) and the SVM algorithm. The findings from experiments indicate that the suggested approach assures a more accurate model that is also fairly simple to understand and use [99]. Additionally, Lin et al. (2013) aimed to show that the interactions of single nucleotide polymorphism (SNP) which play an significant role for understanding reasons of the complex disease. Hence, in order to describe a subset of the key SNPs and discover patterns of interaction, they presented an integrated strategy that incorporates two techniques, RF and MARS. A predictive subset of SNPs is revealed by this two-stage RF-MARS (TRM) technique using RF (here, RF variable selection is based on out-of-bag classification error rate (OOB) and variable important spectrum (IS)), and MARS to analyze the interaction patterns between the chosen SNPs. Their study showed that $\mathrm{RF}_{\mathrm{OOB}}$ performed more effectively in identifying the crucial variables than MARS and $\mathrm{RF}_{\mathrm{IS}}$. To investigate the SNP-SNP interactions in a large-scale genetic variation, $\mathrm{TRM}_{\mathrm{OOB}}$ is preferable for this reason [61].

CHAPTER 3

APPLICATION OF MACHINE LEARNING TECHNIQUES INTO THE SENTIMENT INDEXES

3.1 Application of Machine Learning Techniques into Investor Sentiment Index

In our thesis, we indicate the application of machine learning models into the sentiment indexes. First of these indexes is called as *Investor Sentiment Index* and *Consumer Confidence Index*, respectively. Following part, we begin with introducing these indexes and then see the results of each of indexes' application.

3.1.1 Human Factor: Investor Sentiment

Numerous financial decisions that have a significant impact on people's lives are made at different periods of the economy [38]. Financial performance of investors typically suffers when they are emotionally receptive and have weak impulse control. With the help of some techniques, it can be determined which parts of the brain are particularly active during decision-making processes, particularly when it comes to financial decisions [48, 88, 93]. Behavioral finance is the study of how emotions and attitudes affect investor behavior. Economists take into account a variety of actual facts about investor behavior [28, 39, 48]. Thinking methods can be influenced by feelings. Positive moods encourage more original thinking, although these solutions may be riskier, whilst low moods may encourage more cautious thinking. Optimistic or gloomy expectations of this nature can linger and influence asset values for protracted periods of time, eventually resulting in crises [18, 104]. *Sentiment* is the collective term for all of these emotional situations as presented previously. Financial experts, academic researchers, and the media all take the term "sentiment" into consideration in different ways, and many of them conclude that investor sentiment is significant from an economic standpoint. Some scholars identify investor sentiment as a propensity for making investments on noise rather than facts. In other respects, among the some scholars consider the Sentiment term as “investor optimism or pessimism”. This phrase is emotionally charged, thus news organizations will occasionally refer to it as *investor anxiety* or *risk-avoidance* [47, 102].

While investigating assumptions and descriptions about Sentiment, it is seen that it may be possible to measure the Sentiment. Options traders utilize a variety of sentiment indicators for a number of reasons, varying from daily indexes to metrics developed for scholarly articles [47].

Some frequently used sentiment indicators include investor intelligence surveys, market liquidity, implied volatility of index options, ratio of odd-lot sales to purchases, closed-end fund discount consumer confidence indices, and net mutual fund redemptions. These metrics show that neither academic scholars nor experienced traders can provide a concise, all-encompassing explanation of sentiment. The former uses measurements of investor sentiment to promote market efficiency or to explain the cause and impact of certain movements. The latter adopts Sentiment indicators as a probable trading tool [47, 102]. Therefore, there are disagreements over how to measure it. The two methods of “direct survey data” and “indirect market-based proxies for sentiment” are reviewed for gauging sentiment. The former is the *driven by markets* strategy, which aims to learn indirectly about sentiment via financial substitutes. The put-call ratio and closed-end fund discount are frequently cited as sentiment indicators derived from reactions in the market to show this. The latter strategy involves using surveys and questionnaires that are sent to investors to *measure sentiment directly*. Examples include the Stock Market Confidence Index (YSMI) from Yale School of Management, the Consumer Confidence Index (MCCI) from University of Michigan, the Global Market Sentiment Survey from CFA Institute, and according to the American Association of Individual Investors’ (AAII) Weekly Investor Sentiment Survey (AAII) [47, 102]. In order to assess Sentiment, Baker and Wurgler (2006) favored creating an index rather than employing a single measure of investment sentiment by using 10 proxies. Inputs that are used for Sentiment Index can be listed as follows: the closed-end fund discount, NYSE share turnover, the number and average first-day returns on initial public offerings (ipo)s, the equity share in new issues (s), and the dividend premium, industrial production index (indpro), nominal durables consumption (consdur), nominal nondurables consumption (consnon), nominal services consumption (consserve) , NBER recession indicator (recess), employment (employ), and consumer price index (cpi) [11, 12, 47].

The economic and financial wellness of countries are highly based on the behavior and sentiment of those countrys' financial sector. The well-functioning financial institutions is a fundamental construction of economic development. The influence of these financial institutions and their role in improving economics, are played a significant role to build the finance industry in the countries [10]. To address the actual issues facing the financial industry, the disciplines of stochastic calculus, numerical analytics, scientific computers and financial mathematics are also joined. The application of these computational techniques to finance and sentiment of investors is important for the business world to create and investigate strategic planning by providing perspective into what might become in the future if a strategy is performed, and forecasting the risks related to financial instruments [46]. Even though on a large scale, used traditional statistical approaches in constructing predictions in financial and investor sentiment fields these methods of treatment are founded on constricting presumptions such linearity, normalcy, and independence between independent and dependent variables. Therefore, because of these fields' stochastic and complex nature, it is a compelling task to forecast with the help of classical statistical techniques [6, 59, 64]. That is why, these traditional models are given rise to a growing interest in machine learning techniques [10]. Numerous researchers examine a lot of techniques from Machine Learning and Data Mining to solve problems in finance [46]. The success of machine learning techniques is based on their property to model nonlinear systems with minimal initial assumptions and advanced forecasting accuracy [77]. Many investors' decisions and opinions in the fields of financial industry are searched and applied by machine learning techniques. This provides help the finance industry to comprehend the data and obtain using the data for a competitive edge [46]. On the other hand, Hybrid models which combine the several machine learning techniques prone to improve steadily more [30, 46]. The association that machine learning generated is proved by prospering for hybrid methods. However, the process should always take into consideration the significance of data integrity, as the ambiguity of the resilience of machine learning algorithms is constantly required for data [46].

From this point of view, various approaches are employed in the literature via either single machine learning model or hybrid machine learning model: Bahrammirzaee (2010) studied a comparative research review about three well-known artificial intelligence methods in financial sector; which are artificial neural networks (ANN), hybrid intelligence systems, and expert systems. In his study, financial market is classified and explained on three aspects, namely, portfolio management, credit evaluation, planning and financial prediction. This study shows that the artificial intelligent techniques obtain more accurate results than using conventional statistical techniques to address financial issues [10]. In order to predict stock index prices, Lu et al. (2010) assessed the performance of backpropagation neural network (BPN), MARS, backpropagation neural network (BPN), multiple linear regression (MLR), and support vector regression (SVR) models on the Shanghai B-Share stock index. According to empirical findings, In terms of accuracy and prediction error, MARS performs better than MLR, BPN, and SVR [64]. Furthermore, Li (2010) proposed a new approach based on the a two-phase combination model of logistic regression and ANN, for the purpose of building a financial distress warning system in the banking industry while developing markets between 1998 and 2006. This study takes into account a two-phase combination design that combines the advantages of ANN and logistic regression while avoiding computing complexity. In order to demonstrate that the suggested approach performs better than conventional models, some novel treatments are used. The findings show a strong correlation between asset quality, capital, and liquidity determinants and banks' financial difficulties in developing economies. The suggested two-phase combination design, which performs better than the traditional ones in terms of prediction power, gives the best fit grounded on the R-Squared (R^2) and Root Mean Square Error $(RMSE)$ and Root Mean Square Error $(RMSE)$ metrics for the identification of institutions in financial crisis [62]. Kao et al. (2012) suggested Wavelet-MARS-SVR is a new stock price forecasting model that combines support vector regression (SVR), multivariate adaptive regression splines (MARS), and wavelet transform to increase prediction accuracy. The productivity of these underlying methods is utilized by contrasting the Wavelet-MARS-SVR prediction outcomes with those of the other five opposing techniques, namely, Wavelet-MARS, Wavelet-SVR, single SVR, single ARIMA, and single ANFIS. The findings of this research demonstrate that the suggested strategy outperforms other competing models [49]. Jadhav et al. (2016) performed to investigations into data mining methods for the banking sector from 2010 to 2015. According to reviews, in comparison to loan prediction, money laundering, and time series prediction, researchers are particularly interested in stock prediction and credit rating. On the other hand, because of dynamics, uncertainty, and a variety of data, nonlinear mapping techniques are more thoroughly explored than linear ones. Additionally, in this study, it is shown that combination approaches are nearly as precise as the neural network technology in their predictions [46].

3.1.2 Application

After all these studies that are mentioned above, in this study, it is indicated the applicability of three machine learning techniques by only themselves and also by the hybrid model to forecast sentiment level of investors. This study distinguishes the forecasting capabilities of MARS, Random Forest (RF) and Neural Network (NN) models with each other and then by constructing two-stage hybrid models such as MARS-NN, MARS-RF, RF-MARS, RF-NN, NN-MARS, and MARS-NN to predict sentiment levels of investors. Here, our main purpose is dimension reduction, therefore, we use only train parts in our models. To interpret the performance of these approaches, the Sentiment index which is constructed by Baker and Wurgler (2006) by using 6 proxies are adopted. While constructing this index, raw data are produced by 10 proxies. Then, these raw data are used for the application of machine learning techniques. Therefore, inputs that are implemented for Sentiment Index are listed as NYSE share turnover, the closed-end fund discount, the number, and average first-day returns on IPOs (initial public offerings), the dividend premium, the equity which shares new issues, indpro (industrial production index), consserv (nominal services consumption), consdur (nominal durables consumption), consnon (nominal nondurables consumption), CPI, and employ [11, 12, 47]. Finally, these data are monthly and in this study, it is used from the year 2000 to the sixth month of 2022. To evaluate the accuracy of the model, we need to know the meanings of some statistical measure as presented below:

(i) Mean Squared Error (MSE): the data set's the average square of the gap between the actual and anticipated values.

(ii) Root Mean Squared Error (RMSE): the data set's total squared variance among the actual and projected values, expressed as a root square.

(iii) Average Absolute Error (AAE): the typical difference between a set of data' real values and anticipated values, without regard to the direction of the difference.

(iv) R-Squared (RSq): a gauge of the model's ability to match the training set of data [34].

(v) Generalized R-Squared (GRSq): an estimate of the predictive power of the model (calculated over all responses) [34].

Hereby, at first, we begin our analyses with forecasting by the MARS method. There are 12 inputs which are mentioned above and 1 output which is sentiment itself. For this data set, using the R Programme, the highest degree of interactions is taken as 3 since this is the most preferable interaction level in majority of the financial analyses due to its convenience in interpretation. The number of basis functions is set to 100. To apply two-stage case, we ignore less important variables, thereby, we eliminate employ, consdur and consnon variables. Then, we take the remaining data set for modelling.

MARS-RF implies that MARS is used as a first-stage modeling tool while the obtained results are taken for the RF model. On the other hand, RF-MARS describes that RF is performed initially and its outputs are implemented for the MARS model in the second stage. Hence, Table 3.1 shows the outcomes of remaining models, namely, MARS, MARS-RF and MARS-NN. In the tabulated value, MSE denotes the mean square error, RMSE defines root mean square error and AAE denotes average absolute error values. The coefficients and interactions of basis functions are showed detailed in the Appendix A.

Table 3.1: The result of MARS, MARS-RF and MARS-NN models based on sentiment index.

Methods	MARS	MARS-RF	MARS-NN
MSE	0.0050	0.0670	4.2890
RMSE	0.0730	0.2590	2.0710
AAE	7,19E-12	0.0080	2.0708

According to Table 3.1, we observe that the MARS model itself outperforms better comparing with the other two-stage hybrid models. On the other side, between two-stage hybrid models, MARS-RF has also better results than the MARS-NN model.

Secondly, we apply the RF model for the same data in order to predict prediction Sentiment. To apply two-stage case, as we explain in former parts for the second case of two-stage approaches, we need to eliminate meaningless variables. Thus, we ignore less important variables, which are taken as consserve and cpi. Statistical results of these models are indicated in Appendix A. Initially, we merely take the remaining data set for the MARS model (RF-MARS) and then for the NN model (RF-NN). As it is seen from the Table 3.2, we observe that instead of one-stage RF model, RF-MARS two-stage hybrid model outperforms better this time.

Table 3.2: The Result of RF, RF-MARS and RF-NN Models based on sentiment index.

Methods	RF	RF-MARS	RF-NN
MSE	0.0260	0.0040	1.0450
RMSE	0.0730	0.0670	1.0220
AAE	0.0030	4,41E-12	1.0230

Thirdly, we apply the NN model to forecast Sentiment. The application is done again for train part, however, to check we also show 80% for the training and 20% for the test period. According to the estimated weights of variables, only employ, consnon and conssserve variables are remained since they are the only ones with value 1. Later, we implement the remaining data both for the MARS model (NN-MARS) and RF model (NN-RF).

According to Table 3.3, it is seen that we have better outcomes by using NN-MARS

Table 3.3: The result of NN, NN-MARS and NN-RF Models based on sentiment index.

Methods	NN	NN-MARS	NN-RF
MSE	0.4730	0.0110	0.0310
RMSE	0.6880	0.1040	0.1760
AAE	0.1923	1,35E-12	0.0050

model as a two-stage model.

The results of both one-stage and two-stage machine learning algorithms' statistical results and detailed information about all variables for are indicated in Appendix A.

To conclude, we can say that the MARS model itself outperforms better comparing with the other two-stage hybrid models. On the other hand, RF-MARS two-stage hybrid model which is RF-MARS, has also better results with respect to the single RF model and the RF-NN model. Third best results is obtained by the NN-MARS model. Hence, based on these outcomes, we can observe that by using MARS model only itself and by employing this model as a second stage provide us better accuracy. Furthermore, employing MARS as a first-stage analysis tool and the resulting outputs as RF's inputs are contributed to the achievement of the model [57]. As we say in the former parts, the process should always take into consideration the importance of data structure. On conclusion, MARS single model, RF-MARS and NN-MARS hybrid models achieve better performance for the sentiment data regarding other models.

3.2 Application of Machine Learning Techniques into Consumer Confidence Index

3.2.1 Another Sentiment Index: Consumer Confidence Index

The principles of sentiment are presented by economic conditions such that most of the variance in consumer sentiment is caused by economic situations, either directly or indirectly. There are political and economic repercussions to investors' economic optimism or pessimism. Customer expenditures and hence the direction of the economy's future are predicted by the degree of consumer confidence. Investors, who are in a similar way, positioned economically but have different biased thoughts bring forward quite different sentiments about the future of economics. When the current situation appears favorable, people are more upbeat about the economy both now and in the future. Economic judgments become more negative, particularly when inflation or unemployment rates rise. The general public's expectations about the economic future rise when major economic indicators indicate favorable times will soon arrive [31]. The Consumer Confidence Index plays a significant role in informing decision-makers and economic forecasters about the current and upcoming state of the economy. These indicators serve a special significance in influencing both commercial and governmental policy. The consumer confidence index measures how optimistic consumers are feeling about the state of the economy based on their spending and saving habits, which contribute to national economic expansion. Positive improvements in consumer confidence should be the source of economic growth, while negative changes should prevent it. Numerous researchers have attempted to determine how macroeconomic factors and consumer confidence indices are related. Consumers' replies to inquiries regarding the present and future state of the national and personal economy form the foundation of consumer confidence indices [45].

The expectations of economic actors have a substantial impact on the path that macroeconomic and financial indicators will take. Following financial and macroeconomic indicators ensures advance knowledge of the future direction that indicators will monitor. Because of this, confidence indices that reflect expectations operate as the primary indicators in addition to economic indicators for policy makers, participants in the financial markets, and representatives of the real estate sector. Keynes linked "situation of long-term expectations" and "confidence situation" with with being vulnerable to changes in the economy's feeling. According to analysis of Keynes, the susceptibility of consumers and producers to economical improvements acts a crucial part in the statement's economic variations. Consumer confidence and macroeconomic factors have a direct relationship. Considerations, Feelings and the economic decision-making process are not only impressed by numerous macroeconomic variables but also they are affected by psychological, political and sociological choices. The expectations and actions of economic decision-making units are influenced by the confidence index, which is a key indicator for the economy [14]. There are many studies about the effect of macroeconomic variables and consumer confidence index to each other.

Afshar et al.(2007) used quarter data for the USA from 1980 to 2005 to evaluate the links between the consumer, investor, and business confidence indices and economic oscillation. VAR and a vector error correction model were used. Stock returns, purchasing managers' and consumer confidence index all reveal the enormous gap in the Gross Domestic Product (GDP). The results generally approve the considerations that demonstrate in terms of economic swings, confidence indices play a significant impact [4, 14].

Korkmaz et. al. (2009) investigated the causal relationship among the BIST 100 index and Real Sector Confidence Index return by two-stage dynamic association test. They first used the EGARCH model to estimate the relationship between the variables, and then they investigated the causal link between the mean and variance of the model's standardized error terms [14, 53]. In order to determine the impact of the consumer and real sector confidence indices on the Turkish economy, Arısoy (2012) created two different VAR models using monthly data for the variables Consumption Expenditures, Industrial Production Index, Consumer Confidence Index, Real Sector Confidence Index, Employment Rate, and BIST Index between 2005: 01 and 2012: 01. The study's findings revealed that advances in industrial production and stock index are influenced by the Real Sector Confidence Index and the Consumer Confidence Index, respectively [8, 14].

Using the Consumer Confidence Index for the USA and the Euro-zone, consumption expenditures, real disposable income, financial and real estate wealth, real stock prices, short-term interest rates, unemployment rates, and quarter data from the 1985–1985 and 2010–2010 periods, Dees et al. (2013) studied the connection between the variables using VAR and Threshold Models. They indicated that the US consumer confidence index serves as a "security channel" that ensures the shocks' transitivity and that it influences the Eurozone consumer confidence index [14, 32].

It is beneficial to understand the connections between some macro and financial variables and the Consumer Confidence Index, a measure of consumer confidence in economies. A gain in investor and consumer confidence in the nation's economy may result from positive improvements in the factors that may impact such confidence levels. Başarır et al. (2019) purposed to employ a VAR model to analyze the relation between the chosen macroeconomic and financial variables and Consumer Confidence Index in Turkey. As a consequence of the study, it was discovered that there was a causal relationship between the consumer confidence index and the industrial production index as well as between the consumer confidence index and the BIST100, CPI, and USD Exchange Rates. Furthermore, when the VAR model's consequences are reviewed, it is discovered that the USD exchange rate shock has a detrimental impact on consumer confidence index, which in turn negatively affects the dollar exchange rate and BIST100 index [14].

In our study, we aimed to see the effect of economic variables to consumer confidence levels. We work with machine learning techniques as we did in the previous part (3.1.2).

3.2.2 Application

In this study, it is indicated the applicability of three machine learning techniques by using economic variables: The Consumer Confidence Index (CCI), Unemployment Index, Consumer Price Index, USD/TRY Index. All these datasets are monthly, starting from 2005 to up to now. The CCI, Unemployment Index and CPI are taken from Turkish Statistical Institute (TURKSTAT) and USD/TRY exchange rates are taken from the Central Bank of Turkey (TCMB, EVDS Data Central).

We control the influence of all variables to each other, thus, at each step, we address another of them as a output variable. We begin our analyses by modeling the data via the MARS method. We use 3 inputs and 1 output which is CCI, as mentioned beforehand. Moreover, we apply the highest degree of interactions as 3 as our previous analyses, but, we set the number of basis functions to 50 since we observe that increasing basis functions after 50 (based on our assessment of the basis function's numbers from 50 to 100) don't make any difference. As seen in Table 3.4, the most efficient result is obtained by using the unemployment rate as an output. Then, we apply the RF model for the prediction of effect of our variables to each other. For the RF model, similar to MARS, the most efficient result is found for the unemployment rate as an output with 0.832 for MSE, 0.912 for RMSE and 0.012 for AAE. Finally, we apply the NN model and it is seen that the most efficient model is constructed when the currency index (USD/TRY) is taken as output with 0.004 MSE, 0.064 RMSE and 0.063 AAE. The results are summarized in Table 3.4. From the tabulated value, we observe that typically, the lowest statistical measures are obtained for the neural network model. Afterwards, as applied to our first sentiment data, here, we detect the application of two-stage case for the consumer confidence index (CCI). Similar to previous analyses, we ignore less important variables and eliminate the unemployment variable. Then, we take the remaining data set for the RF (MARS-RF) and NN models (MARS-NN). Furthermore, as applied in the one-stage part, we also use our other variables (UN, CPI, USD/TRY) as output in the two-stage part. From the assessment of the coefficients and importance in Table 3.5, it is seen that when unemployment rate is output, CPI is discarded. On the other hand, when CPI is output, CCI is eliminated and lastly, when the USD/TRY is taken as output, UN is removed. Furthermore, for the MARS–RF, the best result is obtained when the UN is output with 1.184 MSE, 1.088 for RMSE and 0.009 for AAE. Furthermore, for MARS-NN model, the most effective result is acquired when the CPI is output where the MSE is found as 0.011 and the RMSE is found as 0.103.

Table 3.4: The result of sole MARS, RF and NN models based on CCI.

Method	Variables	MSE	RMSE	AAE
MARS	CCI	2.8130	1.6770	1.3077
	UN	0.3930	0.6270	0.5060
	CPI	53.9000	7.3480	5.5580
	UTRY	30.1600	5.4920	4.1518
RF	CCI	7.2610	2.6950	0.0360
	UN	0.8320	0.9120	0.0120
	CPI	1070.9	32.7250	2.0350
	UTRY	1664.8	40.8030	4.4670
NN	CCI	3.1040	1.7620	1.7618
	UN	6.6100	2.5710	2.5713
	CPI	0.0120	0.1090	0.1091
	UTRY	0.0040	0.0640	0.0630

Later, we apply the RF model for the prediction of CCI under two-stage models. We implement RF-MARS and both RF-NN and NN-RF models in the analyses whose model performances are presented in Table 3.6. In the first model (RF-MARS), when

Table 3.5: The result of MARS-RF and MARS-NN models based on CCI.

Method	Variables	MSE	RMSE	AAE
MARS-RF	CCI	5.9950	2.4490	0.077
	UN	1.1840	1.0880	0.0090
	CPI	714.7000	26.7350	1.660
	UTRY	708.3000	26.6140	3.3170
MARS-NN	CCI	4.4410	2.1070	2.1075
	UN	5.0420	2.2450	2.2453
	CPI	0.0110	0.1030	0.1032
	UTRY	0.0300	0.0520	0.0520

unemployment rate (UN) is taken as output, CCI is discarded since its associated regression coefficient is found as unimportant to explain UN. Furthermore, when CPI and USD/TRY are outputs separately, CCI is eliminated for both cases. Hence, among the RF-MARS models, the best result is found when the CCI is set to output. On the other hand, from the RF-NN model, the most efficient model is found when USD/TRY is output. The associated MSE is computed as 0.052, RMSE is calculated as 0.003 and AAE is computed as 0.052. Moreover, when we change the order of RF and NN models that is under NN-RF model we observe that MSE values for all alternative models increase with respect to RF-NN model. Therefore, we conclude that even though we implement the same models for the data, the order of the two-stage construction affects the results. On the other side, from the application of NN model into MARS model, i.e., NN-MARS, as seen in the last part of Table 3.6, when UN is set to output, CCI and CPI are eliminated; when CPI is taken as output, CCI and UN are discarded, and when the output is chosen as USD/TRY, UN and CPI are eliminated since their contributions to the outputs are found unimportant. From this comparison, the best accuracy is found when the UN is set to output. All these results for both one-stage and two-stage cases are detailed shown in Appendix B.

To conclude, when we assess all the tabulated values, we can say that the MARS model itself outperforms better comparing with all two-stage models. On the other hand, among two-stage models, MARS-RF has better results with respect to the single RF model and the NN-RF model. Furthermore, we observe that single NN model manages less error values than two-stage models. Accordingly, on summary, MARS single model, NN single model, and MARS-RF, MARS-NN, RF-NN two-stage models achieve better results in this kind of consumer confidence data based on the MSE

Table 3.6: The result of RF-MARS, RF-NN, NN-RF and NN-MARS models based on CCI.

Method	Variables	MSE	RMSE	AAE
RF-MARS	CCI	3.7040	1.9240	1.5054
	UN	0.6540	0.8100	0.6272
	CPI	169.5200	13.0200	9.8673
	UTRY	73.5800	8.5780	5.4770
RF-NN	CCI	5.7810	2.4040	2.4050
	UN	10.7120	3.2730	3.2740
	CPI	0.1070	0.0120	0.1073
	UTRY	0.0520	0.0030	0.0520
NN-RF	CCI	5.9910	2.4480	0.0780
	UN	2.0040	1.4160	0.0580
	CPI	365.0200	19.1050	0.7664
	UTRY	19036.7000	137.9740	1.7969
NN-MARS	CCI	3.7040	1.9240	1.5053
	UN	1.4610	2.1360	1.1203
	CPI	285.9000	16.9100	13.0318
	UTRY	12960.0200	113.8420	77.0450

model selection criterion. In all these analyses, this criterion is chosen for the overall comparison of all models since this is the unique criterion that is computed for all MARS, NN and RF based models. Statistical results of all these outputs are written in Appendix B.

CHAPTER 4

FINANCIAL VOLATILITY MODELS

While trading in financial markets, one of the key role is to try to seize the movements of the underlying asset. These movements are called *volatility*. The volatility, represented by the symbol σ_t, is the conditional standard deviation of the return on the underlying asset. The volatility possesses some remarkable properties. The volatility swings with time is one of those crucial characteristics. Daily data do not immediately reveal this, since each trading day only includes one observation. But, there are more features of the volatility. For instance, it seems in clusters, it stops at particular spans and does not continue to expand indefinitely. Moreover, when the fundamental asset's value declines, the volatility responds differently than when the price of the underlying asset rises. Additionaly, it is contingent upon the trading in each day and between the days (the over night volatility) [40]. Therefore, the aim of the volatility analysis is to describe the factors that produce and affect volatility. Here, forecasting benefits greatly from the time series structure. The estimate approach provided for autoregressive conditional heteroskedasticity (ARCH) and generalized ARCH (GARCH) models can be easily utilized if exogenous or predefined variables are present. Therefore, both the variance and the mean can be thought of as parts of the estimation problem. To find the optimum formulation, we can also complete description analyses and hypothesis testing. It should not be a surprise that the volatility might be interpreted as a response to news. However, the timing of the news may not have been unexpected. This results in the volatility's predictable elements, including economic updates. Other news events have an impact on the amount of news events. For instance, the volatility seen in Asian markets earlier in the day as well as the volatility seen in the United States the day before may have an impact on the magnitude of return moves on the American stock market [35].

Market volatility can cause changes in the way financial assets are distributed in terms of risk [81]. These changes naturally affect the sentiment of investors. If we bind this approach with the behavioral finance, financial choices are typically predicated on a trade-off between risk and reward, as can be shown. For that reason the econometric study of risk is a crucial component of risk management, portfolio optimization, option pricing, and asset pricing. Thus, the ARCH and GARCH models have been carried out to a diversified amount of time series analyses and these applications in finance have been successful in particular [35].

In order to investigate the connection between investor sentiment and market volatility on the Johannesburg Stock Exchange, Rupande et al. (2019) employed a daily sentiment composite index built from a collection of proxies and GARCH models. The findings indicate a considerable relationship between investor confidence and stock return volatility, demonstrating how behavioral finance may effectively explain how stock returns behave on the Johannesburg Stock Exchange. Additionally, their study stressed how noise trading activity, which is driven by investor sentiment, increases financial market volatility on the South African market [81].

In our study, we aim to use MARS and other data mining techniques to our sentiment and finacial datasets as an alternative of volatility models under certain assumption. Furthermore, there are some studies which forecast both with volatility models and machine learning models;

Dias et al. (2014) suggested using model-based Regime Switching Model (RSM) clustering to find 21 European stock markets' results. They define three regimes: what are known as "bull and bear regimes", a stable regime with returns that are near zero, which ends up being the regime that occurs the most frequently. For the statistical study of these financial time series, the mixture Gaussian Hidden Markov Model (MGHMM) is suggested. The bull and bear regimes "GJR-AR(1)-GARCH(1,1)" model with Student-t innovations, which gathers both asymmetry and non-normality, is a model that is commonly employed to evaluate stock market indices. The findings show that their proposed RSM outperforms the GARCH model in terms of in-sample forecasting. Both models show similar performance when it comes to out-of-sample prediction [33].

Katris (2019) examined and evaluated time series and machine learning algorithms for prediction unemployment over various time periods in a number of nations. His paper is about twenty-two countries' unemployment to be predicted using FARIMA (Fractional ARIMA), FARIMA/GARCH, ANN, SVR and MARS models utilizing monthly seasonally adjusted data for unemployment. From the outcomes, no single model is found to be completely reliable, and while choosing an approach, it is advised to consider both the forecasting horizon and the locality. While neural network approaches produce results equivalent to FARIMA-based models for the longer period ($h = 12$), FARIMA models were found to be clearly the preferred approach for forecasts one step ahead. Holt-Winters model was discovered to be better appropriate for $h = 3$. [50]. From this point of view, we consider to compare the results of volatility models with machine learning based model.

Below, we present the details of the most well-known volatility models, namely, ARCH, GARCH and EGARCH, discuss shortly how they can be compatible with MARS.

4.1 Autoregressive Conditional Heteroskedastcity (ARCH) Model

To capture the return on an asset, the Autoregressive Conditional Heteroskedastic (ARCH) models are employed. The ARCH model is predicated on two assumptions regarding the features of the volatility for any time-based financial series. The former presumption holds that there are clusters of high volatility, and that the movement of an asset's return depends on its previous values. However, it is unrelated across the whole time series. According to the latter hypothesis, a quadratic function of the earlier delayed values can explain the distribution of asset returns (a_t), as it is dependent on earlier values [40]. The input collection present at time $(t-1)$ is used to build the model.
The prior m lag innovations determine the conditional variance [40] via

$$\sigma_t^2 = \omega + \alpha_t a_{t-1}^2 + ... + \alpha_m a_{t-m}^2. \tag{4.1}$$

In Equation (4.1), σ_t^2 denotes conditional variance, a_t denotes error term (return residuals), ω and α are the parameters of the model. By squaring them, we can observe that substantial changes in return innovation exert a more significant impact on the conditional variance. This implies that significant disturbances have a tendency to occur in succession, which is similar to how volatility clusters behave [40].

Thus, the ARCH (m) model can be described as

$$a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \omega + \alpha_t a_{t-1}^2 + ... + \alpha_m a_{t-m}^2, \tag{4.2}$$

where ϵ_t is white noise and $\epsilon_t \sim N(0, 1)$ under the assumptions of the independent and identically distribution (iid), $\omega > 0$ and $\alpha_i \geqslant 0$ for $i > 0$.

4.2 Generalized Autoregressive Conditional Heteroskedasticity (GARCH) Model

There are various criteria that must be followed in the ARCH model such that the model's prediction of the volatility is accurate. For that reason, it is possible to consider and recommend about transforming the ARCH model to create a generalized ARCH model. To explain the GARCH model, we begin with the continuously compounded log return series r_t. Here, let the novelty at time t (a_t) be denoted by $a_t = r_t - \mu_t$ [40]. At this point, we are able to edit e a_t in a GARCH (m, s) model by

$$a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \omega + \sum_{i=1}^{s} \alpha_i a_{t-i}^2 + \sum_{j=1}^{m} \beta_j \sigma_{t-j}^2, \tag{4.3}$$

where ϵ_t is iid and ϵ_t $N(0, 1)$, a_t is the model's residual at time t, σ_t is conditional standard deviation (volatility) at time t; m is the order of the ARCH component model and s represents the order of the GARCH component model, ω and $\alpha_1, ..., \alpha_m$ are the parameters of the ARCH component model and $\beta_1, ..., \beta_s$ are the parameters of the GARCH component model. Furthermore, $\omega > 0$, $\alpha_i \geqslant 0$, $\beta_j \geqslant 0$ and $\sum_{i=1}^{max(m,s)}(\alpha_i + \beta_i) < 1$ [40].

The limitation on the ARCH and GARCH parameters (α_i, β_i) is that the conditional standard deviation increases (σ_t) and the volatility (a_t) is finite. The GARCH parameter (β_i) disappears and the remaining portion transforms into an ARCH (m) model if $s = 0$ [40].

Equation (4.3) can be also expressed as

$$a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \omega + \alpha a_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{4.4}$$

in which $0 \leqslant \alpha, \beta \leqslant 1$, $(\alpha + \beta) < 1$.
According to this model, large values for a_{t-1}^2 and σ_{t-1}^2 likely to produce large values for σ_t^2. When clusters of volatility occur, this is true [40].

4.3 Exponential GARCH (EGARCH) Model

Even though GARCH is an improved version of ARCH, managing financial time series may still provide some challenges for the GARCH model. That's why, a new model, called exponential GARCH (EGARCH), is recommended. The suggested change is the inclusion of a measured creation to the model, which can account for the inequalities in the asset's return volatility [40].
Thereby, let a_t still be the novelty of the return at time t. The EGARCH (m, s) model can then be expressed as

$$a_t = \sigma_t \epsilon_t, \quad \ln(\sigma_t^2) = \omega + \sum_{i=1}^{s} \alpha_i \frac{|a_{t-i}| + \theta_i a_{t-i}}{\sigma_{t-i}} + \sum_{j=1}^{m} \beta_j \ln(\sigma_{t-1}^2). \tag{4.5}$$

In Equation (4.5), a_t is the model's residual at time t, σ_t is conditional standard deviation (volatility) at time t; s represents the order of the ARCH component model and m defines the order of the GARCH component model. Moreover, ω and $\alpha_1, ..., \alpha_s$ are the parameters of the ARCH constituent algorithm and $\beta_1, ..., \beta_m$ are the GARCH constituent model's parameters.
Thus, EGARCH $(1, 1)$ is represented as

$$a_t = \sigma_t \epsilon_t, \quad \ln(\sigma_t^2) = \omega + \alpha(|a_{t-1}| - E(|a_{t-1}|)) + \theta a_{t-1} + \beta \ln(\sigma_{t-1}^2), \tag{4.6}$$

where ϵ_t and $|a_{t-1}| - E(|a_{t-1}|)$ are iid and have mean zero. when the error term's distribution in the EGARCH is Gaussian, then $E(|\epsilon_t|) = \sqrt{2/\pi}$ which gives

$$\ln(\sigma_t^2) = \omega + \alpha(|a_{t-1}| - \sqrt{2/\pi}) + \theta a_{t-1} + \beta \ln(\sigma_{t-1}^2). \tag{4.7}$$

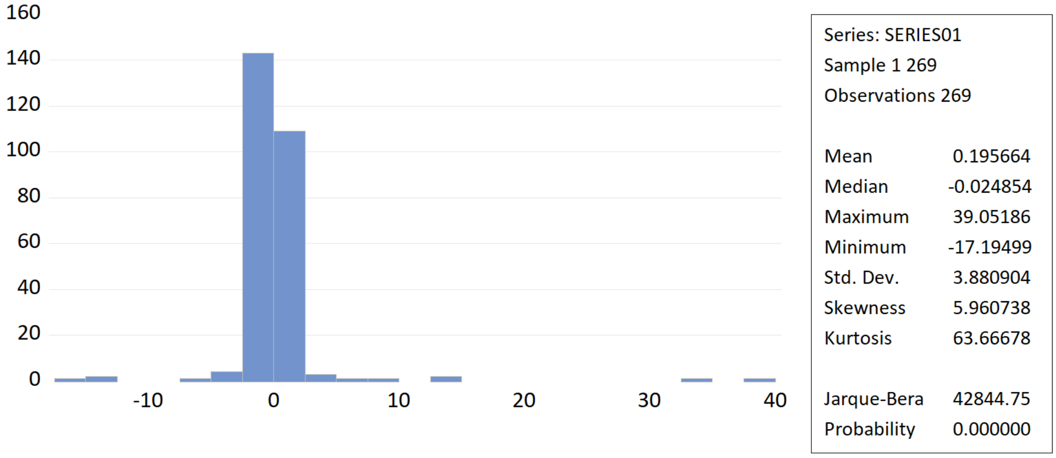
There is single quality that has to be emphasized. Here, the adverse volatility shocks typically have a larger effect and thus, θ is frequently considered to be negative. Indeed as the model employs logarithms, it gives rise to challenges when attempting to estimate an unbiased forecast [40].

Hereby, in our study, as previously applied in the thesis of Kalaycı [47], the stochastic model is converted to a MARS-similar mathematical model. By using this finding, we consider to check whether these volatility models can be written similar to a MARS structure or as a model within the generalized additive model. Alternatively, we also think to detect whether MARS can be written as a similar structure of the ARCH/-GARCH model if the error are at least defined as normal distribution. By this way, we can investigate whether MARS based model in time series analyses can be informative in behavioral finance datasets.

4.4 Application of Volatility Models

From the beginning of the thesis, we make the application of several different methods to observe the sentiment levels of investors. First, we apply machine learning techniques (MARS, RF, NN) both by only themselves and as two-stage. Now, we apply volatility models for each of the variable that we have studied in this thesis. In the analyses via volatility models, we begin with Sentiment Index. Here, we use 228 measurements. At first stage, we convert this dataset to return series by taking its first difference, and check the descriptive statistics of this transformed data. Then, in this updated dataset, we control the outliers by using their scatter plot. In the analyses we empirically accept ± 4 standard deviation as the indication of the outliers since the kurtosis and skewness of the data show a heavily-tailed distribution [75] . Therefore, we assign the 31st, 56th, 66th, the 73rd, the 177th, the 180th, the 186th, the 219th, the 237th and the 249th observations as extreme values. Hence, we adjust positive shock via +1 (referring to the 31st, 56th, 73rd, 180th, 219th, 237th observations) and negative shock (referring to the 66th, 104th, 186th and 249th observations) via -1 values. Afterwards, we check for the stationary case, and then, we determine the lag by checking AR (Autoregressive) and MA (Moving Average) models. Since, we prefer less lags, we take the model for the AR(1) and MA(1), and then, we compute for the

ARCH, GARCH and EGARCH models. Finally, we obtain the results for ARCH=1, ARCH=2 GARCH=1 and EGARCH=1 based on the AIC (Akaike information criterion) and BIC (Bayesian information criterion) model selection criteria as listed in Table 4.1. All the statistical results, graphs and tables are indicated comprehensive in the Appendix part C.1.1.



Series: SERIES01
Sample 1 269
Observations 269

Mean	0.195664
Median	-0.024854
Maximum	39.05186
Minimum	-17.19499
Std. Dev.	3.880904
Skewness	5.960738
Kurtosis	63.66678
Jarque-Bera	42844.75
Probability	0.000000

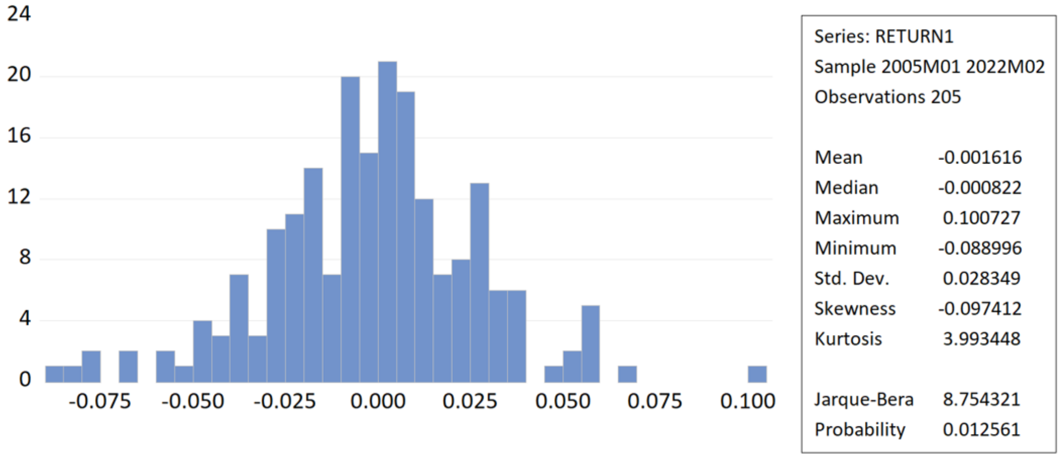
Figure 4.1: The descriptive statistics and histogram of the return series of sentiment index.

Table 4.1: The performance of model selection criteria for volatility models with AR(1) and MA(1) model by using sentiment index

Fitted Volatility Model	AIC	BIC
ARCH(1)	1.9744	2.0414
GARCH(1,1)	1.97408	2.0548
EGARCH(1,1)	1.8787	1.9725
EGARCH(1,2)	4.4958	4.6264

In the second stage of the analyses, we apply these volatility models to consumer confidence index. In these data, we have 206 observations. In the analyses, similar to sentiment index data, we initially convert the series to return series by taking its first difference, and check then for the descriptive statistics as shown in Figure 4.2. Furthermore, information for the explanatory data variables are also showed in Appendix C.1.2. Since CCI is explained by the TURKSTAT three weeks later than other macroeconomic variables that we used in our model, instead of taking CCI variable with 1-lag we take CCI itself, the reason of this is explained in the Appendix C.2.2.1. On the other hand, since TURKSTAT is changed the sub-indexes of CCI, in case of any biasness situation, model is checked, the results are shown in Appendix C.2.2.2.

In this updated dataset, we control the outliers by using their scatter plot.



```
Series: RETURN1
Sample 2005M01 2022M02
Observations 205

Mean          -0.001616
Median        -0.000822
Maximum        0.100727
Minimum       -0.088996
Std. Dev.      0.028349
Skewness      -0.097412
Kurtosis       3.993448

Jarque-Bera    8.754321
Probability    0.012561
```

Figure 4.2: The descriptive statistics and histogram of the return series of consumer confidence index.

Accordingly, by taking ±3 standard deviation as the indication of the outliers similar to the previous analysis, we assign the 31st and the 165th observations as outliers. Then, we replace positive shock by $+1$ (131st) and negative shock (165th) by -1 values. Afterwards, we check for the stationary case, and determine AR(2) and MA(1) as the best fitted modal for the presentation of the variance in the selected volatility model. For the computation, we find ARCH=1 and EGARCH=2 as the optimal model based on AIC and BIC model selection criteria Table 4.2. All the statistical graphs, analysis and results are showed detailly in the Appendix part C.1.2. In addition to these descriptive statistics of sentiment index and consumer confidence index, other variables which are unemployment index, consumer price index and usd/try currency index, are analysed individually in the Appendix part C.1.3 with all the details of graphs and tables.

Table 4.2: The performance of model selection criteria for volatility models with AR(1) and MA(2) model by using consumer confidence index.

Fitted Volatility Model	AIC	BIC
ARCH(1)	4.5716	6.4327
EGARCH(1,2)	4.496	4.626

4.4.1 Application of Volatility Models with Multiple Input Variables

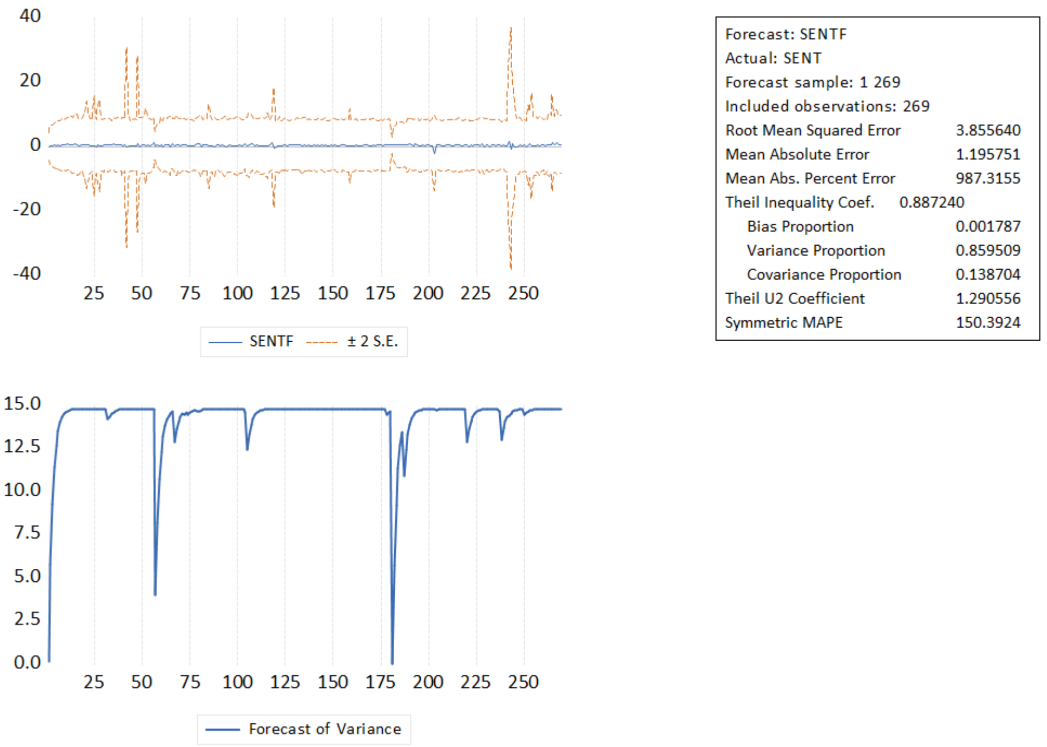
For the application of volatility models with the multiple input variables, at first, we apply GARCH models by taking sentiment index as an output, and other variables (NYSE share turnover, the closed-end fund discount, the number, and average first-day returns on IPOs (initial public offerings), the dividend premium, the equity which shares new issues, indpro (industrial production index), consserv (nominal services consumption), consdur (nominal durables consumption), consnon (nominal nondurables consumption), cpi (consumer price index), and employ (employment)) as an inputs. We obtain results for ARCH=1 and EGARCH=1. All of the statistical results, tables, equations and graphs are demostrated in the Appendix part C.2.1. Furthermore, among our 270 variables, we forecast the last 23 variables.

Table 4.3: The performance of model selection criteria for volatility models with AR(1) and MA(1) model by using sentiment index with the multiple input variables.

	Under Normal		Under Student-t	
Fitted Volatility Model	AIC	BIC	AIC	BIC
ARCH(1)	4.0000	4.1871	2.8953	2.2958
EGARCH(1,1)	3.2831	3.4969	2.1512	2.3785

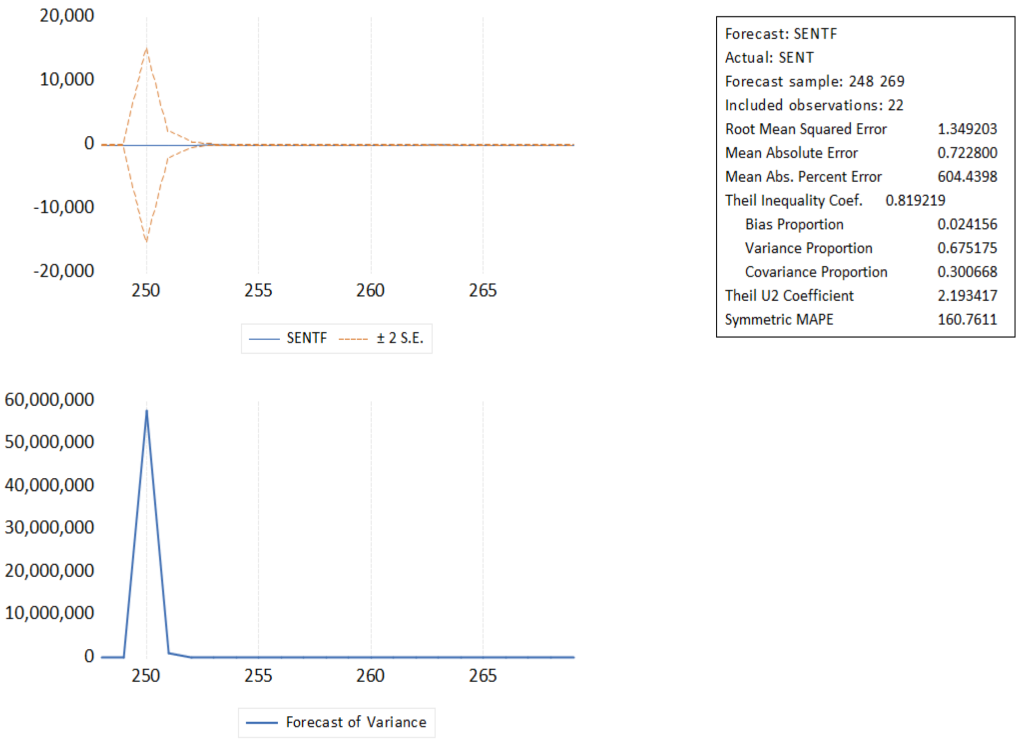
At the first graph which is shown in Figure 4.3, we forecast sample for full data, and at the second graph, we have forecast results for modified sample. Here, return on sentiment is generally stable. Whereas, from the beginning of 2020 and until the end of 2021, volatility level seems increased, which might thought stem from the Covid-19 Pandemic effect. The effect of Covid-19 was strongly felt by worldwide especially in terms of economical and financial fields. The global stock markets and economy experienced a significant surge in volatility when the COVID-19 pandemic emerged in February 2020 [26]. International investors taking risky positions in the financial markets lead to a high number of financial transactions that result in an unprecedented level of instability in the prices of financial assets. The pandemic-induced multiple crashes and significant fluctuations in financial returns are negatively affecting global investments. These unanticipated crashes and fluctuations are posing a significant challenge to financial investors worldwide [51].

As a second case, we apply GARCH models by taking CCI as an output, and other variables (UN, CPI, USDTRY) as an inputs. We obtain results for ARCH=1. Among



```
Forecast: SENTF
Actual: SENT
Forecast sample: 1 269
Included observations: 269
Root Mean Squared Error          3.855640
Mean Absolute Error              1.195751
Mean Abs. Percent Error          987.3155
Theil Inequality Coef.     0.887240
     Bias Proportion             0.001787
     Variance Proportion         0.859509
     Covariance Proportion       0.138704
Theil U2 Coefficient             1.290556
Symmetric MAPE                   150.3924
```

Figure 4.3: The forecasting results of the return series for entire sentiment index data.



```
Forecast: SENTF
Actual: SENT
Forecast sample: 248 269
Included observations: 22
Root Mean Squared Error          1.349203
Mean Absolute Error              0.722800
Mean Abs. Percent Error          604.4398
Theil Inequality Coef.     0.819219
     Bias Proportion             0.024156
     Variance Proportion         0.675175
     Covariance Proportion       0.300668
Theil U2 Coefficient             2.193417
Symmetric MAPE                   160.7611
```

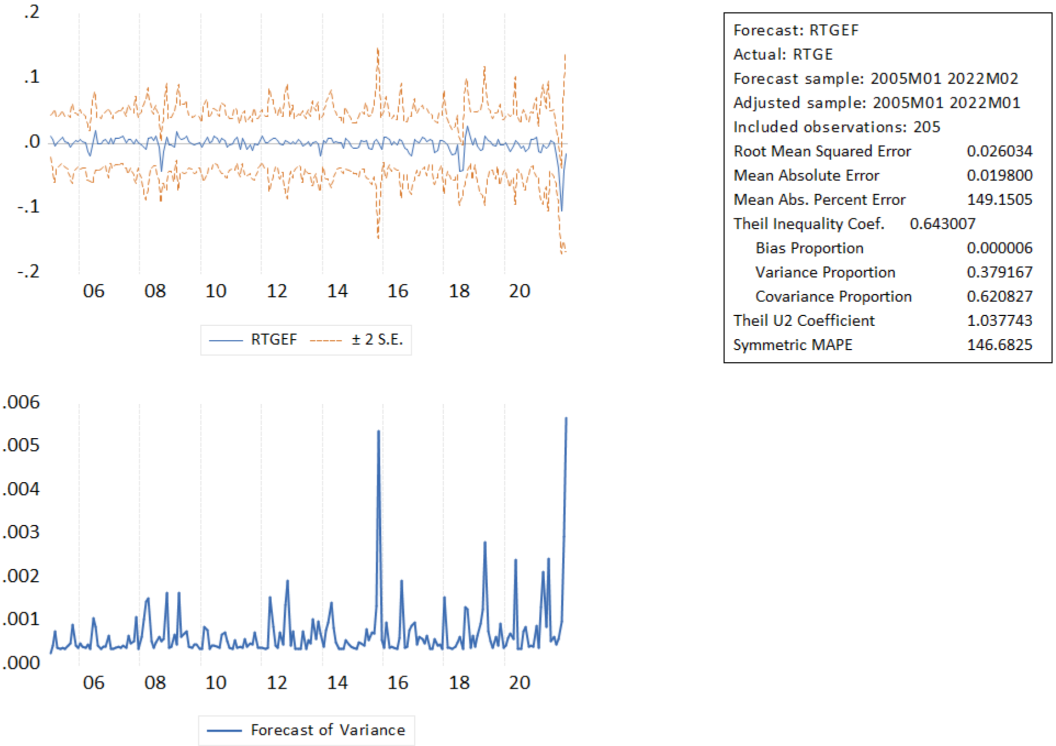
Figure 4.4: The forecasting result of the return series of sentiment index for the last 20 variables.

Table 4.4: The performance of model selection criteria for volatility models with AR(1) and MA(1) model by using consumer confidence index with the multiple input variables.

	Under Normal		Under Student-t	
Fitted Volatility Model	AIC	BIC	AIC	BIC
ARCH(1)	-4.5372	-4.4399	-4.5275	-4.4140

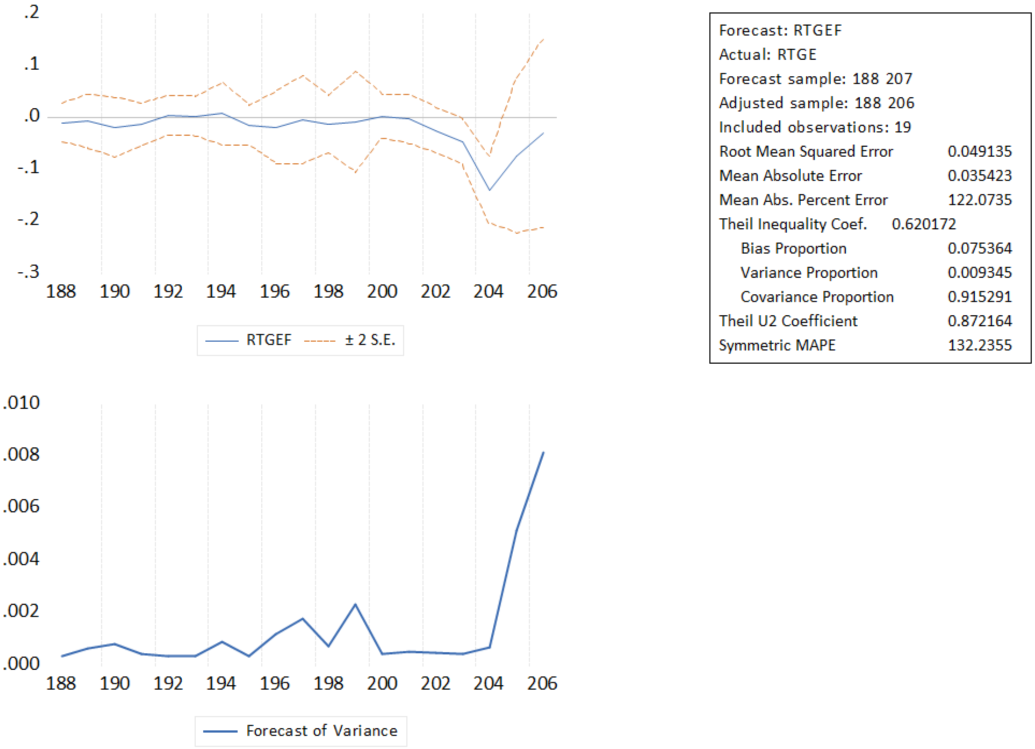
our 206 variables, we forecast for the last 20 variables, as previously implemented.

In Figure 4.6, the first graph shows us forecast sample for full data and at the second graph, we forecast results for modified sample. All of the information about the equation, graphs and statistical results are analysed in the Appendix part C.2.2. This time, return on CCI is observed as volatile. Accordingly, from the beginning of 2018, where the currency crises occurred, it still continues due to the decisions about the macro policy, which makes most of the macroecenomic variables uncertain and resulting in the volatile level of consumers confidence level to increase.



```
Forecast: RTGEF
Actual: RTGE
Forecast sample: 2005M01 2022M02
Adjusted sample: 2005M01 2022M01
Included observations: 205
Root Mean Squared Error        0.026034
Mean Absolute Error            0.019800
Mean Abs. Percent Error        149.1505
Theil Inequality Coef.    0.643007
    Bias Proportion            0.000006
    Variance Proportion        0.379167
    Covariance Proportion      0.620827
Theil U2 Coefficient           1.037743
Symmetric MAPE                 146.6825
```

Figure 4.5: The forecasting results of the return series for entire consumer confidence index data.



```
Forecast: RTGEF
Actual: RTGE
Forecast sample: 188 207
Adjusted sample: 188 206
Included observations: 19
Root Mean Squared Error        0.049135
Mean Absolute Error            0.035423
Mean Abs. Percent Error        122.0735
Theil Inequality Coef.    0.620172
     Bias Proportion           0.075364
     Variance Proportion       0.009345
     Covariance Proportion     0.915291
Theil U2 Coefficient           0.872164
Symmetric MAPE                 132.2355
```

Figure 4.6: The forecasting result of the return series of consumer confidence index for the last 20 variables.

CHAPTER 5

MARKOV SWITCHING MODEL AND MS-GARCH MODEL

5.1 Markov Switching Model

Numerous economic and financial time series seem to go through phases where their behavior quickly changes from what was previously seen. The average value of a series will determine how it behaves over time, volatility, or how closely its present value resembles its past value [27]. The modification of the behavior might happen permanently, generally called as a ‘structural break’ in a series. It might alter as well for a while before bringing back to its previous habit or adopting yet another pattern of conduct, and the second case is mostly named as ‘regime shift’ or ‘regime switch’ [27].

Numerous economic time series occasionally exhibit dynamic interruptions in their behavior, which are related to occurrences like financial crises or changes in governmental policy [42]. From the side of economists, during the economy is struggling, when insufficient use of economic dynamics are dominated by production factors rather than their propensity to grow through time, it is appealing to rely on a variety of economic variables to function substantially differently [42]. To indicate how it might be described the outcomes of a effective alteration in the behavior of a single variable, y_t is defined. If we suppose that the general historical behavior could be defined with a first-order autoregression [42], we obtain the following equation:

$$y_t = c_1 + \phi y_{t-1} + \varepsilon_t \tag{5.1}$$

with $\varepsilon_t \sim N(0, \sigma^2)$, which sufficiently defines the observed data for $t = 1, 2, ..., t_0$. In Equation (5.10), ϕ is autoregressive variable and c_1 is an intercept. At date t_0 there is a remarkable alteration in the series' standard range, for that reason it is preferred to define the data with regard to the subsequent equation:

$$y_t = c_2 + \phi y_{t-1} + \varepsilon_t \tag{5.2}$$

for $t = t_0 + 1, t_0 + 2, ...$, c_2 is an intercept. Fixing the intercept value from c_1 to c_2 could help the model improve its predictions. Nevertheless, it falls short as a probability theory that may have generated the data. That the transition from c_1 to c_2 at date t_0 was a deterministic occurrence that anyone could have predicted with precision by looking beyond from day $t = 1$ is definitely not desirable [42]. Alternatively, there must have been some wrongfully presumable forces that generated the alteration. Thus, it should be taken into account that there is a larger model around them both rather just asserting that expression execute the data up to date t_0 and later than that date [42] via

$$y_t = c_{s_t} + \phi y_{t-1} + \varepsilon_t, \tag{5.3}$$

In Equation (5.3), s_t is a random variable that, in consequence of organizational alterations, occurs in this sample to suppose the value $s_t = 1$ for $t = 1, 2, ..., t_0$ and $s_t = 2$ for $t = t_0 + 1, t_0 + 2, ...$ An absolute representation of the probability theory running a model based on probability would therefore be required for the observed data of what brought about the alteration from $s_t = 1$ to $s_t = 2$. Here, the probabilistic model which cause and rules these movements is a Markov process [21].
Accordingly, s_t is the eventuation of a two-state Markov chain with [42]

$$P(s_t = j|s_{t-1} = i, s_{t-2} = k, ..., y_{t-1}, y_{t-2}, ...) = P(s_t = j|s_{t-1} = i) = p_{ij}, \tag{5.4}$$

here p_{ij} denotes the transition likelihood of from state i at time $t - 1$ to state j at time t.

It is considered that s_t is not immediately observed but rather only serves to suggest its activity through the behavior of y_t, the two state transition probabilities, p_{11} and p_{22}, along with the variance of the Gaussian innovation σ^2, the autoregressive coefficient ϕ, the two intercepts c_1 and c_2, and the necessary parameters to completely express the probability law running y_t [42].

The calculation in Equation (5.4) implies that the value of the most recent regime is the only factor that influences the chance of a regime transition and not on the past.

However, nothing in the method defined below rules out examining more common probabilistic determinations. However, the straightforward the most appropriate initial point seems to be Markov chain Equation (5.4), and it is best to act as though the transition from c_1 to c_2 was a deterministic outcome [42].

There are lots of non-linear models in the econometrics, however, in literature, few kinds of model have had remarkable effect in finance and Markov regime switching model is one of them [21]. One of the primary reasons why researchers have shown significant interest in regime-switching models is due to their capacity to effortlessly capture the various modes of the financial market [83, 84].

The switching mechanism in the Markov regime switching model differs from other switching models in that it is regulated by an unobservable variable arising from a *Hidden Markov Model* (HMM). In addition, financial time series have several structured facts that can be effectively recreated by a HMM. As a result, the Markov regime switching model has become one of the most widely used nonlinear time series models in the literature. For that reason, we would rather analyze this model [19].

5.2 The Markov Switching GARCH (MS-GARCH) Model

There are numerous explanations for why financial series show significant behavioral breakdowns; depression, recession, bankruptcies, market panics, as well as changes in government policies or investor expectations from regime change [15]. Each regime has a different volatility structure. From different fields of researchers and practitioners, the GARCH model is carried out widely [3] in such a way that if the series exhibit structural breaks, standard GARCH models can generate biased outcomes [22]. At this point, more appropriate model should be considered. For that reason, since each regime has a different volatility structure which means that each state of the chain regime enables a different GARCH behavior and also to prevent biasness, by uniting GARCH models with a Markov switching chain, widens the dynamic formulation of the model and within possibility permits advanced forecasts of the volatility [3, 22]. In such situations there is a model called *Markov-Switching GARCH* (MS-GARCH) models, where a distinct latent variable can cause parameters to change over time

[22]. Thereby, to estimate a model that supports parameter regime switching, the MS-GARCH model is used. The conditional variance of each regime can have a varied persistence in this expansion of the GARCH model [3, 13, 15, 22].

In literature, firstly, Cao et al. (1994) studied regime switching framework for modelling volatility estimated ARCH specifications. An ARCH model's conditional variance solely takes into account previous observations. Here, when it is compared to GARCH model according to computational tractability side, integration over all K^N routes (K denotes the number of regimes, N is the number of observations) is required for the assessment of the probability function for the MS-GARCH model in in specific, making this computation complicated. To cope with this problem Gray (1996), Dueker (1997) and the hoc approximation method employed by Klaassen (2002) relies on eliminating conditional variances from previous regimes [3, 22]. Haas et al. (2004) presented new MS-GARCH model which aim to solve some difficulties. One of the difficulties is about computation and other problem is about the understanding of dynamic features. The findings propose that up and coming volatility model is an unincorporated switching GARCH process with a probably conditional mixing density that is skewed [41].

For stationarity analysis of MS-GARCH processes, in order to create a comprehensive strategy, Abramson et al. (2007) used finite state-space Markov chains to monitor the transition between each regime's active GARCH model of order (p, q) [3].

Sajjad et al. (2008) proposed an asymmetric MS-GARCH model to compute Value-at-Risk (VaR) for both short and long positions. The purpose of their model is to enhance existing VaR methods by considering not only regime change but also skewness or leverage effects. The result of this study shows that MS-GARCH specifications obviously transcend other models in approximating the VaR for both long and short positions [82].

Bauwens et al. (2010) studied MS-GARCH model wherein the conditional mean and variance change in time from one GARCH process to another. The switching is ruled by a hidden Markov chain [13].

Augustyniak (2014) enhanced an approach which is based on Monte Carlo Expectation–Maximization algorithm and significance sampling to compute the maximum likelihood estimator and asymptotic variance–covariance matrix of the MS-GARCH model [9].

Billio et al. (2016) presented a new produced Metropolis algorithms built on the integration of multi-pronged approaches by designing influential sample methods for GARCH models with Markov switching under Bayesian inference [16].

Ardia et al. (2018) aimed to compare the forecasting abilities of single-regime and MS models from the perspective of risk management. For daily, weekly, and ten-day equity log-returns, they were able to obtain more accurate Value-at-Risk anticipated shortfall and left-tail distribution predictions than their single-regime counterparts [7].

The most suitable model or combination of models for modeling the volatility of the four most well-known cryptocurrencies was selected by Caporale et al. (2019). Each of these cryptocurrencies on computed a one-step later prediction of VaR and Expected Shortfall (ES) based on a rotating window. The result of their study showed that employing standard GARCH models might lead to inaccurate VaR and ES estimations, and thus concluded affectless risk-management, portfolio optimization. Furthermore, it was found that two-regime GARCH models generated superior VaR and ES estimates than single-regime approaches [22].

Wang et al. (2022) carried out the development of estimated power of renewable energy stock volatility by improving MSGARCH-MIDAS (Mis Data Sampling) models long-terms and short-terms. The models that allow for regime-switching in the both short- and long-volatility parts simultaneously outperform other competing models for short-term prediction by using several out-of-sample tests. But, as a result, at longer horizon Markov regime-switching performs a greater influence on forecasting accuracy [94]. In our study, to see at which points our model is subject to any regime change, we implement the MS model to our macroeconomic variables.

The number of states (or regimes) which is N, depending on the present state of the

HMM:

$$Y_t = \mu_1 + \varepsilon_t \qquad for\ state\ 1, \tag{5.5}$$

$$Y_t = \mu_2 + \varepsilon_t \qquad for\ state\ 2, \tag{5.6}$$

$$. \tag{5.7}$$

$$. \tag{5.8}$$

$$. \tag{5.9}$$

$$Y_t = \mu_N + \varepsilon_t \qquad for\ state\ N, \tag{5.10}$$

with $\varepsilon_t \sim N(0, {\sigma_1}^2)$ for state 1, $\epsilon_t \sim N(0, {\sigma_2}^2)$ for state 2,...,$\epsilon_t \sim N(0, {\sigma_1}^N)$ for state N [19].

Because the underlying Markov chain is hidden, one is unable to directly see what state the HMM is in, but must instead derive its operation from the observed behavior of Y_t. A probabilistic model of what causes the change from state $S_t = i$ to state $S_t = j$ is necessary to achieve the probability law regulating the observed data Y_t. The transition probabilities of a N state HMM can be utilized to determine this [19, 42].

$$p_{ij} = P(S_t = j|S_{t-1} = i) \quad (i, j \in \omega = 1, 2, ..., N). \tag{5.11}$$

The transition probability (5.11) is only dependent on the past through the value of the most current state, according to the Markov property mentioned in (5.4). This is a key feature of the structure of a Markov regime switching model, as switching the states of the fundamental HMM is a stochastic process within itself. At this point, we determine the states by employing HMM. We apply MS model to our Consumer Confidence Index (CCI) data. This index is assigned a value between 0 and 200. Consumers who score above 100 are optimistic, while those who score below 100 are pessimistic. In this case, we had need to determine two states; optimistic or pessimistic, however, from the psycological side, people are generally seperated according to their motivational state as low motivated, medium motivated or high motivated [25], or from the market side, states are generally set as bear market, bull market and mixed market [24]. From this point of view, apart from optimisim or pessimism level of consumers, we also want to set another state as an 'neutral' because of the consumers who does not participate the poll or who feels impartial. As a result, we determine three states: optimistic, pessimistic and neutral. Here, we could arrange the states not only with three states but much more states such as very optimistic, moderately optimistic, moderately low optimistic, low optimistic, realistic, low pessimistic, moderately low pessimistic, etc. We plan to analyse and research states of consumers with this kind of detailed information in the future. All explanations, construction of states with HMM, initial, tranmission and emission probabilities of HMM with *Expectation-Maximization(EM) Algorithm*, *Viterbi Algorithm* and *Baum-Welch Algorithm* are computed and showed in the Appendix D.

As a results, we obtain three regimes, their coefficient results and statistical results. Afterwards, since each regime contains different volatility nature, we apply GARCH models into each of these regimes to interpret these various volatility behaviors. By this way, we aim to learn whether using only the GARCH model or by MS-GARCH model is more accurate and preferable. In the next part, we present the outcomes of our application.

5.2.1 Application of MS-GARCH Model

Application for the MS-GARCH model is done by the same datasets as we used in previous parts. Hereby, the worked economic variables are, the Consumer Confidence Index (CCI), Unemployment Index, Consumer Price Index and USD/TRY Index. All these datasets are monthly, starting from 2005 to up to second month of 2022. For that reason, we have 206 variables in total.
Firstly, a linear model is fitted to see how the covariate input variables explains the variable response in CCI.

Table 5.1: Residuals obtained with MS model

Min	First Quantile	Median	Third Quantile	Maximum
-16.4722	-1.8558	0.5912	2.8481	9.4005

Table 5.2: Coefficients of variables obtained with MS model

Variables	Estimate	Standard Error	t value	$Pr(>\|t\|)$
(Intercept)	106.9733	2.0829	51.357	$< 2e - 16$
UN	-1.4791	0.1988	-7.440	2.83e-12
CPI	0.01609	0.01186	1.357	0.1762
USD/TRY INDEX	-0.029	0.0086	-3.347	0.0009

From the findings it is seen that the covariate is really significant, but, the data behavior is not sufficiently explained by the model.

Table 5.3: Statistical results of LM

Criteria	Value
RSS	4.578
Multiple R-squared	0.5676
Adjusted R-squared	0.5612
F-statistic	88.390
p-value	$< 2.2e - 16$

Table 5.4: Residuals

AIC	BIC	logLik
971.6293	1075.498	-473.8146

Table 5.5: Coefficients of Regime 1

Regime 1	Estimate	Standard Error	t value	$Pr(>\|t\|)$
(Intercept)	101.5730	3.4492	29.4483	$< 2.2e - 16$
UN	-1.6340	0.1753	-9.3212	$< 2.2e - 16$
CPI	0.0864	0.0211	4.0948	4.225e-05
USD/TRY INDEX	-0.0847	0.0166	-5.1024	3.354e-07

Table 5.6: Statistical Results of Regime 1

Criteria	Value
RSS	1.6615
Multiple R-squared	0.928

Table 5.7: Coefficients of Regime 2

Regime 2	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	59.4937	5.3390	11.1432	$< 2.2e - 16$
UN	-0.3665	0.3071	-1.1934	0.2327
CPI	0.3135	0.0444	7.0608	1.656e-12
USD/TRY INDEX	-0.2482	0.0468	-5.3034	1.137e-07

Table 5.8: Statistical results of Regime 2

Criteria	Value
RSS	2.3412
Multiple R-squared	0.8377

According to the Table 5.6, 5.8, 5.10, it is seen that *RSS* value is the best for the Regime 1, then Regime 3 and Regime 2. In line with these results *Multiple R-squared* has the highest value in Regime 1, then Regime 3 and Regime 2. On the other hand, according to the Table 5.5, 5.7 and 5.9, which shows the coefficients of each regimes, UN and USD/TRY INDEX has inverse relation, while CPI has linear relation with

Table 5.9: Coefficients of Regime 3

Regime 3	Estimate	Std. Error	t value	$Pr(> \|t\|)$
(Intercept)	100.5808	1.6414	61.2774	$< 2.2e - 16$
UN	-0.2537	0.1984	-1.2787	0.201
CPI	0.0081	0.0078	1.0385	0.299
USD/TRY INDEX	-0.0347	0.0054	-6.4259	1.311e-10

Table 5.10: Statistical Results of Regime 3

Criteria	Value
RSS	2.1233
Multiple R-squared	0.9038

Table 5.11: Transition probabilities

	to		
From	Regime 1	Regime 2	Regime 3
Regime 1	0.9036	0.0888	0.02808
Regime 2	0.0395	0.9063	0.0072
Regime 3	0.0569	0.0049	0.9647

CCI. From Table 5.11, it is seen that the probability of staying at regime 1, regime 2 and regime 3 are higher than transtioning from one of the regime at time $t - 1$ to another regime at time t or vice versa. In both, the R-squared have high values. Finally, the transition probabilities matrix has high values which indicate that is hard to shift from one regime to the other. The model specifies sufficiently the periods of each state. On the other hand, there are numerous factors that cause significant changes in financial trends, such as economic downturns, financial crises, government policy changes, and shifts in investor expectations due to political transitions. Each of these factors impacts the volatility of the market in a unique way. To analyze these changes and allow for varying volatility, we employ the Markov-switching GARCH (MS-GARCH) model. This model, which is a GARCH model furtherance, allows us to vary the persistence of each regime's conditional variance [22].

Here, Figure 5.3 shows that smoothed probabilities of the three regimes, which are indicated in Figure 5.1, are built on the original plotted graph of CCI. By this way, we can see and analyze which regime change occured at which point more clearly.

CCI starts from 2005 to the second month of 2022. In Figure 5.3, from the beginning of 2005 to beginning of 2008, which is shown as gray area, is belongs to Regime 1.

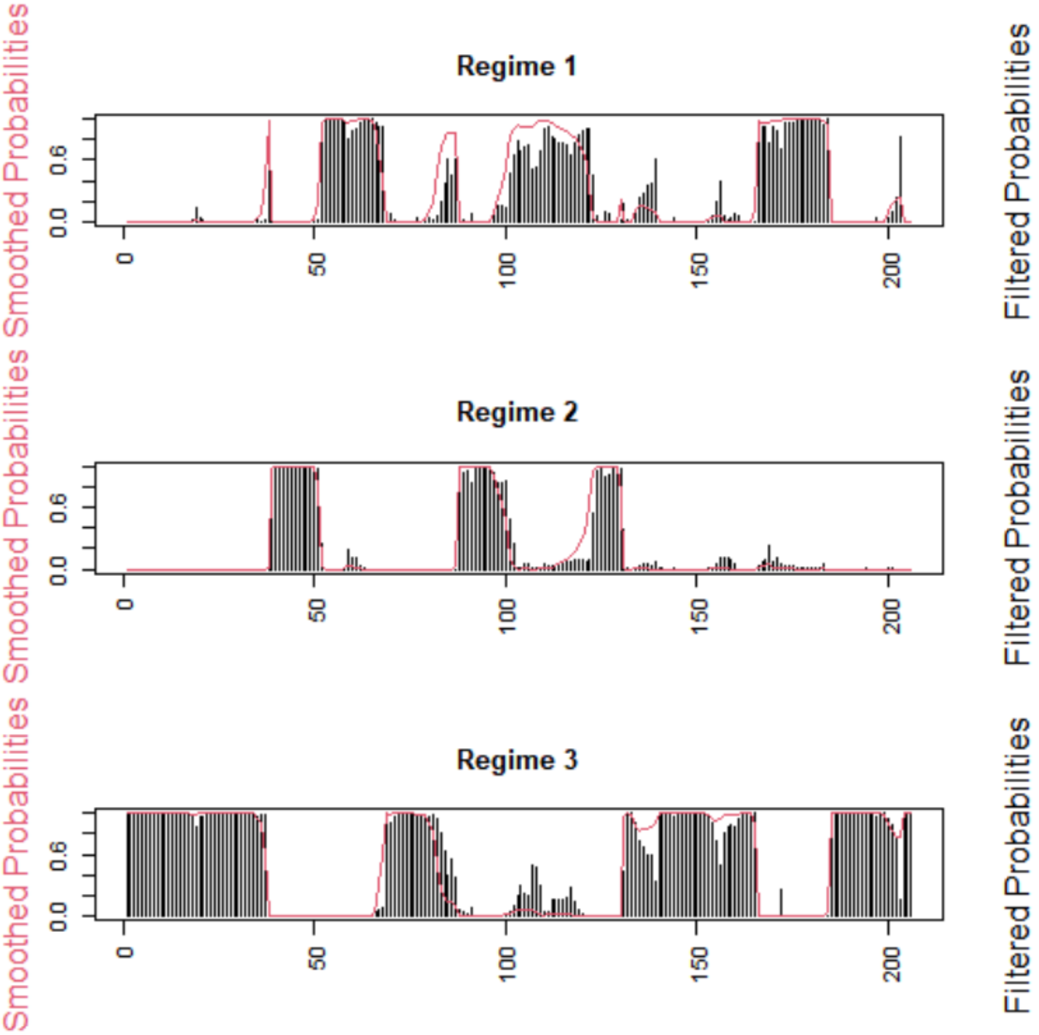


Figure 5.1: Smoothed probability of optimism and pessimism.

At this time interval, CCI seems moderately high. We can deduce from this perspective that there occurs no situation that affects consumer confidence. However, from the beginning of 2008 to beginning of 2009, which is shown as green area on a graph (Regime3), regime break occurs and there is a sharp decrease in CCI. Within that period there was a 2008 Global Financial Crisis which is also named as *Great Recession*. The onset of the 2008 financial crisis was triggered by the availability of easy credit and lenient lending policies which contributed to the formation of a housing bubble. As soon as the bubble eventually burst, the banks were left with vast amounts of valueless subprime mortgage investments. This led to the Great Recession, which resulted in numerous individuals losing their employment, their savings, and their homes. Specifically, the collapse of Lehman Brothers, recognized as the most significant bank failure in history, has severely damaged confidence in banks, financial markets, financial instruments, and rating agencies across the globe. As a result, this huge economical disaster lead consumer confidence levels to decrease. Although, this crisis is occured in US, because of the export-import relationship, all the developed and developing countries were affected indirectly. So, the banking and finance sector in Türkiye was not directly influenced by the 2008 Economic Crisis. But, because of the decrease in foreign capital inflows and tightining global foreign trade volume, economic growth was negatively get damaged. From this point of view, we can come to terms that all these financial damages has an affect on our CCI and we can say that CCI may decrease sharply because of these reasons.

Moreover, after the first months of 2009, we see recovery process until the mid-2010 which is indicated as yellow area (Regime 2). Furthermore, from the mid-2010 until the end of 2011, our regime again changes back to Regime 1 (gray area), since it seems better, higher and more optimistic level. Afterwards, this situation changes, first confidence moderately decreases (Regime 2) and then, sharply decreases (Regime 3). Then, again confidence level moderately increases until the end of 2014 (Regime 2), and affectingly decreases until the end of 2015 (Regime 3). This regime changes were brought about some kaotic incidents and demonstrations, which give rise to economical distress. In these process, there is a highly increase in unemployment rates and currency rates. Since there was widely uncertainy prevailed in the economy, consumers' confidence declined.

On the other hand, from the end of 2015 to tenth month of 2018, there is temperatively high confidence level (Regime 1). At this point, we need to emphasize that by saying moderately high in Regime 1, we do not guarantee that consumers' confidence levels are always high. It is obviously seen from the Figure 5.3 that between these intervals, there are some sharply decreased points, although it is in Regime 1. This shows us that each regime is not about being entirely optimisim or pessimism, instead, it shows us the general trend to stay in the same situation. From passing Regime 1 to Regime 2, we observe a sharp decline in the confidence level. This was because of the dolarization crisis which occured in August 2018. The reflection of this crisis is happened as an increase in inflation, unemployment and currency rate. All these are given rise in declining the confidence level. However, since it is recovered conservatively afterwards, the alteration is moved from Regime 1 to Regime 2 instead of Regime 3.

From the eleventh of 2018 to mid-2020, we see the confidence at intermediate level (Regime 2). This interval includes Covid-19 pandemic affect which caused most of the economies to be influenced enormously. However, this time interval is not only comprehended by the pandemic effect. Therefore, the confidence level of consumers' are balanced. Similarly, for the next step there was a recuperation, thus, regime is changed from 2 to 1.

To evaluate this different GARCH behaviors in these regimes, we incorporate GARCH models to each of our regime changes. We applied it to the raw data for the locations corresponding to the 3 regime paths.

Table 5.12: The performance of model selection criteria for volatility models of each regime.

		Under Normal		Under Student-t	
	Fitted Volatility Model	AIC	BIC	AIC	BIC
Regime 1	GARCH(1,1)	-4.134	-3.990	-4.153	-4.009
Regime 2	ARCH(1,0)	-4.4697	-4.4191	-4.4563	-4.3805
Regime 2	GARCH(1,1)	-	-	-4.4585	-4.3574
Regime 2	EGARCH(1,1)	-4.4833	-4.3822	-4.4771	-4.3507
Regime 3	GARCH(1,1)	-	-	-3.4684	-3.2835

Accordingly, for the first regime, there is a volatility effect, we obtained for GARCH (1,1) under both normal and student-t distribution. However, for Regime 2, there was a GARCH effect under only student-t distribution, because probability value is higher (0.306)than 0.05 for the normal distribution. But, EGARCH effect was found under both with normal and student-t distribution. Finally, Regime 3 had also GARCH effect under only student-t distribution, because, here again, probability value is higher (0.361) than 0.05 for the normal distribution.

Here, our purpose was to compare sole the GARCH model or MS-GARCH model,has better performances. By comparing their outcomes (AIC, BIC values) as represented in Table 5.12 and Table 4.2, it was defected that MS-GARCH model fitted better to the data.

As a result, since all regimes includes different volatility structure in it, by applying volatility models to each of our regimes (MS-GARCH), we improved the accuracy of our findings.

Figure 5.2: Smoothed probabilities of three Regimes built on plotted CCI.

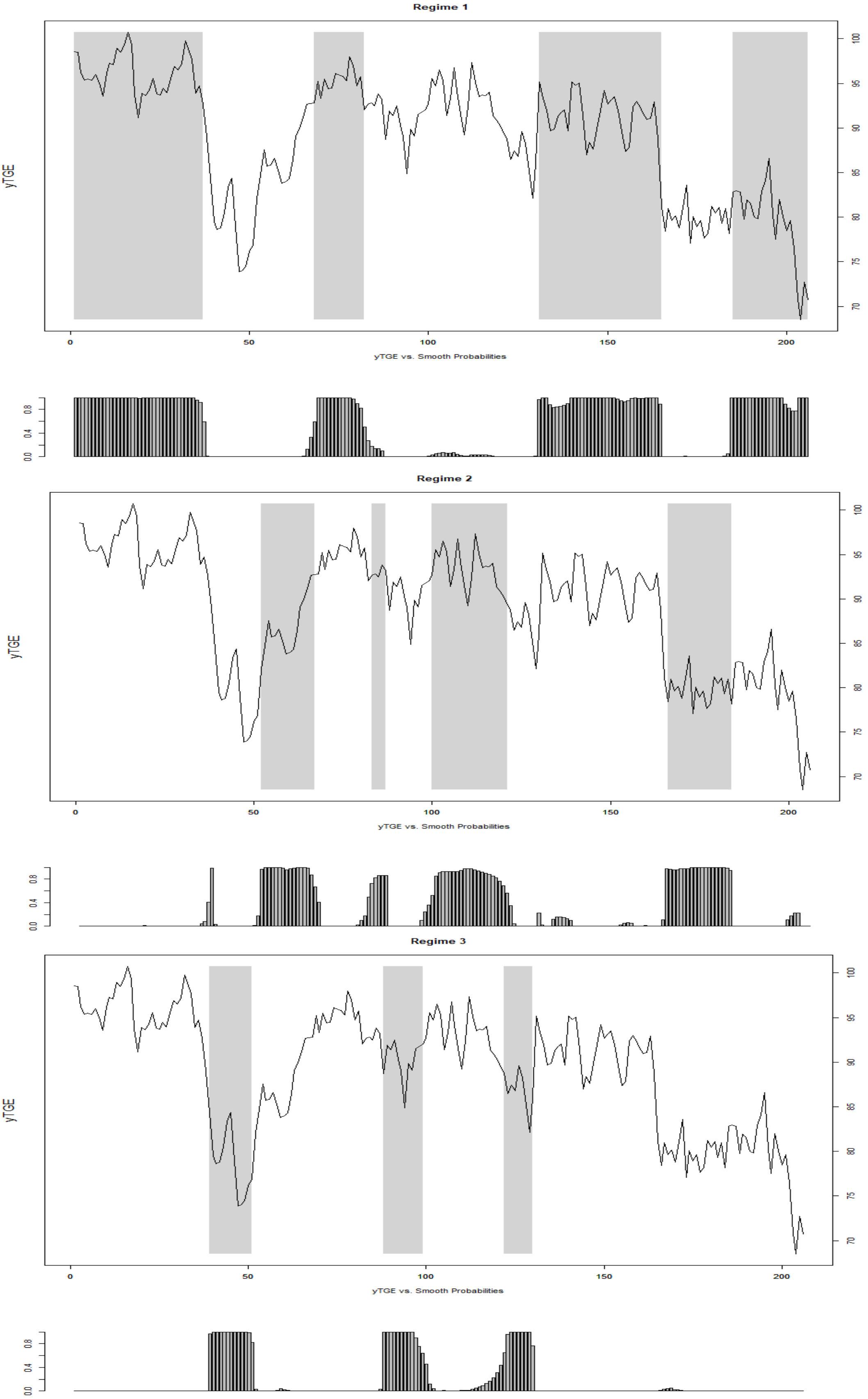
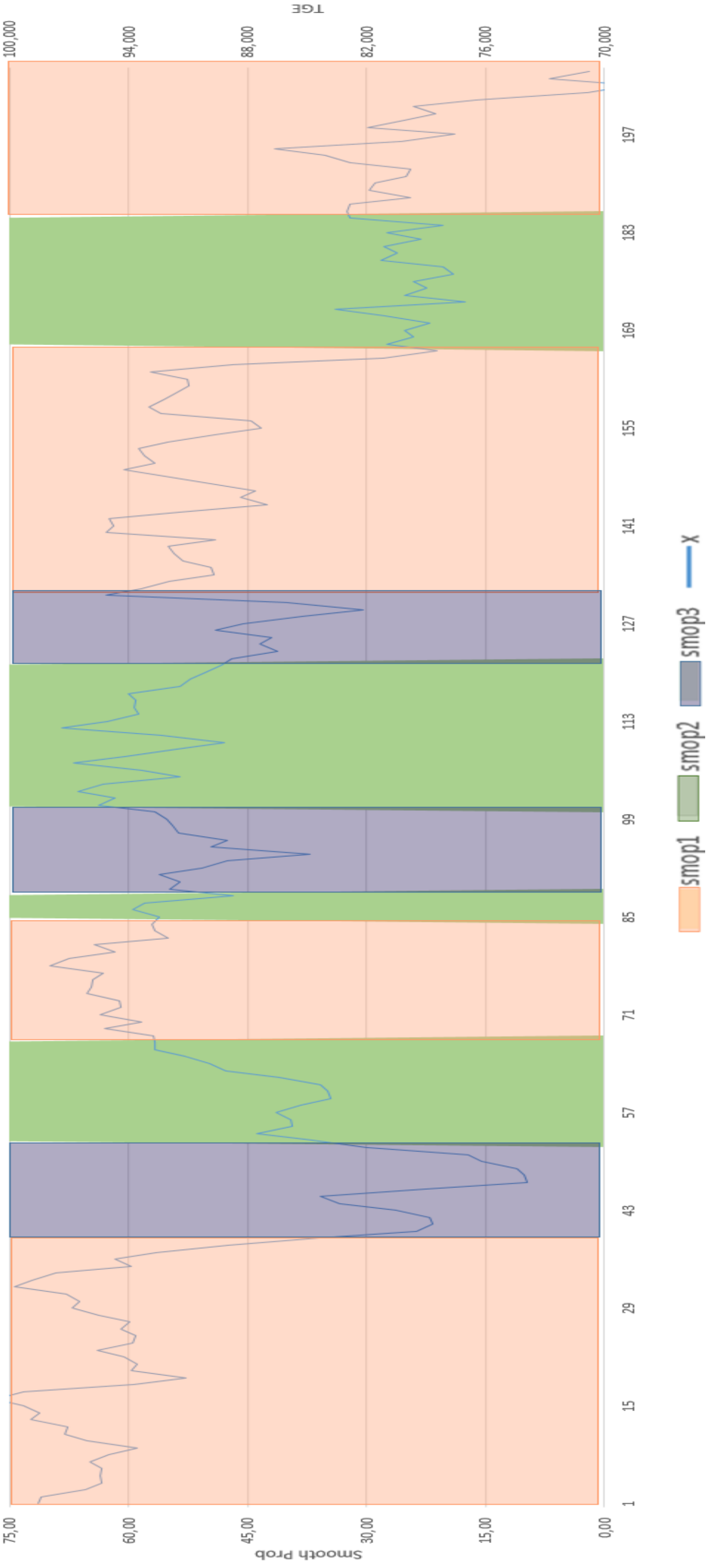


Figure 5.3: Smoothed probabilities of three Regimes built on plotted CCI.



CHAPTER 6

CONCLUSION

Various recent real financial applications have nonlinear and complex behaviors. Since classical statistical methods depend on some restrictive assumptions and applications, methods are required to deal with these stochastic problems [6, 10].

Machine Learning methods are well-known and beneficial tools for prediction problems and have already been successfully applied to numerous financial associated problems.

In this study, apart from pure financial related problems, we focus on behavior of investors introduced as Sentiment. The goal of this study is to compare the forecasting ability of sentiment index by using single MARS, RF, NN models, and two-stage MARS-NN, MARS-RF, RF-MARS, RF-NN, NN-MARS, and NN-RF hybrid models. Results show that MARS single model, RF-MARS and NN-MARS two-stage models outperform better in this kind of sentiment data. On the other hand, we consider to work with another behavioral data that also shows the investors' sentiment levels. The principles of sentiment are presented by economic conditions such that most of the divergence in consumer sentiment stems from either explicitly or implicitly from economic circumstances. Investors' economic and political outlooks are influenced by how optimistic or pessimistic they are. When the current situation appears favorable, people are more upbeat about the economy both now and in the future. Particularly, when inflation rates rise or unemployment increases, economic assessments become more pessimistic. The general anticipations of the public about the future of the economy rise when major economic indicators indicate favorable times will soon arrive, too [31]. The consumer confidence index measures how optimistic consumers are feeling about the state of the economy based on their savings and expenditure habits, which contribute to national economic expansion.

From this perpective, we aimed to see the effect of economic variables to consumer confidence levels. We indicated the applicability of three machine learning techniques by using economic variables: The Consumer Confidence Index (CCI), Unemployment Index, Consumer Price Index, USD/TRY Index. All these datasets are monthly, starting from 2005 to up to second month of 2022. The CCI, Unemployment Index and CPI were taken from Turkish Statistical Institute (TURKSTAT) and USD/TRY exchange rates were collected from the Central Bank of Turkey (TCMB, EVDS Data Central).

The results of the application showed that MARS model itself outperformed better comparing with the other two-stage models. On the other hand, two-stage model, which is MARS-RF, had also better results with respect to the single RF model and the NN-RF model. Adopting MARS as a first-stage modeling tool and the results that were achieved being the inputs to RF was contributed to the achievement of the model [57]. As said previously, the process should always take into consideration the importance of data structure. According to general outcomes, MARS single model, NN single model, and MARS-RF, RF-MARS, NN-MARS two-stage models achieved better results in this kind of sentiment and consumer confidence data and thereby, their model selection criteria worked better regarding other models.

Furthermore, we obtained findings for the volatility models for each of the variable that we have studied in this thesis. According to the results, the best performance was obtained for the CPI and especially for the USD/TRY exchange rate. This case was not suprising, because of the fact that these were the most volatile dataset the financial and economical stucture of Türkiye. On the other hand, among the sentiment indexes which were the first sentiment index that we introduced in the beginning of the thesis and the Consumer Confidence Index; it had better results according to the AIC and BIC values and obtained more model according to the CCI. These were the assesments for the univariate volatility models. We also performed volatility models for multivariate case. We obtained more kinds of volatility models in Sentiment Index (ARCH, GARCH and EGARCH) and only one (ARCH) in Consumer Confidence

Index. However, model selection criteria had better results for the CCI based on its AIC and BIC values.

At this point, there might be some modifications in behavior which can cause to 'regime shift' or 'structural break'. CCI contains these kind of shifts in it. In our findings, there were three regimes at total. We employed the first MS model to define regimes and then, we applied volatility models to each of them (MS-GARCH). Here, to avoid from biasness, we preferred to use MS-GARCH model.

According to the the transition probabilities matrix which was obtained by using MS model, it is hard to shift from on regime to the other. Furthermore, the model specified sufficiently the periods of each state. Additionaly, by applying volatility models to the each of these regimes, we obtained better and more accurate outcomes by only using standard GARCH model.

As the extension of the study, we plan to apply machine learning models (MARS, NN, RF, etc.) to each of regimes that we obtained by using MS model. By this way, we can conclude that we obtain better and more accurate results by only using machine learning or by separately applying these techniques to each of these regimes. In addition to the sentiment level of consumers and investors, we can delve into the effect of economical crises and financial bubbles [54]. Furthermore, we can extend these models by employing additional mathematical models; PCA, SVM, Generalized Additive Model (GAM), CMARS and RCMARS [23, 55, 70, 71, 73, 91, 95, 97]. Moreover, the particular relevance of this study for development and the developing countries will be discussed, worked out and submitted for future research.

REFERENCES

[1] Central Bank of the Republic of Turkey - Electronic Data Delivery System, `https://evds2.tcmb.gov.tr/`, accessed: [April 2022].

[2] Tüketici güven endeksi Şubat 2022, `https://data.tuik.gov.tr/Bulten/Index?p=Tuketici-Guven-Endeksi-Subat-2022-45803`, accessed: [March 2022].

[3] A. Abramson and I. Cohen, On the stationarity of markov-switching garch processes, Econometric Theory, 23(3), pp. 485–500, 2007.

[4] T. Afshar, G. Arabian, R. Zomorrodian, et al., Stock return, consumer confidence, purchasing managers index and economic fluctuations, Journal of Business & Economics Research (JBER), 5(8), 2007.

[5] C. C. Aggarwal, *Data mining: the textbook*, Springer, 2015.

[6] B. S. Arasu, M. Jeevananthan, N. Thamaraiselvan, and B. Janarthanan, Performances of data mining techniques in forecasting stock index–evidence from india and us, Journal of the National Science Foundation of Sri Lanka, 42(2), pp. 177–191, 2014.

[7] D. Ardia, K. Bluteau, K. Boudt, and L. Catania, Forecasting risk with markov-switching garch models: A large-scale performance study, International Journal of Forecasting, 34(4), pp. 733–747, 2018.

[8] İ. Arısoy, Türkiye ekonomisinde iktisadi güven endeksleri ve seçilmiş makro değişkenler arasındaki ilişkilerin var analizi, Maliye Dergisi, (162), pp. 304–315, 2012.

[9] M. Augustyniak, Maximum likelihood estimation of the markov-switching garch model, Computational Statistics and Data Analysis, 76, pp. 61–75, 2014.

[10] A. Bahrammirzaee, A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems, Neural Computing and Applications, 19(8), pp. 1165–1195, 2010.

[11] M. Baker and J. Wurgler, Investor sentiment and the cross-section of stock returns, The Journal of Finance, 61(4), pp. 1645–1680, 2006.

[12] M. Baker and J. Wurgler, Investor sentiment in the stock market, The Journal of Economic Perspectives, 21(2), pp. 129–151, 2007.

[13] L. Bauwens, A. Preminger, and J. V. Rombouts, Theory and inference for a markov switching garch model, The Econometrics Journal, 13(2), pp. 218–244, 2010.

[14] Ç. Başarır, İ. M. Bicil, and Ö. Yılmaz, The relationship between selected financial and macroeconomic variables with consumer confidence index, Journal of Yaşar University, 14, pp. 173–183, 2019.

[15] M. Bildirici and Ö. Ersin, Modeling markov switching arma-garch neural networks models and an application to forecasting stock returns, The Scientific World Journal, 2014.

[16] M. Billio, R. Casarin, and A. Osuntuyi, Efficient gibbs sampling for markov switching garch models, Computational Statistics and Data Analysis, 100, pp. 37–57, 2016.

[17] J. A. Bilmes et al., A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models, International Computer Science Institute, 4(510), p. 126, 1998.

[18] S.-K. Bormann, Sentiment indices on financial markets: What do they measure?, Technical report, Economics Discussion Papers, 2013.

[19] S. Brandel, Markov regime switching model implementation to the stockholm stock market & comparison with equal weight portfolio, 2017.

[20] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, CA, 1984.

[21] C. Brooks, Rats handbook to accompany introductory econometrics for finance, Cambridge Books, 2008.

[22] G. M. Caporale and T. Zekokh, Modelling volatility of cryptocurrencies using markov-switching garch models, Research in International Business and Finance, 48, pp. 143–155, 2019.

[23] A. Çevik, G.-W. Weber, B. M. Eyüboğlu, K. K. Oğuz, and A. D. N. Initiative, Voxel-mars: a method for early detection of alzheimer's disease by classification of structural brain mri, Annals of Operations Research, 258, pp. 31–57, 2017.

[24] P. Chen, D. Yi, and C. Zhao, Trading strategy for market situation estimation based on hidden markov model, Mathematics, 8(7), p. 1126, 2020.

[25] W. Chen, X. Wei, and K. Zhu, Engaging voluntary contributions in online communities: A hidden markov model, Mis Quarterly, 42(1), pp. 83–100, 2017.

[26] T. Cheng, J. Liu, W. Yao, and A. B. Zhao, The impact of covid-19 pandemic on the volatility connectedness network of global stock market, Pacific-Basin Finance Journal, 71, p. 101678, 2022.

[27] B. Chris, *Introductory econometrics for finance*, Cambridge Books, 2008.

[28] D. Colander, *The complexity vision and the teaching of economics, 2000.*

[29] E. D. Dar, V. Purutçuoglu, and E. Purutçuoglu, Detection of hiv-1 protease cleavage sites via hidden markov model and, Numerical Solutions of Realistic Nonlinear Phenomena, 31, p. 171, 2020.

[30] J. De Andrés, P. Lorca, F. J. de Cos Juez, and F. Sánchez-Lasheras, Bankruptcy forecasting: A hybrid approach using fuzzy c-means clustering and multivariate adaptive regression splines (mars), Expert Systems with Applications, 38(3), pp. 1866–1875, 2011.

[31] S. De Boef and P. M. Kellstedt, The political (and economic) origins of consumer confidence, American Journal of Political Science, 48(4), pp. 633–649, 2004.

[32] S. Dees and P. S. Brinca, Consumer confidence as a predictor of consumption spending: Evidence for the united states and the euro area, International Economics, 134, pp. 1–14, 2013.

[33] J. G. Dias, J. K. Vermunt, and S. Ramos, Clustering financial time series: New insights from an extended hidden markov model, European Journal of Operational Research, 243(3), pp. 852–864, 2015.

[34] N. Dorf and R. Documentation, Find the optimal testing configuration for non-informative two-stage hierarchical testing.

[35] R. Engle, Garch 101: The use of arch/garch models in applied econometrics, Journal of economic perspectives, 15(4), pp. 157–168, 2001.

[36] S. Ereeş, S. EREES, and N. DEMİREL, Omitted variable bias and detection with reset test in regression analysis, Anadolu University Journal of Science and Technology B-Theoretical Sciences, 2(1), pp. 1–19, 2012.

[37] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, volume 1, Springer series in statistics Springer, Berlin, 2001.

[38] C. Frydman and C. F. Camerer, The psychology and neuroscience of financial decision making, Trends in Cognitive Sciences, 20(9), pp. 661–675, 2016.

[39] C. D. Frydman, *Essays in neurofinance*, Ph.D. thesis, California Institute of Technology, 2012.

[40] Å. Grek, Forecasting accuracy for arch models and garch (1, 1) family: Which model does best capture the volatility of the swedish stock market?, 2014.

[41] M. Haas, S. Mittnik, and M. S. Paolella, A new approach to markov-switching garch models, Journal of financial Econometrics, 2(4), pp. 493–530, 2004.

[42] J. D. Hamilton, *Regime switching models*, Springer, 2010.

[43] M. R. Hassan and B. Nath, Stock market forecasting using hidden markov model: a new approach, in *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pp. 192–196, IEEE, 2005.

[44] Y. Huang, Estimation and testing of nonparametric hidden markov model with application in stock market, Communications in Statistics-Theory and Methods, 49(24), pp. 5917–5929, 2020.

[45] T. U. Islam and M. N. Mumtaz, Consumer confidence index and economic growth: An empirical analysis of EU countries., EuroEconomica, 35(2), 2016.

[46] S. Jadhav, H. He, and K. W. Jenkins, An academic review: applications of data mining techniques in finance industry, 2017.

[47] B. Kalaycı, *Identification of coupled systems of stochastic differential equations in finance including investor sentiment by multivariate adaptive regression splines*, MSc. Thesis, Middle East Technical University, 2017.

[48] B. Kalaycı, A. Özmen, and G.-W. Weber, Mutual relevance of investor sentiment and finance by modeling coupled stochastic systems with MARS, Annals of Operations Research, 295(173), pp. 1–24, 2020.

[49] L.-J. Kao, C.-C. Chiu, C.-J. Lu, and C.-H. Chang, A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting, Decision Support Systems, 54(3), pp. 1228–1244, 2013.

[50] C. Katris, Prediction of unemployment rates with time series and machine learning techniques, Computational Economics, 55(2), pp. 673–706, 2020.

[51] M. Khan, U. N. Kayani, M. Khan, K. S. Mughal, and M. Haseeb, Covid-19 pandemic & financial market volatility; evidence from garch models, Journal of Risk and Financial Management, 16(1), p. 50, 2023.

[52] J.-S. Kim and E. W. Frees, Omitted variables in multilevel models, Psychometrika, 71(4), pp. 659–690, 2006.

[53] T. Korkmaz and E. Çevik, Reel kesim güven endeksi ile İMKB 100 endeksi arasındaki dinamik nedensellik ilişkisi, İstanbul Üniversitesi İşletme Fakültesi Dergisi, 38(1), pp. 24–37, 2009.

[54] E. Kürüm, G.-W. Weber, and C. Iyigun, Early warning on stock market bubbles via methods of optimization, clustering and inverse problems, Annals of Operations Research, 260(1-2), pp. 293–320, 2018.

[55] S. Kuter, Z. Akyurek, and G.-W. Weber, Retrieval of fractional snow covered area from modis data by multivariate adaptive regression splines, Remote Sensing of Environment, 205, pp. 236–252, 2018.

[56] J. Lee and M. Shin, Stock forecasting using hidden markov processes, ht tp://cs229. stanford. edu/proj2009/ShinLee. pdf, 2009.

[57] T.-S. Lee and I.-F. Chen, A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines, Expert Systems with Applications, 28(4), pp. 743–752, 2005.

[58] T.-S. Lee, C.-C. Chiu, Y.-C. Chou, and C.-J. Lu, Mining the customer credit using classification and regression tree and multivariate adaptive regression splines, Computational Statistics & Data Analysis, 50(4), pp. 1113–1130, 2006.

[59] T.-S. Lee and C. Yang, Incorporating financial ratios and intellectual capital in bankruptcy predictions, in *Proceedings of the National Taiwan University International Conference in Finance, Taiwan, December*, pp. 20–21, Citeseer, 2004.

[60] N. Li, *Hidden Markov model and financial application*, Ph.D. thesis, The University of Texas at Austin, 2016.

[61] H.-Y. Lin, Y. Ann Chen, Y.-Y. Tsai, X. Qu, T.-S. Tseng, and J. Y. Park, Trm: A powerful two-stage machine learning approach for identifying snp-snp interactions, Annals of human genetics, 76(1), pp. 53–62, 2012.

[62] S. L. Lin, A two-stage logistic regression-ann model for the prediction of distress banks: Evidence from 11 emerging countries, African Journal of Business Management, 4(14), pp. 3149–3168, 2010.

[63] A. Lojowska, D. Kurowicka, G. Papaefthymiou, and L. van der Sluis, Advantages of arma-garch wind speed time series modeling, in *2010 IEEE 11th International Conference on Probabilistic Methods Applied to Power Systems*, pp. 83–88, 2010.

[64] C.-J. Lu, C.-H. Chang, C.-Y. Chen, C.-C. Chiu, and T.-S. Lee, Stock index prediction: A comparison of mars, bpn and svr in an emerging market, in *2009 IEEE International conference on Industrial Engineering and Engineering Management*, pp. 2343–2347, IEEE, 2009.

[65] L. Ma, C. Hu, R. Lin, and Y. Han, Arima model forecast based on eviews software, in *IOP Conference Series: Earth and Environmental Science*, volume 208, p. 012017, IOP Publishing, 2018.

[66] F. Özkurt Yerlikaya, *Refinements, Extensions and Modern Applications of Conic Multivariate Regression Splines*, PhD. thesis, Middle East Technical University, Ankara, Turkey, 2013.

[67] F. Özkurt Yerlikaya and G. Weber, Identification of stochastic differential equations by conic optimization of multivariate adaptive regression splines, Preprint 2013-20, Middle East Technical University, Ankara, Turkey, 2013.

[68] A. Özmen, *Robust Conic Quadratic Programming in Applied to Quality Improvement - A Robustification of CMARS*, MSc thesis, Middle East Technical University, Ankara, Turkey, 2010.

[69] A. Özmen, *Robust Optimization of Spline Models and Complex Regulatory Networks*, Springer, 2016.

[70] A. Özmen, G.-W. Weber, and İ. Batmaz, The new robust CMARS (RCMARS) method, vectors, 1, pp. 362–368, 2010.

[71] A. Özmen, G. W. Weber, İ. Batmaz, and E. Kropat, Rcmars: Robustification of cmars with different scenarios under polyhedral uncertainty set, Communications in Nonlinear Science and Numerical Simulation, 16(12), pp. 4780–4787, 2011.

[72] A. Özmen, G.-W. Weber, Z. Çavuşoğlu, and Ö. Defterli, The new robust conic gplm method with an application to finance: prediction of credit default, Journal of Global Optimization, 56(2), pp. 233–249, 2013.

[73] A. Özmen, Y. Yılmaz, and G.-W. Weber, Natural gas consumption forecast with MARS and CMARS models for residential users, Energy Economics, 70, pp. 357–381, 2018.

[74] R. N. Paramanik and V. Singhal, Sentiment analysis of indian stock market volatility, Procedia Computer Science, 176, pp. 330–338, 2020.

[75] R. K. Pearson, Outliers in process modeling and identification, IEEE Transactions on control systems technology, 10(1), pp. 55–63, 2002.

[76] Ó. Pérez, M. Piccardi, J. García, M. Á. Patricio, and J. M. Molina, Comparison between genetic algorithms and the baum-welch algorithm in learning hmms for human activity classification, in *Workshops on Applications of Evolutionary Computation*, pp. 399–406, Springer, 2007.

[77] V. Plakandaras, T. Papadimitriou, and P. Gogas, Forecasting daily and monthly exchange rates with machine learning techniques, Journal of Forecasting, 34(7), pp. 560–573, 2015.

[78] L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE, 77(2), pp. 257–286, 1989.

[79] V. Ravi, D. Pradeepkumar, and K. Deb, Financial time series prediction using hybrids of chaos theory, multi-layer perceptron and multi-objective evolutionary algorithms, Swarm and Evolutionary Computation, 36, pp. 136–149, 2017.

[80] L. J. Rodríguez and I. Torres, Comparative study of the baum-welch and viterbi training algorithms applied to read and spontaneous speech recognition, in *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 847–857, Springer, 2003.

[81] L. Rupande, H. T. Muguto, and P.-F. Muzindutsi, Investor sentiment and stock return volatility: Evidence from the johannesburg stock exchange, Cogent Economics & Finance, 7(1), p. 1600233, 2019.

[82] R. Sajjad, J. Coakley, and J. C. Nankervis, Markov-switching garch modelling of value-at-risk, Studies in Nonlinear Dynamics & Econometrics, 12(3), 2008.

[83] E. Savku, Advances in optimal control of markov regime-switching models with applications in finance and economics, PhD.Thesis, Middle East Technical University, 2017.

[84] E. Savku and G.-W. Weber, A stochastic maximum principle for a markov regime-switching jump-diffusion model with delay and an application to finance, Journal of Optimization Theory and Applications, 179, pp. 696–721, 2018.

[85] J. G. Saw, M. C. Yang, and T. C. Mo, Chebyshev inequality with estimated mean and variance, The American Statistician, 38(2), pp. 130–132, 1984.

[86] P. Sephton, Forecasting recessions: Can we do better on MARS, Review, 83, 2001.

[87] D. Seçilmiş, *Deterministic Modeling and Inference of Biochemical Networks*, Ph.D. thesis, Middle East Technical University, 2017.

[88] M. Z. Shariff, J. Al-Khasawneh, and M. Al-Mutawa, Risk and reward: A neurofinance perspective, International Review of Business Research Papers, 8(6), pp. 126–133, 2012.

[89] E. M. Sledjeski, L. C. Dierker, R. Brigham, and E. Breslin, The use of risk assessment to predict recurrent maltreatment: A classification and regression tree analysis (cart), Prevention science, 9(1), pp. 28–37, 2008.

[90] R. Syah, M. K. Nasution, M. Elveny, and H. Arbie, Optimization model for customer behavior with mars and kyc system, Journal of Theoretical and Applied Information Technology, 98(13), 2020.

[91] P. Taylan, G.-W. Weber, and A. Beck, New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and technology, Optimization, 56(5-6), pp. 675–698, 2007.

[92] A. Tenyakov, Estimation of hidden markov models and their applications in finance, 2014.

[93] K. Tseng, Behavioral finance, bounded rationality, neuro-finance, and traditional finance, Investment Management and Financial Innovations, (3, Iss. 4), pp. 7–18, 2006.

[94] L. Wang, J. Wu, Y. Cao, and Y. Hong, Forecasting renewable energy stock volatility using short and long-term markov switching garch-midas models: either, neither or both?, Energy Economics, 111, p. 106056, 2022.

[95] G.-W. Weber, I. Batmaz, G. Köksal, P. Taylan, and F. Yerlikaya-Özkurt, Cmars: a new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization, Inverse Problems in Science and Engineering, 20(3), pp. 371–400, 2012.

[96] G.-W. Weber, P. Taylan, K. Yıldırak, and Z.-K. Görgülü, Financial regression and organization, Special Issue on Optimization in Finance, DCDIS-B, 17(16), pp. 149–174, 2010.

[97] G.-W. Weber, A. Tezel, P. Taylan, A. Soyler, and M. Çetin, Mathematical contributions to dynamics and optimization of gene-environment networks, Optimization, 57(2), pp. 353–377, 2008.

[98] J. Wurgler, Website of j. wurgler, `https://pages.stern.nyu.edu/~jwurgler/`, 2022.

[99] D. Yao, J. Yang, and X. Zhan, A novel method for disease prediction: hybrid of random forest and multivariate adaptive regression splines, Journal of Computers, 8(1), pp. 170–177, 2013.

[100] F. Yerlikaya-Özkurt, *Refinements, Extensions and Modern Applications of Conic Multivariate Regression Splines*, PhD. thesis, Middle East Technical University, Ankara, Turkey, 2013.

[101] F. Özkurt Yerlikaya and G. Weber, Identification of stochastic differential equations by conic optimization of multivariate adaptive regression splines, Preprint 2013-20, Middle East Technical University, Ankara, Turkey, 2013.

[102] C. Zhang, Defining, modeling, and measuring investor sentiment, University of California, Berkeley, Department of Economics, 2008.

[103] Y. Zhang, D. Zhao, and J. Liu, The application of baum-welch algorithm in multistep attack, The Scientific World Journal, 2014, 2014.

[104] M. Zouaoui, G. Nouyrigat, and F. Beer, How does investor sentiment affect stock market crises evidence from panel data, Financial Review, 46(4), pp. 723–747, 2011.

APPENDIX A

APPENDIX

In this part, some of the statistical results of Machine Learning models (MARS, NN and RF) are given.

As a conclusion of pure MARS method, all the basis functions and related statistical results are given in the following.

A.1 MARS Method

A.1.1 Statistical results of Sole MARS Model for Sentiment Index

In this part, before showing the results for MARS algorithm, we create a table in the following for the short names of the variables: According to the sole MARS model

Table A.1: List of the short names of all variables

Variables	Short names
closed-end fund discount	cefd
number of initial public offerings	nipo
return of initial public offerings	ripo
dividend premium	pdnd
the equity which shares new issues	s
industrial production index	indpro
nominal services consumption	consserve
nominal durables consumption	consdur
nominal nondurables consumption	consnon
consumer price index	cpi
employment	employ

results,statistical values and basis functions are as in the following.

- Selected 39 of 71 terms, and 11 of 12 predictors.
- Termination condition: RSq changed by less than 1e-06 at 71 terms.
- Importance: consserv, employ, consdur, indpro, cefd, s, consnon, ripo, cpi, nipo, pdnd, recess-unused.
- Number of terms at each degree of interaction: 1, 9, 17, 12.
  Here, 1 represents the intercept, 9 represents the basis functions with one variable $(h(1-nipo), h(cefd-4.27), ..., h(employ-130427))$,
  17 shows the basis fucntions with two variables, $(h(5149.4-consserv)*employ, h(pdnd--0.05)*h(nipo-1), ..., h(consserv-5149.4)*h(137793-employ))$,
  and lastly 12 indicates basis functions with three variables, $(h(pdnd--0.05)*h(nipo-1)*cefd, ..., h(2.9-ripo)*h(4.27-cefd)*h(5222.5-consserv))$

Table A.2: Statistical Results of the MARS model

GCV	0.0129
RSS	1.4510
GRSq	0.9775
RSq	0.9906
RMSE	0.0730
MSE	0.0050

According to the Table A.2, it is seen that the values $RMSE$, MSE and GCV show the accuracy of model is sufficient. According to the Table A.3, we obtain the following equation:

$$\begin{aligned} y = &-0.6041 - 0.1432*h(1-nipo) - 0.0649*h(cefd-4.27) + ... \\ &- 0.0001*h(cefd-4.27)*h(indpro-100.748)*h(140839-employ) \\ &- 0.0005*h(s-0.11)*h(102.909-indpro)*h(7618-consserv). \end{aligned} \tag{A.1}$$

Here, in Table A.4, residiuals of the MARS, mean quantile and third quantile are presented.

Table A.3: Coefficients of Basis Functions obtained with MARS model

Basis Functions	Coefficients
(Intercept)	-0.6041
h(1-nipo)	-0.1432
h(cefd-4.27)	-0.0649
h(102.909-indpro)	-0.05376
h(indpro-102.909)	-0.25786
h(consdur-1036.2)	0.00107
h(consdur-2008.5)	-0.0032
h(5149.4-consserv)	-0.2089
h(130427-employ)	-0.0008
h(employ-130427)	0.0000
h(5149.4-consserv)*employ	0.0000
h(pdnd- -0.05)*h(nipo-1)	0.0028
h(3.4-ripo) * h(cefd-4.27)	0.0016
h(ripo-3.4) * h(cefd-4.27)	0.0003
h(nipo-1) * h(s-0.17)	-0.2260
h(cefd-4.27) * h(indpro-100.748)	0.0207
h(4.27-cefd) * h(5222.5-consserv)	-0.0098
h(4.55-cefd) * h(employ-130427)	0.0000
h(4.27-cefd) * h(184.6-cpi)	0.0627
h(0.11-s) * h(102.909-indpro)	0.7900
h(s-0.11) * h(102.909-indpro)	0.7205
h(0.14-s) * h(consdur-1036.2)	-0.01563
h(90.474-indpro) * h(1036.2-consdur)	0.0008
h(99.9141-indpro) * h(employ-130427)	0.0000
h(consdur-1036.2) * h(consserv-9596.4)	0.0000
h(consserv-5149.4) * h(employ-137793)	0.0000
h(consserv-5149.4) * h(137793-employ)	0.0000
h(pdnd- -0.05) * h(nipo-1) * cefd	-0.0007
h(4.27-cefd) * consnon * h(5222.5-consserv)	0.0000
h(90.4879-indpro) * h(5149.4-consserv) * employ	0.0000
h(indpro-90.4879) * h(5149.4-consserv) * employ	0.0000
h(2.9-ripo) * h(4.27-cefd) * h(5222.5-consserv)	0.0008
h(ripo-2.9) * h(4.27-cefd) * h(5222.5-consserv)	-0.0001
h(5-nipo) * h(cefd-4.27) * h(100.748-indpro)	0.0005
h(cefd-4.27) * h(indpro-100.748) * h(140839-employ)	-0.0001
h(cefd-4.27) * h(100.748-indpro) * h(131596-employ)	0.0000
h(s-0.11) * h(102.909-indpro) * h(7618-consserv)	-0.0005

Table A.4: Residuals of the MARS model

Min	First Quantile	Median	Third Quantile	Maximum
-0.2404	-0.0429	0.0005	0.0444	0.2512

On the other hand, according to the Figure A.1, which is the plot for prediction of sentiment data, it is seen that there is linear expression.

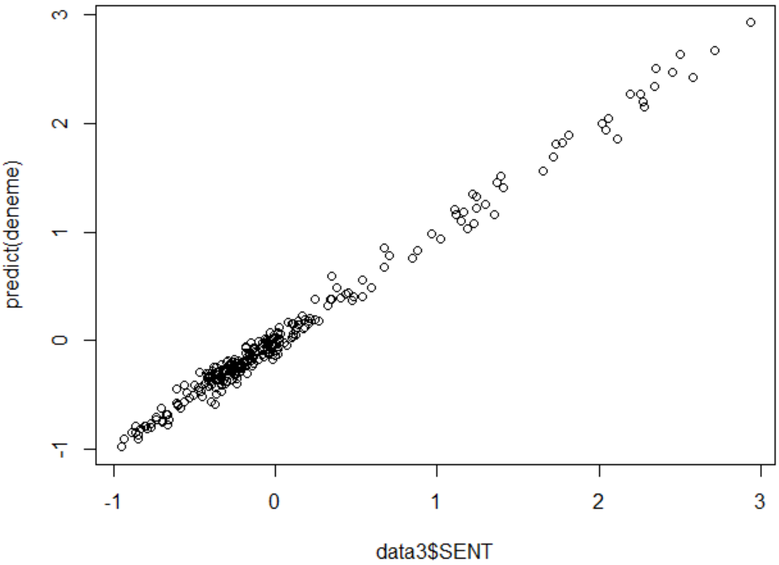


Figure A.1: Visualization of Output (sentiment) and Predicted Output

A.1.2 Statistical results of RF-MARS Model for Sentiment Index

- Selected 48 of 71 terms, and 9 of 10 predictors.
- Termination condition: RSq changed by less than 1e-06 at 71 terms.
- Importance: consnon, employ, consdur, indpro, cefd, s, pdnd, nipo, ripo, recess-unused.
- Number of terms at each degree of interaction: 1, 11, 23, 13.

Here, 1 represents the intercept, 11 represents the basis functions with one variable $(h(5 - nipo), h(nipo - 5), ..., h(130427 - employ))$,
23 shows the basis fucntions with two variables, $(h(4.27 - cefd) * employ, ..., -0.05) * h(nipo - 1), ..., h(2004.2 - consnon) * h(employ - 130427))$,

and lastly 13 indicates basis functions with three variables, $(h(cefd - 1.01) * h(1036.2 - consdur) * consnon*, ..., h(indpro - 99.0974) * h(consnon - 1694.3) * h(employ - 137993))$

Table A.5: Statistical Results of RF-MARS model

GCV	0.0142
RSS	1.2070
GRSq	0.9753
RSq	0.9922
RMSE	0.0670
MSE	0.0040

It seen in Table A.5, that the values $RMSE$, MSE and GCV show the accuracy of model is sufficient.

According to the Table A.6, we obtain the following equation:

$$\begin{aligned} y &= 2.6526 - 0.0187 * h(5 - nipo) - 0.0014 * h(nipo - 5) + ... \\ &+ 0.2816 * h(cefd - 1.01) * h(s - 0.09) * h(1036.2 - consdur) \\ &- 0.0002 * h(6.09 - cefd) * h(102.909 - indpro) * h(1825.1 - consnon). \end{aligned} \tag{A.2}$$

Here, in Table A.7, residiuals of the RF-MARS, mean quantile and third quantile are presented.

A.1.3 Statistical results of NN-MARS Model for Sentiment Index

- Selected 39 of 55 terms, and 3 of 3 predictors.
- Termination condition: Reached maximum RSq 1.0000 at 55 terms.
- Importance: consserv, employ, consnon.
- Number of terms at each degree of interaction: 1, 8, 22, 8.
  Here, 1 represents the intercept, 8 represents the basis functions with one variable $(h(consnon - 1681), h(1775.3 - consnon), ..., h(employ - 140568))$,

22 shows the basis functions with two variables, $(h(1775.3-consnon)*employ,$ $h(consserv-7499.1)*h(149269-employ), ..., h(consserv-7499.1)*h(149269-employ)),$

and lastly 8 indicates basis functions with three variables, $(h(1645-consnon)*h(7499.1-consserv)*h(132694-employ), ..., h(consnon-2646.6)*h(9302.6-consserv)*h(employ-140568))$

It is seen in Table A.8, that the values $RMSE$, MSE and GCV indicate the accuracy of model is sufficient.

According to the Table A.9, we obtain the following equation:

$$\begin{aligned} y = &-66.2700 - 0.0210 * h(consnon - 1681) - 0.6340 * h(1775.3 - consnon) + ... \\ &+ 0.0090 * h(140568 - employ) - 0.0100 * h(employ - 140568). \end{aligned} \tag{A.3}$$

Here, in Table A.10, residiuals of the NN-MARS, mean quantile and third quantile are presented.

A.2 RF Method

A.2.1 Statistical results of Sole RF Model for Sentiment Index

According to the sole RF Model results, statistical values as in the following.

Here, according to the Table A.11, number of variability is highest when the number of tree is 500, afterwards, it started to decrease, thus it is stopped at the $tree = 500$. Furthermore, when the number of tree is taken as $\%95.49$ of variance is explained.

In Table A.12, since higher values of Incremental MSE show us that a variable has a stronger impact on reducing the MSE, $consserv$ has the highest effect.

A.2.2 Statistical results of MARS-RF Model for Sentiment Index

Here, Table A.13, number of variability is highest when the number of tree is 500, afterwards, it started to decrease, thus it is stopped at the $tree = 500$. Furthermore, when the number of tree is taken as $\%88.61$ of variance is explained.

A.2.3 Statistical results of NN-RF Model for Sentiment Index

In Table A.14, number of variability is highest when the number of tree is 500, afterwards, it started to decrease, thus it is stopped at the $tree = 500$. In fact, when the number of tree is taken as $\%94.59$ of variance is explained.

A.3 NN Method

A.3.1 Statistical results of Sole NN Model for Sentiment Index

According to the sole NN Model results, statistical values as in the following.

It is seen in Table A.15, the most affective result is obtained for the $employ$ variable to the first hidden layer.

A.3.2 Statistical results of MARS-NN Model for Sentiment Index

In Table A.16, the most affective result is obtained for the $cefd$ variable to the first hidden layer.

A.3.3 Statistical results of RF-NN Model for Sentiment Index

The most affective result is obtained for the $cefd$ variable to the first hidden layer which is shown in Table A.17.

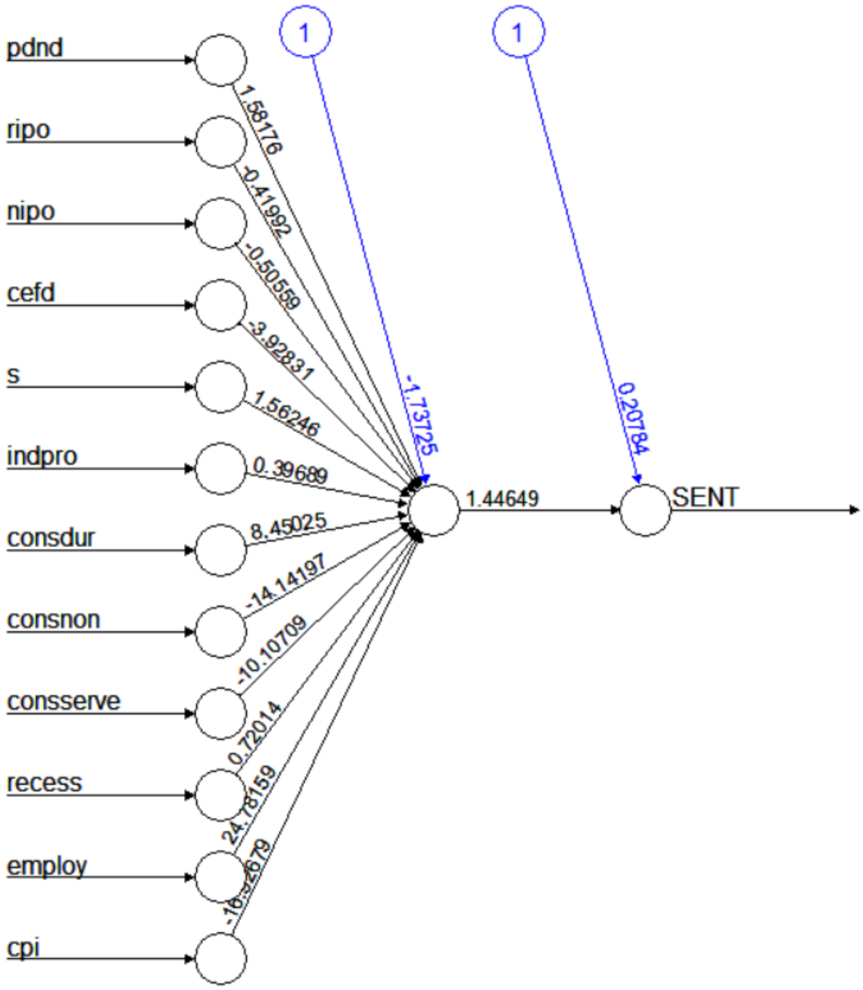


Figure A.2: Visualization of sole Neural Network Model

Table A.6: Coefficients of Basis Functions obtained with RF-MARS model

Basis Functions	Coefficients
(Intercept)	2.6526
h(5-nipo)	0.0187
h(nipo-5)	-0.0014
h(4.27-cefd)	-7.2897
h(cefd-4.27)	-0.045
h(102.909-indpro)	-0.2011
h(consdur-1002.4)	-0.0670
h(1036.2-consdur)	0.0209
h(consdur-1036.2)	0.0612
h(consdur-2008.5)	-0.0047
h(1694.3-consnon)	-0.4167
h(130427-employ)	-0.0005
h(4.27-cefd) * employ	0.0001
indpro * h(1694.3-consnon)	-0.0057
h(1694.3-consnon) * employ	0.0000
h(pdnd- -0.94) * h(1036.2-consdur)	0.0005
h(ripo-30) * h(cefd-4.27)	0.0006
h(4.27-cefd) * h(indpro-95.7455)	-0.0151
h(4.27-cefd) * h(95.7455-indpro)	0.0310
h(cefd-6.09) * h(102.909-indpro)	0.0031
h(cefd-1.01) * h(1036.2-consdur)	0.0535
h(3.28-cefd) * h(1036.2-consdur)	-0.0086
h(cefd-3.28) * h(1036.2-consdur)	0.0039
h(4.27-cefd) * h(consnon-1788.1)	-0.0004
h(4.27-cefd) * h(1788.1-consnon)	0.0037
h(0.14-s) * h(consdur-1036.2)	-0.0086
h(100.14-indpro) * h(consdur-1036.2)	-0.0039
h(indpro-100.14) * h(consdur-1036.2)	0.0044
h(101.704-indpro) * h(consdur-1002.4)	0.0039
h(indpro-101.704) * h(consdur-1002.4)	-0.0042
h(99.0974-indpro) * h(consnon-1694.3)	0.0001
h(indpro-99.0974) * h(consnon-1694.3)	-0.0006
h(consdur-1036.2) * h(consnon-2848.5)	0.0000
h(1036.2-consdur) * h(130982-employ)	0.0000
h(2004.2-consnon) * h(employ-130427)	0.0000
h(cefd-1.01) * h(1036.2-consdur) * consnon	0.0000
h(cefd-1.01) * h(1036.2-consdur) * employ	0.0000
h(cefd-6.09) * h(0.1-s) * h(102.909-indpro)	0.3513
h(cefd-1.01) * h(s-0.1) * h(1036.2-consdur)	-0.2813
h(cefd-1.01) * h(s-0.09) * h(1036.2-consdur)	0.2816
h(cefd-4.27) * h(indpro-99.3817) * h(consdur-1272.7)	0.0000
h(cefd-6.09) * h(102.909-indpro) * h(consdur-983.9)	0.0000
h(6.09-cefd) * h(102.909-indpro) * h(consnon-1825.1)	0.0000
h(6.09-cefd) * h(102.909-indpro) * h(1825.1-consnon)	-0.0002
h(1.27-cefd) * h(2004.2-consnon) * h(employ-130427)	0.0000
h(indpro-99.0974) * h(consnon-1694.3) * h(employ-140377)	0.0000
h(indpro-99.0974) * h(consnon-1694.3) * h(140377-employ)	0.0000
h(indpro-99.0974) * h(consnon-1694.3) * h(employ-137993)	0.0000

Table A.7: Residuals of RF-MARS model

Min	First Quantile	Median	Third Quantile	Maximum
-0.2369	-0.0455	-0.0038	0.0456	0.1981

Table A.8: Statistical Results of NN-MARS model

GCV	0.0262
RSS	2.9410
GRSq	0.9544
RSq	0.9809
RMSE	0.1040
MSE	0.0110

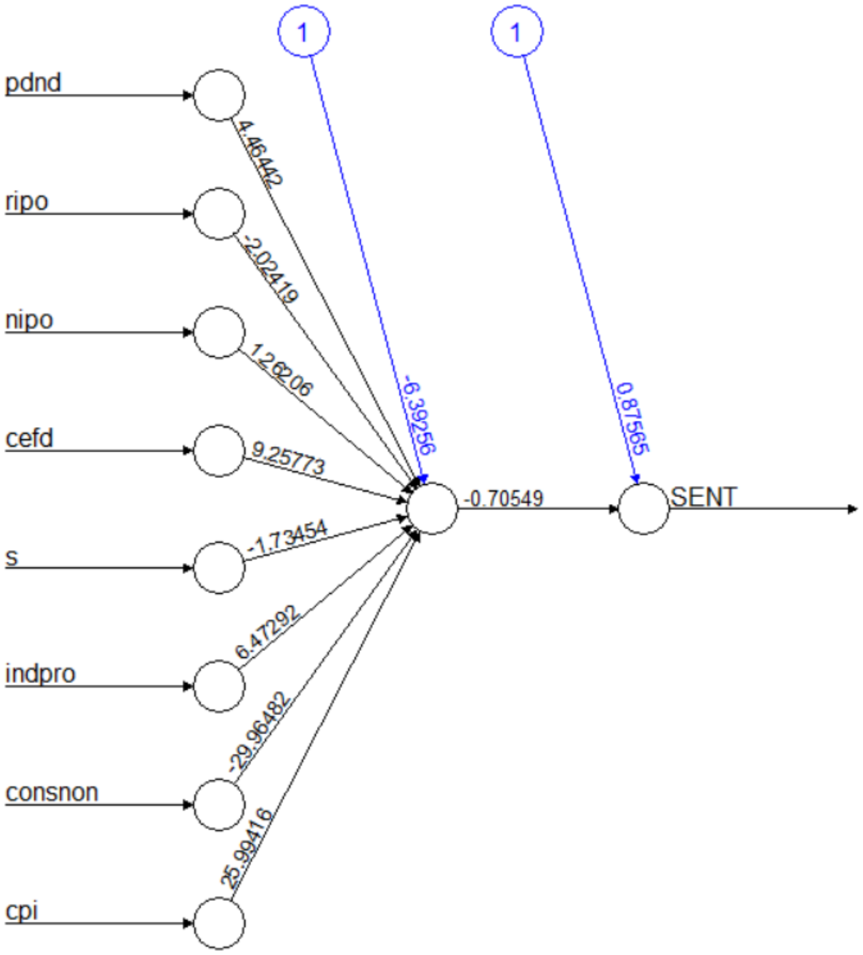


Figure A.3: Visualization of MARS-NN Model.

Table A.9: Statistical Results of Basis Functions obtained with NN-MARS model

Basis Functions	Coefficients
(Intercept)	-66.2700
h(consnon-1681)	-0.0210
h(1775.3-consnon)	-2.6340
h(consserv-4735.8)	0.0550
h(consserv-5345)	-0.00040
h(consserv-7499.1)	-0.0630
h(employ-130841)	-0.0030
h(140568-employ)	0.0090
h(employ-140568)	-0.0100
h(1775.3-consnon) * employ	0.0000
h(consserv-4565.3) * employ	0.0000
h(consnon-1775.3) * h(consserv-8840.3)	0.0000
h(consnon-1775.3) * h(8840.3-consserv)	0.0000
h(2615-consnon) * h(consserv-4735.8)	0.0000
h(consnon-2615) * h(consserv-4735.8)	0.0000
h(consnon-2697.3) * h(consserv-9692.5)	0.0000
h(consnon-2697.3) * h(9692.5-consserv)	0.0000
h(consnon-1775.3) * h(employ-146388)	0.0000
h(consnon-1775.3) * h(146388-employ)	0.0000
h(consnon-1775.3) * h(employ-133752)	0.0000
h(2373.5-consnon) * h(employ-130841)	0.0000
h(consnon-2373.5) * h(employ-130841)	0.0000
h(2646.6-consnon) * h(employ-140568)	0.0000
h(consnon-2646.6) * h(employ-140568)	0.0000
h(consnon-2697.3) * h(employ-145071)	0.0000
h(6239-consserv) * h(employ-130841)	0.0000
h(consserv-6239) * h(employ-130841)	0.0000
h(7499.1-consserv) * h(employ-132694)	0.0000
h(7499.1-consserv) * h(132694-employ)	0.0000
h(consserv-7499.1) * h(employ-149269)	0.0000
h(consserv-7499.1) * h(149269-employ)	0.0000
h(1645-consnon) * h(7499.1-consserv) * h(132694-employ)	0.0000
h(consnon-1645) * h(7499.1-consserv) * h(132694-employ)	0.0000
h(consnon-1775.3) * h(consserv-8840.3) * h(employ-149269)	0.0000
h(consnon-1775.3) * h(consserv-8840.3) * h(149269-employ)	0.0000
h(consnon-1775.3) * h(8840.3-consserv) * h(130623-employ)	0.0000
h(consnon-1788.1) * h(7499.1-consserv) * h(132694-employ)	0.0000
h(consnon-2646.6) * h(consserv-9302.6) * h(employ-140568)	0.0000
h(consnon-2646.6) * h(9302.6-consserv) * h(employ-140568)	0.0000

Table A.10: Residuals of NN-MARS model

Min	First Quantile	Median	Third Quantile	Maximum
-0.2369	-0.0455	-0.0038	0.045	0.1982

Table A.11: Statistical results of RF Model

Number of trees	50
Mean of squared residuals	0.0324
%Var explained	94.3300

Tree	MSE	%Var(y)
100	0.02703	4.74
200	0.0265	4.64
300	0.0258	4.53
400	0.0262	4.59
500	0.0257	4.51

Table A.12: Incremental MSE values of variables of RF Model

input values	% IncMSE
pdnd	10.6218
ripo	7.8820
nipo	9.6510
cefd	12.8831
s	10.8064
indpro	20.2900
consdur	16.0100
consnon	17.2460
consserv	21.1340
recess	4.6495
employ	19.1710
cpi	19.2001

Table A.13: Statistical results of MARS-RF Model

Number of trees	500
Mean of squared residuals	0.0650
%Var explained	88.6100

Tree	MSE	%Var(y)
100	0.0755	13.22
200	0.0695	12.18
300	0.06797	11.91
400	0.02674	11.81
500	0.0670	11.73

Table A.14: Statistical results of NN-RF Model

Number of trees	500
Mean of squared residuals	0.0310
%Var explained	94.5900

Tree	MSE	%Var(y)
100	0.0348	6.11
200	0.0330	5.78
300	0.0314	5.50
400	0.03070	5.38
500	0.0310	5.41

Table A.15: Visualization of variables to first hidden layer with sole Neural Network Model

Result matrix	Values
error	1.9524
reached.threshold	0.0069
steps	470
Intercept	-1.7330
pdnd	1.5817
ripo	-0.4200
nipo	-0.5056
cefd	-3.9283
s to first hidden layer	1.5625
indpro	0.39689
consdur	8.4503
consnon	-14.1420
consserve	-10.1071
recess	0.72014
employ	24.7816
cpi	-16.9268
Intercept	0.2078
SENT	1.4465

Table A.16: Visualization of variables to first hidden layer with MARS-NN Model

Result matrix	Values
error	2,07E+06
reached.threshold	9,44E+03
steps	2,27E+09
Intercept	-6,39E+06
pdnd	4,46E+06
ripo	-2,02E+06
nipo	1,26E+06
cefd	9,26E+06
s	-1,73E+06
indpro	6,47E+06
consnon	-3,00E+07
cpi	2,60E+07
Intercept	8,76E+05
SENT	-7,05E+05

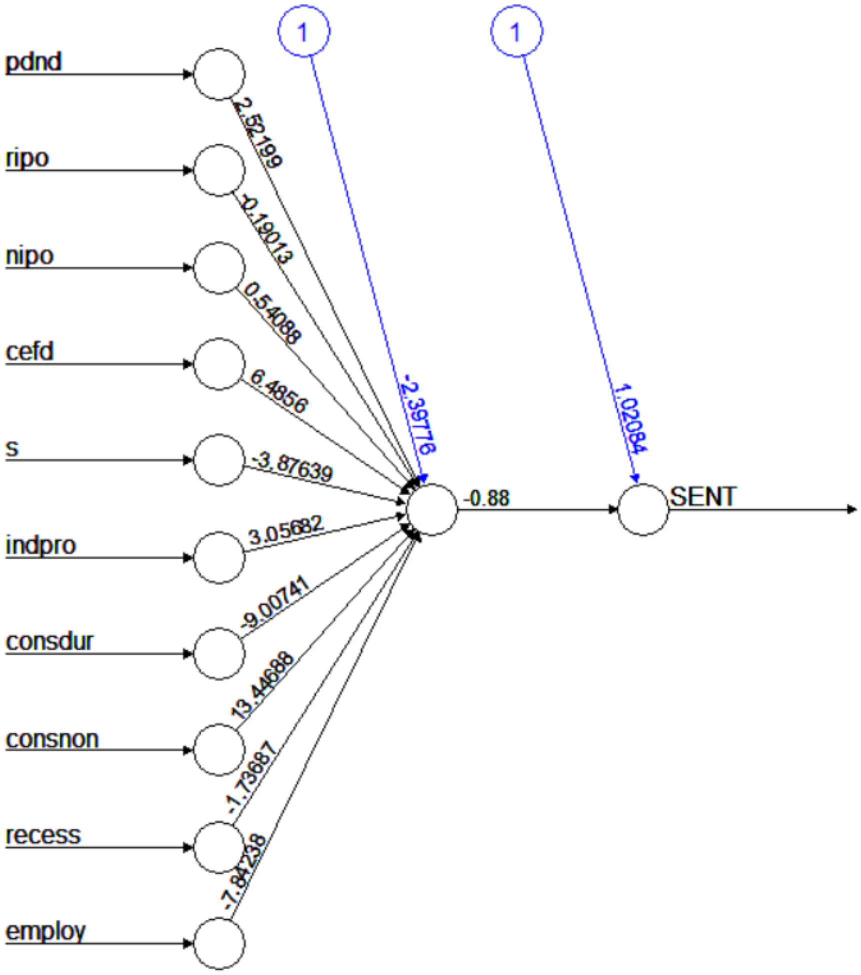


Figure A.4: Visualization of RF-NN Model.

Table A.17: Visualization of variables to first hidden layer with RF-NN Model

error	1,02E+06
reached.threshold	7,62E+03
steps	1,08E+09
Intercept	-2,40E+06
pdnd	2,52E+06
ripo	-1,90E+05
nipo	5,41E+05
cefd	6,49E+06
s	-3,88E+06
indpro	3,06E+06
consdur	-9,01E+06
consnon	1,34E+07
recess	-1,74E+06
employ	-7,84E+06
Intercept	1,02E+06
SENT	-8,80E+05

APPENDIX B

APPENDIX

In this part, some of the statistical results of Machine Learning models (MARS, NN and RF) are given.

As a conclusion of pure MARS method, all the basis functions and related statistical results are given in the following.

B.1 MARS Method

B.1.1 Statistical results of Sole MARS Model for Consumer Confidence Index

According to the sole MARS model results,statistical values and basis functions are as in the following.

- Selected 30 of 58 terms, and 15 of 208 predictors.
- Termination condition: Reached nk 100.
- Importance: USDTRYINDEX, CPI, UN, Time2008-11, Time2008-12, Time2008-04, Time2009-01, Time2008-06, Time2008-05, ...
- Number of terms at each degree of interaction: 1 15 9 5. Here, 1 represents the intercept, 15 represents the basis functions with one variable $(Time2008-12, h(UN-11.2), ..., h(USDTRYINDEX-556.115))$,
  9 shows the basis fucntions with two variables, $(Time2008-03*h(309.78-CPI), ..., Time2015-11*h(309.78-CPI))$,
  and lastly 5 indicates basis functions with three variables, $Time2015-12*$

Table B.1: Coefficients of Basis Functions obtained with MARS model

GCV	6.7990
RSS	579.4
GRSq	0.8583
RSq	0.9408
RMSE	1.6770
MSE	2.8130

$h(USDTRYINDEX) * h(USDTRYINDEX - 163.223), ...Time2012 - 10 * h(CPI - 162.15) * h(163.223 - USDTRYINDEX))$

According to the Table B.1, it is seen that the values $RMSE$, MSE and GCV show the accuracy of model is sufficient.

According to the Table B.2, we obtain the following equation:

$$\begin{aligned} y = {} & 263.5 - 8.6210 * (Time2008 - 12) + 2.1750 * h(UN - 11.2) + ... \\ & - 0.2840 * (Time2008 - 11) * h(162.15 - CPI) - 0.0040 * Time2012 \\ & - 10 * h(CPI - 162.15) * h(163.223 - USDTRYINDEX). \end{aligned} \tag{B.1}$$

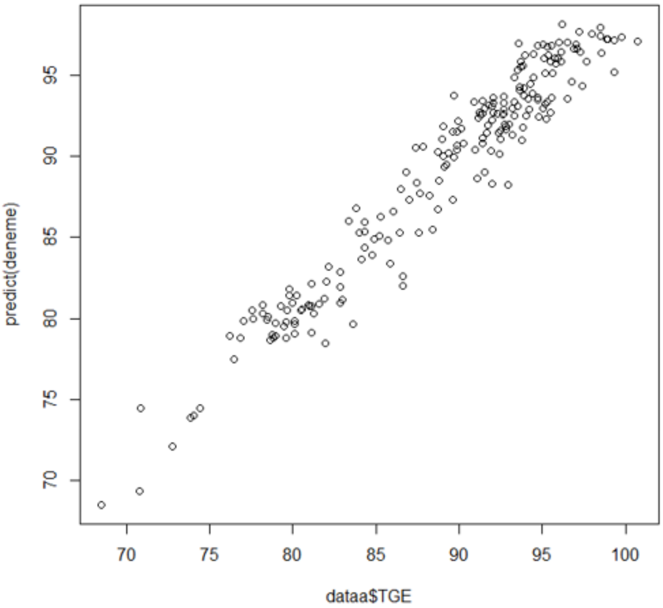


Figure B.1: Visualization of Output (Consumer Confidence) and Predicted Output

Table B.2: Coefficients of Basis Functions obtained with MARS model

Basis Functions	Coefficients
(Intercept)	263.500
Time2008-12	-8.6210
h(UN-11.2)	-2.1750
h(CPI-139.33)	-0.6240
h(CPI-188.67)	-0.7470
h(CPI-207.55)	0.6450
h(CPI-286.33)	-0.4320
h(309.78-CPI)	-1.0280
h(CPI-309.78)	0.7050
h(CPI-327.41)	0.2560
h(CPI-401.27)	0.2020
h(USDTRYINDEX-117.615)	-0.3160
h(USDTRYINDEX-149.754)	-0.4230
h(163.223-USDTRYINDEX)	-0.2250
h(USDTRYINDEX-163.223)	0.7400
h(USDTRYINDEX-556.115)	-0.0310
Time2008-03 * h(309.78-CPI)	-0.0360
Time2008-06 * h(309.78-CPI)	-0.0520
Time2008-10 * h(309.78-CPI)	-0.0490
Time2009-01 * h(UN-11.2)	-2.3460
Time2015-11 * h(309.78-CPI)	0.1640
Time2015-12 * h(USDTRYINDEX) * h(USDTRYINDEX-163.223)	0.1020
h(UN-11.2) * h(CPI-250.45)	0.0120
h(162.15-CPI) * h(163.223-USDTRYINDEX)	0.0160
h(336.48-CPI) * h(USDTRYINDEX)	-0.0040
Time2008-04 * h(162.15-CPI)	-0.0014
Time2008-05 * h(162.15-CPI)	-0.0160
Time2008-07 * h(162.15-CPI)	-0.0110
Time2008-11 * h(162.15-CPI)	-0.2840
Time2012-10 * h(CPI-162.15)*h(163.223-USDTRYINDEX)	-0.0040

B.1.2 Statistical results of RF-MARS Model for Consumer Confidence Index

- Selected 20 of 60 terms, and 2 of 2 predictors.
- Termination condition: Reached nk 100.
- Importance: USDTRYINDEX, CPI.
- Number of terms at each degree of interaction: 1 10 9. Here, 1 represents the intercept, 10 represents the basis functions with one variable $(Time2008-12, h(CPI-125.84), ..., h(USDTRYINDEX-423.157))$,
  9 shows the basis fucntions with two variables, $(h(CPI-140.13)*h(156.934-USDTRYINDEX), ..., h(422.84-CPI)*h(147.426-USDTRYINDEX))$.

Table B.3: Statistical Results of RF-MARS model

GCV	6.3360
RSS	763
GRSq	0.8680
RSq	0.9221

According to the Table B.3, it is seen that the values GCV show the accuracy of model is sufficient.

Table B.4: Coefficients of Basis Functions obtained with RF-MARS model

(Intercept)	316.3700
h(CPI-125.84)	-0.5600
h(CPI-282.58)	-0.4000
h(CPI-348.34)	-0.5700
h(422.84-CPI)	-0.8400
h(CPI-422.84)	-0.1201
h(CPI-465.84)	-0.1600
h(156.934-USDTRYINDEX)	0.7300
h(USDTRYINDEX-156.934)	-0.2600
h(USDTRYINDEX-172.464)	0.3800
h(USDTRYINDEX-423.157)	-0.1600
h(CPI-140.13) * h(156.934-USDTRYINDEX)	-0.0200
h(160.9-CPI) * h(156.934-USDTRYINDEX)	0.0100
h(CPI-160.9) * h(156.934-USDTRYINDEX)	0.1100
h(CPI-163.19) * h(156.934-USDTRYINDEX)	-0.1100
h(CPI-174.07) * h(156.934-USDTRYINDEX)	0.0600
h(CPI-174.07) * h(156.956-USDTRYINDEX)	-0.0400
h(CPI-187.31) * h(156.934-USDTRYINDEX)	-0.0200
h(422.84-CPI) * h(USDTRYINDEX-147.426)	0.0000
h(422.84-CPI) * h(147.426-USDTRYINDEX)	0.0000

From the Table B.4, we obtain the following equation:

$$\begin{aligned} y = 316.37 - 0.5600 * h(CPI - 125.84) - 0.4000 * h(CPI - 282.58) \\ + ... + 0.0600 * h(CPI - 174.07) * h(156.934 - USDTRYINDEX) \\ - 0.0200 * h(CPI - 187.31) * h(156.934 - USDTRYINDEX). \end{aligned} \tag{B.2}$$

B.1.3 Statistical results of NN-MARS Model for Consumer Confidence Index

- Selected 20 of 60 terms, and 2 of 2 predictors.

Table B.5: Statistical Results of NN-MARS model

GCV	6.3360
RSS	763
GRSq	0.8680
RSq	0.9221

- Termination condition: Reached nk 100.
- Importance: USDTRYINDEX, CPI.
- Number of terms at each degree of interaction: 1 10 9.

According to the Table B.5, it is seen that the values GCV show the accuracy of model is sufficient.

Table B.6: Coefficients of Basis Functions obtained with NN-MARS model

(Intercept)	316.3700
h(CPI-125.84)	-0.5600
h(CPI-282.58)	-0.4000
h(CPI-348.34)	-0.5700
h(422.84-CPI)	-0.8400
h(CPI-422.84)	1.7100
h(CPI-465.84)	-0.1600
h(156.934-USDTRYINDEX)	0.7300
h(USDTRYINDEX-156.934)	-0.2600
h(USDTRYINDEX-172.464)	0.3800
h(USDTRYINDEX-423.157)	-0.1600
h(CPI-140.13) * h(156.934-USDTRYINDEX)	-0.0200
h(160.9-CPI) * h(156.934-USDTRYINDEX)	0.0100
h(CPI-160.9) * h(156.934-USDTRYINDEX)	0.1100
h(CPI-163.19) * h(156.934-USDTRYINDEX)	-0.1100
h(CPI-174.07) * h(156.934-USDTRYINDEX)	0.0600
h(CPI-177.04) * h(156.934-USDTRYINDEX)	-0.0400
h(CPI-187.31) * h(156.934-USDTRYINDEX)	-0.0200
h(422.84-CPI) * h(USDTRYINDEX-147.426)	0.0000
h(422.84-CPI) * h(147.426-USDTRYINDEX)	0.0000

According to the Table B.6, we obtain the following equation:

$$\begin{aligned} y &= 316.37 - 0.5600 * h(CPI - 125.84) - 0.4000 * h(CPI - 282.58) \\ &+ ... - 0.0400 * h(CPI - 177.04) * h(156.934 - USDTRYINDEX) \\ &- 0.0200 * h(CPI - 187.31) * h(156.934 - USDTRYINDEX). \end{aligned} \tag{B.3}$$

B.2 RF Method

B.2.1 Statistical results of Sole RF Model for Consumer Confidence Index

Table B.7: Coefficients of Basis Functions obtained with NN-MARS model

Number of trees	500
Mean of squared residuals	7.2606
%Var explained	84.7200

Tree	MSE	%Var(y)
100	7.4560	15.6900
200	7.4810	15.7400
300	7.660	16.1200
400	7.7120	16.2300
500	7.6940	16.1900

Here, in Table B.7, number of variability is highest when the number of tree is 500, afterwards, it started to decrease, thus it is stopped at the $tree = 500$. When the number of tree is taken as $\%84.72$ of variance is explained.

Table B.8: Incremental MSE values of variables of RF Model

input values	% IncMSE
UN	20.0960
CPI	42.0310
USDTRYINDEX	27.2640

According to the Table B.8, since higher values of Incremental MSE show us that a variable has a stronger impact on reducing the MSE, CPI has the highest effect.

B.2.2 Statistical results of MARS-RF Model for Consumer Confidence Index

When the number of tree is taken as 500, $\%87.67$ of variance is explained as can be seen from the Table B.9.

Table B.9: Statistical Results of MARS-RF Model

Number of trees	500
Mean of squared residuals	5.8580
%Var explained	87.6700

Tree	MSE	%Var(y)
100	5.9890	12.6000
200	5.9870	12.6000
300	6.0320	12.6900
400	5.8650	12.3400
500	5.8580	12.3300

Table B.10: Statistical Results of NN-RF Model

Number of trees	500
Mean of squared residuals	6.0193
%Var explained	87.3400

Tree	MSE	%Var(y)
100	6.4790	13.6300
200	6.2620	13.1800
300	6.0500	12.7300
400	6.0660	12.7600
500	6.0190	12.6600

B.2.3 Statistical results of NN-RF Model for Consumer Confidence Index

According to the Table B.10, when the number of tree is taken as 500, $\%87.67$ of variance is explained.

B.3 NN Method

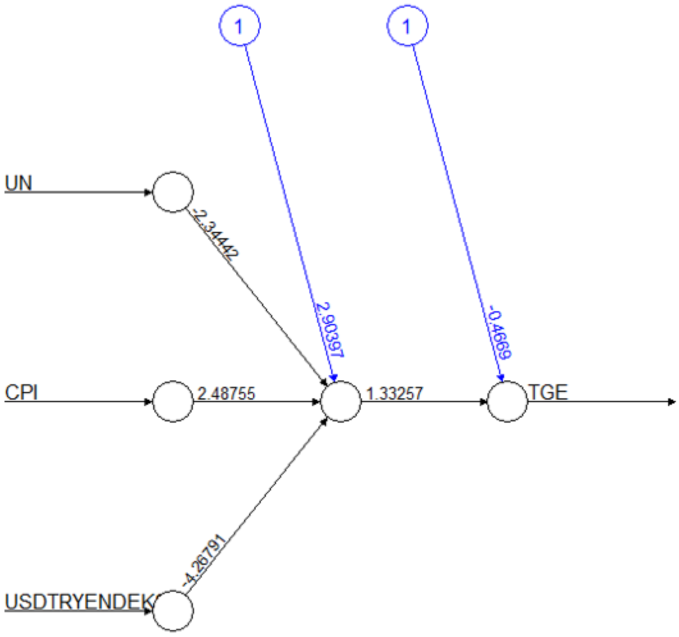
B.3.1 Statistical results of Sole NN Model for Consumer Confidence Index

For the results of sole NN Model, statistical values as in the following.

According to the estimated weights shown in Table B.11, only variable with weight is CCI.

Table B.11: Statistical Results of NN Model

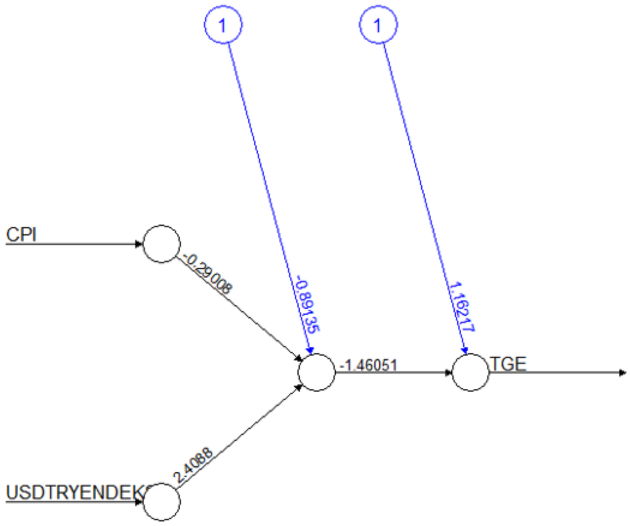
Estimated weights
1.0000000
-0.8073
0.8566
-1.46967



Error: 1.76177 Steps: 725

Figure B.2: Visualization of sole Neural Network Model

B.3.2 Statistical results of MARS-NN Model for Consumer Confidence Index



Error: 2.404282 Steps: 1379

Figure B.3: Visualization of MARS-NN Model.

In Table B.12, we see the estimated weights of the variables obtained by MARS-NN model whose graph is shown in Figure B.3. According to the estimated weights shown in Table B.12, only variable with weight 1 is CCI.

Table B.12: Statistical Results of MARS-NN Model

Estimated weights
0
0
1

B.3.3 Statistical results of RF-NN Model for Consumer Confidence Index

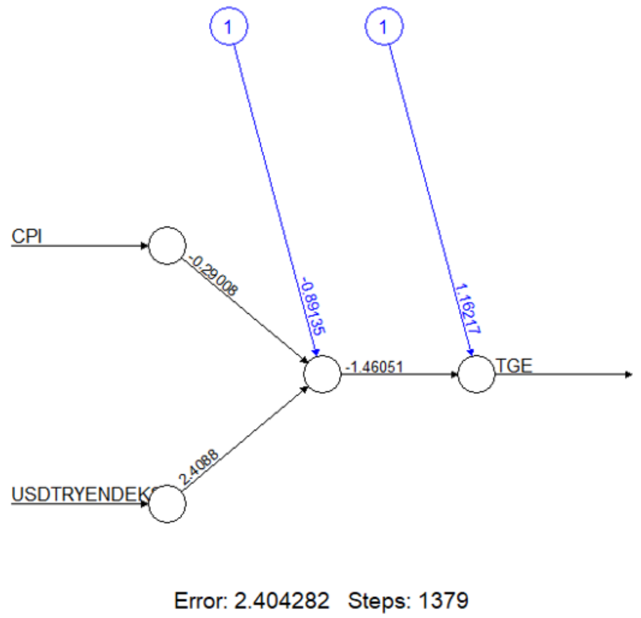


Figure B.4: Visualization of RF-NN Model.

Table B.13: Estimated weights of variables obtained with RF-NN Model

Estimated weights
0
0
1

In Table B.13, we see the estimated weights of the variables obtained by RF-NN model whose graph is shown in Figure B.4.

APPENDIX C

APPENDIX

C.1 Individual Statistical Results of the Variables used in Volatility Models

In this part, details of all the statistical tests, analysis are indicated with the tables and plots of all of the variables which are used in volatility models by employing at eviews are showed [65]. The ARCH, GARCH, and EGARCH models are preferred because of the advantages of capturing volatility clustering, prediction accuracy, and flexibility, as our indices may tend to include volatility dynamics [63].

C.1.1 Statistical Results of Investor Sentiment Index for the Volatility Models

At first, we begin with Investor Sentiment Index. This index has taken from the academic study of two researchers Baker and Wurgler in 2006 [11]. They update this index generally year by year. We use the latest updated version of this index which can be found in their academical website [98]. This index is up to sixth month of 2022. We employ this dataset from the beginning of 2000 to sixth month of 2022. In this thesis, we show the application of this index at first for the machine learning part and then for the volatility analysis. For the application of volatility analyis, at first we need to check the data from the statistical point of view. Here, we set sentiment data as an output and the other 12 proxies as an input. Firstly, we began to control the stationarity of the sentiment dataset. When we first check the raw data by using unit-root test, we obtain p-values higher than 0.05 as a result our dataset is not stationary. Thus, we take the first difference of sentiment data, defined as *return* and check again the unit-root test values. This time we observe p-values less than 0.05 as can be seen from Table C.1. As a result, sentiment data is stationary now.

Table C.1: The statistical performance of sentiment index.

Null Hypothesis: d(sentiment) has a unit root
Exogenous: Constant, Linear trend
Lag Lenth: 3 (Automatic-based on SIC, maxlag=15)

		t-statistic	Prob*
Augmented Dicky-Fuller test statistic		-13.9400	0.0000
Test critical values:	1% level	-3.9932	
	5% level	-3.4269	
	10% level	-3.1367	

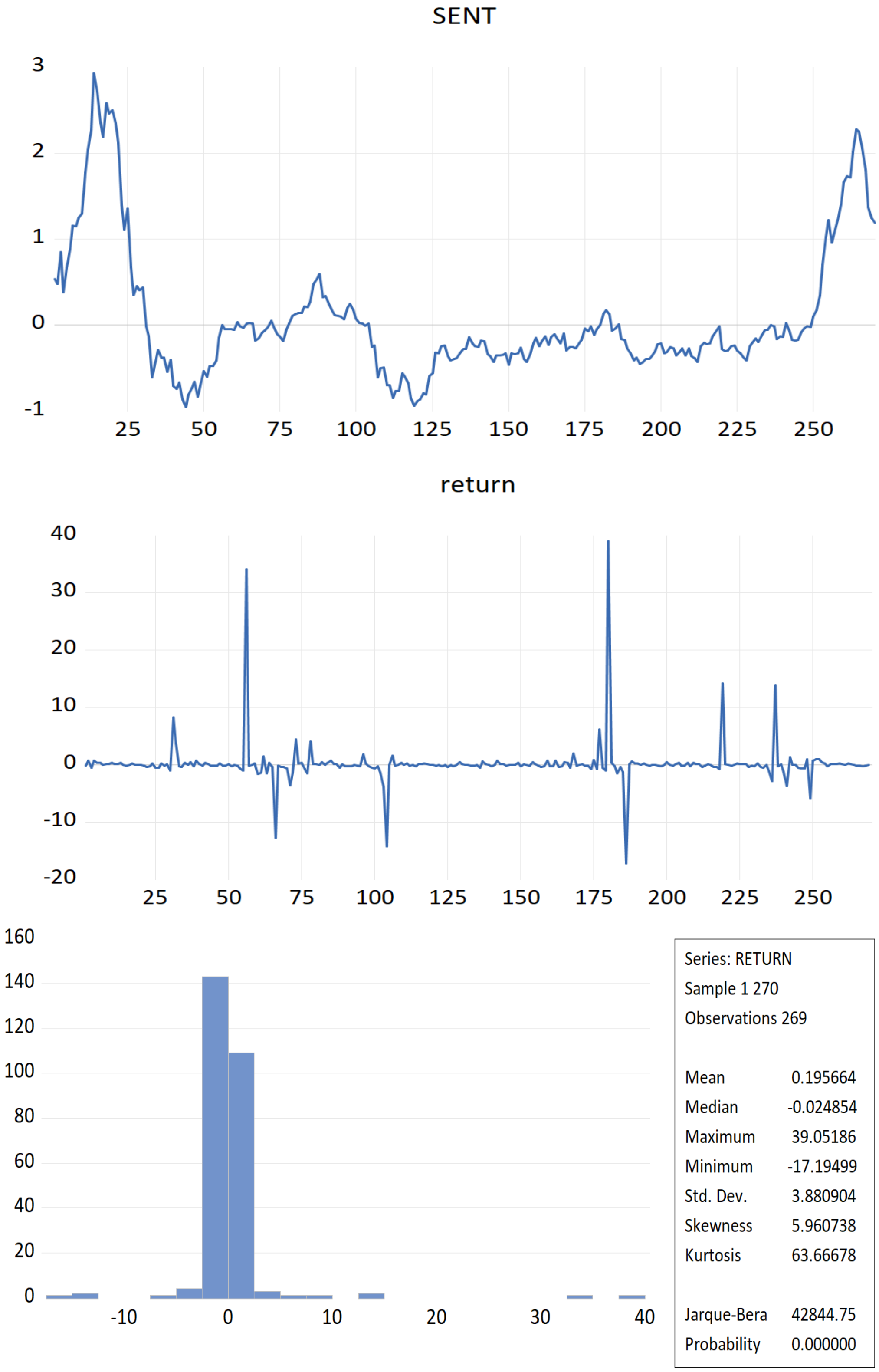
Variable	Coefficient	Std.Error	t-statistic	Prob.
d(sentiment(-1))	-3.0155	0.2164	-13.936	0.0000
d(sentiment(-1),2)	1.2149	0.1792	6.7778	0.0000
d(sentiment(-2),2)	0.5843	0.1222	4.7799	0.0000
d(sentiment(-3),2)	0.2133	0.0608	3.5062	0.0005
C	0.0036	0.5465	0.0065	0.9948
Trend("1")	-5.89E-05	0.0035	-0.0168	0.9866

S.E of regression	4.3283
Sum squared resid.	4833.3333
Log likelihood	-758.3733
F-statistic	207.6212
Prob(F-statistic)	0.0000
Mean dependent var	0.0019
S.D. dependent var	9.6085
Akaike info criterion	5.7906
Schwarz criterion	5.8719

In Table C.1, among from the information criterias, the lower the value of information criteria better the goodness of fit of the data.

According to the graph of returns, we observe outliers. Here, we get rid of this kind of heavily-tailed distribution because of the skewness (5.9607) and kurtosis (63.67) of the data which is explained in Chapter 4. Afterwards, we determine the lags by checking AR (Autoregressive) and MA (Moving Average) models. Since, we prefer less lags, we take the model for the AR(1) and MA(1), and then, we compute for the ARCH, GARCH and EGARCH models.

At first, we observe results for ARCH=1. Details of the statistical results are given below.



Series: RETURN
Sample 1 270
Observations 269

Mean	0.195664
Median	-0.024854
Maximum	39.05186
Minimum	-17.19499
Std. Dev.	3.880904
Skewness	5.960738
Kurtosis	63.66678
Jarque-Bera	42844.75
Probability	0.000000

Figure C.1: The graph of raw sentiment data, first difference taken and histogram of sentiment data .

Table C.2: The statistical results of lags and ARCH=1 of sentiment variable.

Variable	Coefficient	Std.Error	z-statistic	Prob.
C	0.0199	0.0131	1.5231	0.1277
AR(1)	0.5375	0.2739	1.9625	0.0497
MA(1)	-0.5840	0.2526	-2.3125	0.0208

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C	0.0938	0.00909	10.3231	0.0000
$Resid(-1)^2$	2.8625	0.1809	15.8204	0.0000

S.E of regression	1.0809
Sum squared resid.	309.6560
Log likelihood	-259.5790
Mean dependent var	-0.01258
S.D. dependent var	1.0751
Akaike info criterion	1.9744
Schwarz criterion	2.0414

Afterwards, we check for the GARCH and EGARCH models. Details of the statistical results are given below.

Table C.3: The statistical results of lags and GARCH=1 of sentiment variable.

Variable	Coefficient	Std.Error	z-statistic	Prob.
C	-0.0157	0.0107	-1.4628	0.1435
AR(1)	-0.7100	0.1146	-6.1932	0.0000
MA(1)	0.7604	0.1026	7.4059	0.0000

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C	0.0523	0.0109	4.7732	0.0000
$Resid(-1)^2$	2.9898	0.2004	14.9175	0.0000
GARCH(-1)	0.0608	0.01405	4.3418	0.0000

S.E of regression	1.0744
Sum squared resid.	305.8999
Log likelihood	-258.5265
Mean dependent var	-0.0125
S.D. dependent var	1.0751
Akaike info criterion	1.9740
Schwarz criterion	2.0548

Here, for the model specification, ARMA(1,1) model is constructed as follows:

$$SENT_t = 0.00199 + 0.5375SENT_{t-1} - 0.584\varepsilon_{t-1} + \varepsilon_t. \tag{C.1}$$

Table C.4: The statistical results of lags and EGARCH=1 of sentiment variable.

Variable	Coefficient	Std.Error	z-statistic	Prob.
C	-0.0485	0.0195	-2.4896	0.0128
AR(1)	0.5870	0.2031	2.8902	0.0038
MA(1)	-0.6445	0.1701	-3.7879	0.0002

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C(4)	-0.6805	0.0400	-16.9952	0.0000
C(5)	0.8760	0.0509	17.1998	0.0000
C(6)	-0.7878	0.0554	-14.2220	0.0000
C(7)	0.8618	0.0196	43.9294	0.0000

S.E of regression	1.0818
Sum squared resid.	310.1683
Log likelihood	-244.7480
Durbin-Watson stat	1.8198
Mean dependent var	-0.0126
S.D. dependent var	1.0751
Akaike info criterion	1.8787
Schwarz criterion	1.9725

where ϵ is the error term and coefficients are statistically significant. For the variance equation below, $\hat{h}$ represents the variance and $\hat{u}_{t-1})^2$ ARCH effect. Here, ARCH effect is statistically significant since its p-value is less than 0.05.

$$\hat{h}_t = 0.0938 + 2.8625(\hat{u}_{t-1})^2 - 0.37. \tag{C.2}$$

C.1.2 Statistical Results of Consumer Confidence Index for the Volatility Models

Secondly, other main index that we concentrate on is the Consumer Confidence Index (CCI). These data is taken from the Turkish Statistical Institute (TURKSTAT) [2]. Here, we use this dataset from the beginning of 2005 to second month of 2022. We show the application of indexes at first for the machine learning part and then for the volatility analysis. For the application of volatility analyis, similarly as we did in the sentiment index, at first we control the data from the statistical point of view. Therefore, we start to check the stationarity of the CCI dataset. When we first search for the raw data by using unit-root test, we obtain p-values higher than 0.05 as a result our dataset is not stationary. Hence, we take the first difference of CCI data, defined as *return* and check again the unit-root test values. This time we observe p-values less than 0.05 as can be seen from Table C.5. As a result, CCI data is stationary.

When we look at the plot of returns, we observe some outliers. Here, we eliminate this heavily-tailed distribution because of the skewness (-0.0974) and kurtosis (3.9935) of the data which is explained in Chapter 4. Afterwards, we determine the lags by calculating AR (Autoregressive) and MA (Moving Average) models. Since, we prefer less lags, we take the model for the AR(2) and MA(1), and then, we compute for the ARCH, GARCH and EGARCH models. At first, we observe results for ARCH=1. Details of the statistical results as in the following.

Table C.5: The statistical performance of consumer confidence index variable.

Null Hypothesis: d(cci) has a unit root
Exogenous: none
Lag Lenth: 0 (Automatic-based on SIC, maxlag=14)

		t-statistic	Prob*
Augmented Dicky-Fuller test statistic		-13.4750	0.0000
Test critical values:	1% level	-2.5764	
	5% level	-1.9424	
	10% level	-1.6156	

Variable	Coefficient	Std.Error	t-statistic	Prob.
d(cci(-1))	-0.9466	0.0705	-13.4750	0.0000

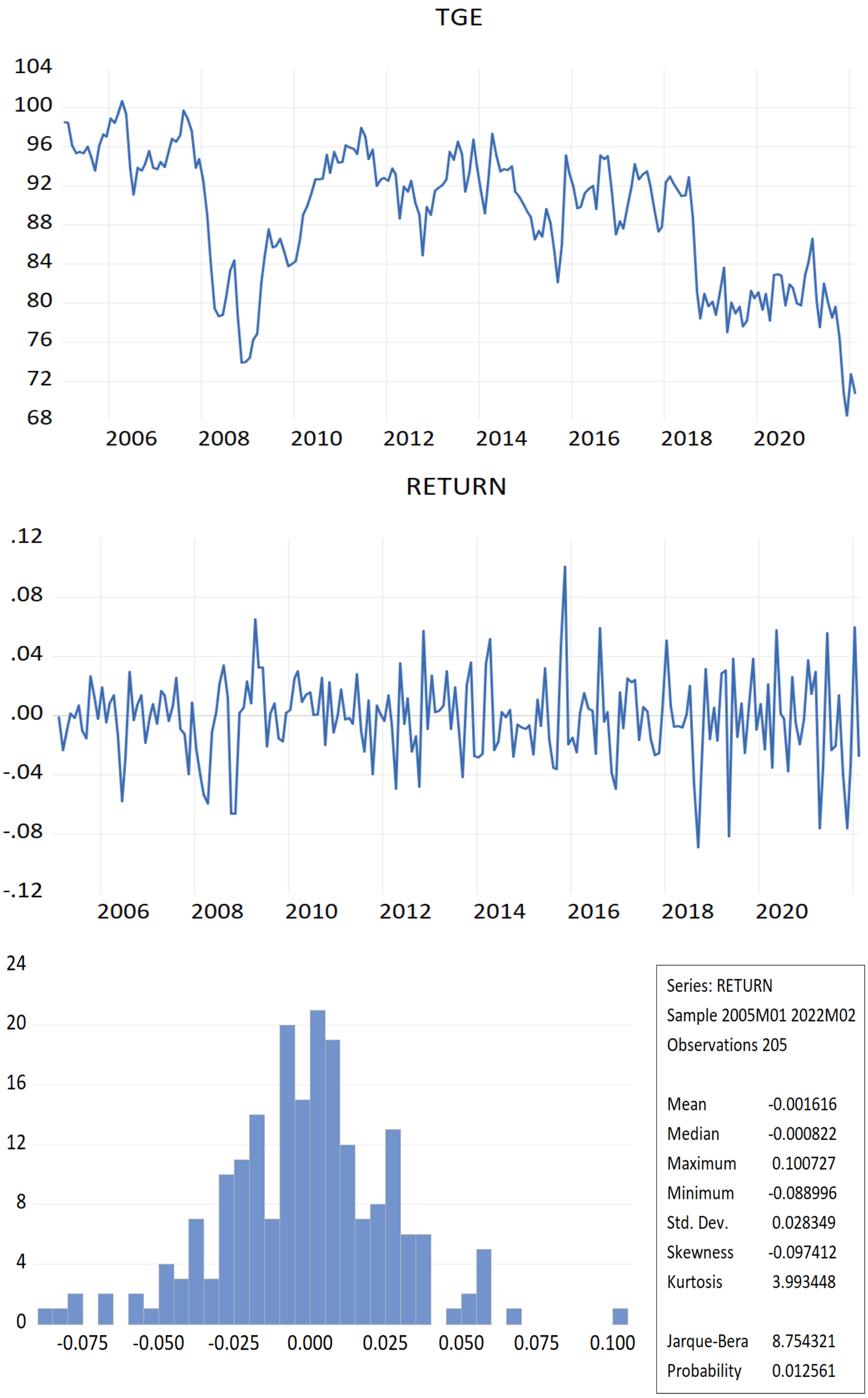
S.E of regression	0.0284
Sum squared resid.	437.3780
Log likelihood	-758.3700
Mean dependent var	-0.0001
S.D. dependent var	0.0391
Akaike info criterion	-4.2783
Schwarz criterion	-4.2619

For the model specification, ARMA(2,1) is constructed as follows:

$$CCI_t = -87.863 + 0.92826CCI_{t-2} + 0.3858\epsilon_{t-1} + \varepsilon_t. \tag{C.3}$$

Here ε is the error term and coefficients are statistically significant. For the variance equation, $\hat{h}$ indicates the variance and $(\hat{u}_{t-1})^2$ represents the ARCH effect. ARCH effect is statistically significant since its p-value is less than 0.05:

$$\hat{h}_t = 3.70594 + 2.9898(\hat{u}_{t-1})^2. \tag{C.4}$$



Series: RETURN
Sample 2005M01 2022M02
Observations 205

Mean	-0.001616
Median	-0.000822
Maximum	0.100727
Minimum	-0.088996
Std. Dev.	0.028349
Skewness	-0.097412
Kurtosis	3.993448
Jarque-Bera	8.754321
Probability	0.012561

Figure C.2: The graph of raw consumer confidence data, first difference taken and histogram of consumer confidence data.

Table C.6: The statistical results of lags and ARCH=1 of consumer confidence variable.

Variable	Coefficient	Std.Error	z-statistic	Prob.
C	-87.8630	4.0728	-21.5730	0.0000
AR(2)	0.9282	0.0458	20.2420	0.0000
MA(1)	0.9512	0.0358	26.5148	0.0000

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C	3.7059	0.5645	6.5644	0.0000
$Resid(-1)^2$	0.3858	0.1374	2.8080	0.0050

S.E of regression	2.3877
Sum squared resid.	1140.2400
Log likelihood	-455.9780
Mean dependent var	-88.7726
S.D. dependent var	6.8609
Akaike info criterion	4.5416
Schwarz criterion	6.4232

We continue by checking for the GARCH and EGARCH models. Details of the statistical results are given below. For the mean equation, ARMA(2,1) is constructed

Table C.7: The statistical results of lags and EGARCH=2 of consumer confidence index variable.

Variable	Coefficient	Std.Error	z-statistic	Prob.
C	-71.3926	13.0821	-5.4572	0.0000
AR(2)	0.9826	1.67E-14	5.89E+13	0.0000
MA(1)	0.9893	1.03E-13	9.64E+12	0.0000

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C(4)	0.4815	0.0623	7.7275	0.0000
C(5)	-0.3668	0.0720	-5.0883	0.0000
C(6)	0.2329	0.0596	3.9066	0.0001
C(7)	0.5054	0.0444	11.3720	0.0000
C(8)	0.3836	0.0397	9.6439	0.0000

S.E of regression	2.3970
Sum squared resid.	1149.1520
Log likelihood	-448.3250
Mean dependent var	-88.7726
S.D. dependent var	6.8609
Akaike info criterion	4.4958

as follows:

$$CCI_t = -71.3926 + 0.98268CCI_{t-2} + 0.9893\varepsilon_{t-1} + \varepsilon_t, \tag{C.5}$$

where ϵ is the error term and coefficients are statistically significant. For the variance Equation below, $\hat{h}$ shows the variance and $(\hat{u}_{t-1})^2$ demonstrates the ARCH effect. Here, EGARCH effect is statistically significant since its p-value is less than 0.05. On the other hand, in our variance equation C.6, there are both ARCH and GARCH terms. They are all statistically significant. Besides, according to the coefficients of each term, we can say that there is negative relation between the past variance and the recent variance in absolute value from the ARCH term -0.37. For the leverage effect size, we see that the term is 0.233, since this is positive, we can say that the good news tend to increase volatility more than bad news. For the GARCH terms, they are also positive and statistically significant, past volatility helps to forecast future volatility.

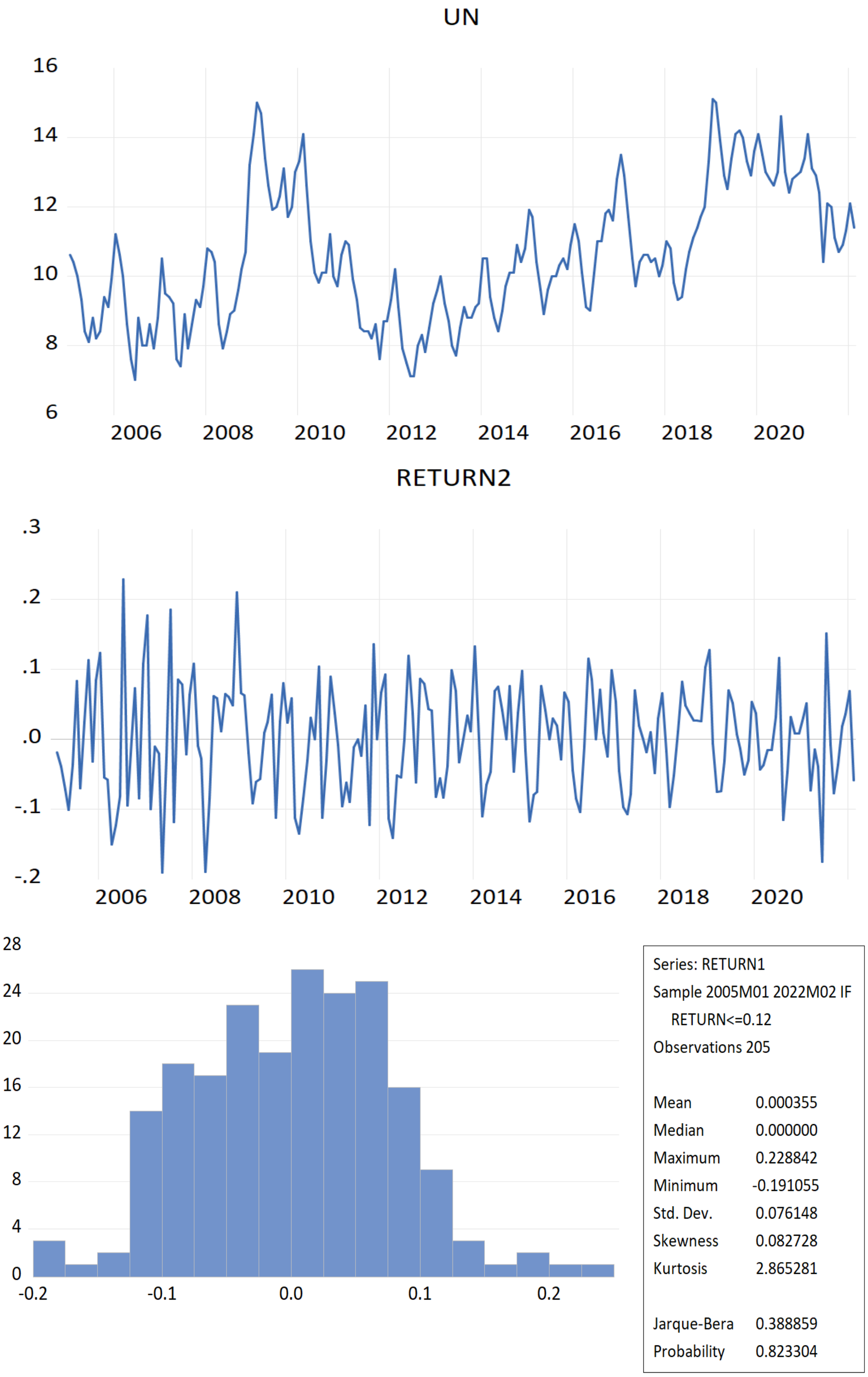
$$\log(h_t) = -0.4815 - 0.37\left|\frac{h_{t-1}}{\sqrt{h_{t-1}}}\right| + 0.233\frac{h_{t-1}}{\sqrt{h_{t-1}}} + 0.5054\log(h_{t-1}) + 0.3836\log(h_{t-2}). \tag{C.6}$$

C.1.3 Statistical Results of Unemployment Index for the Volatility Models

While we were searching for the volatility case of CCI, we aim to check the volatility behavior of other macroeconomic variables (UN, CPI, USD/TRY) as we also use in the machine learning applications. Here, we continue to show the statistically details of Unemployment rate (UN). This data is taken from the Turkish Statistical Institute (TURKSTAT) [2]. We employ this dataset from the beginning of 2005 to second month of 2022 as we did in CCI dataset. Firstly, we control the data from the statistical point of view. That is why, we start to check the stationarity of the UN dataset. When we first search for the raw data by using unit-root test, we obtain p-values higher than 0.05 as a result our dataset is not stationary. Hence, we take the first difference of UN data, defined as *return2* and check again the unit-root test values. This time we observe p-values less than 0.05 as can be seen from Table C.8. Finally, we obtain UN data as stationary.

For the mean equation, ARMA(1,1) is constructed as follows:

$$UN_t = -10.5538 + 0.8595UN_{t-1} + 0.2986\varepsilon_{t-1} + \varepsilon_t. \tag{C.7}$$



```
Series: RETURN1
Sample 2005M01 2022M02 IF
    RETURN<=0.12
Observations 205

Mean          0.000355
Median        0.000000
Maximum       0.228842
Minimum      -0.191055
Std. Dev.     0.076148
Skewness      0.082728
Kurtosis      2.865281

Jarque-Bera   0.388859
Probability   0.823304
```

Figure C.3: The graph of raw unemployment data, first difference taken and histogram of unemployment data.

Table C.8: The statistical performance of unemployment variable.

Null Hypothesis: un has a unit root
Exogenous: constant
Lag Lenth: 0 (Automatic-based on SIC, maxlag=14)

		t-statistic	Prob*
Augmented Dicky-Fuller test statistic		-2.8525	0.053
Test critical values:	1% level	-3.4643	
	5% level	-2.8764	
	10% level	-2.5747	

Variable	Coefficient	Std.Error	z-statistic	Prob.
C	-10.5538	0.4963	-21.2674	0.0000
AR(1)	0.8595	0.0399	21.4953	0.0000
MA(1)	0.2986	0.072	4.1466	0.0000
σ^2	0.5765	0.0572	10.069	0.0000

S.E of regression	0.7668
Sum squared resid.	118.188
Log likelihood	-235.379
F-statistic	356.1740
prob(F-statistic)	0.0000
Mean dependent var	-10.5368
S.D. dependent var	1.9129
Akaike info criterion	2.3354
Schwarz criterion	2.4002

where ε is the error term and coefficients are statistically significant. There is no ARCH or GARCH effect for the Unemployment variable. Therefore, UN data is not exhibit the specific characteristics captured by ARCH and GARCH models.

C.1.4 Statistical Results of Consumer Price Index for the Volatility Models

Among from our macroaconomic variables, we also check the volatility case of Consumer Price Index (CPI). This data is taken from the Turkish Statistical Institute (TÜİK) [2]. Same as in the other variables, we employ this dataset from the beginning of 2005 to second month of 2022. At first, we again research the data from the statistical point of view. Thus, we begin to control the stationarity of the CPI dataset. When we first search for the raw data by using unit-root test, we obtain p-values higher than 0.05 as a result our dataset is not stationary. Hence, we take the first difference of CPI data, defined as *returns* and check again the unit-root test values. This time we observe p-values less than 0.05 as can be seen from table **??**. Afterwards, we obtain CPI data as stationary. Afterwards, we determine the lags by checking AR (Autoregressive) and MA (Moving Average) models. Since, we prefer less lags, we take the model for the AR(1) and MA(2), and then, we compute for the ARCH, GARCH and EGARCH models. Firstly, we observe results for ARCH=1. Details of the statistical results are given below.

Table C.9: The statistical performance of consumer price index variable.

Null Hypothesis: d(cpi) has a unit root
Exogenous: constant, linear trend
Lag Lenth: 1 (Automatic-based on SIC, maxlag=14)

		t-statistic	Prob*
Augmented Dicky-Fuller test statistic		-7.7966	0.0000
Test critical values:	1% level	-4.0043	
	5% level	-3.4323	
	10% level	-3.1399	

Variable	Coefficient	Std.Error	t-statistic	Prob.
cpi(-1)	-0.5789	0.0742	-7.7966	0.0000
d(cpi(-1))	0.2073	0.0800	2.5891	0.0103
c	0.0131	0.0016	0.7909	0.4299

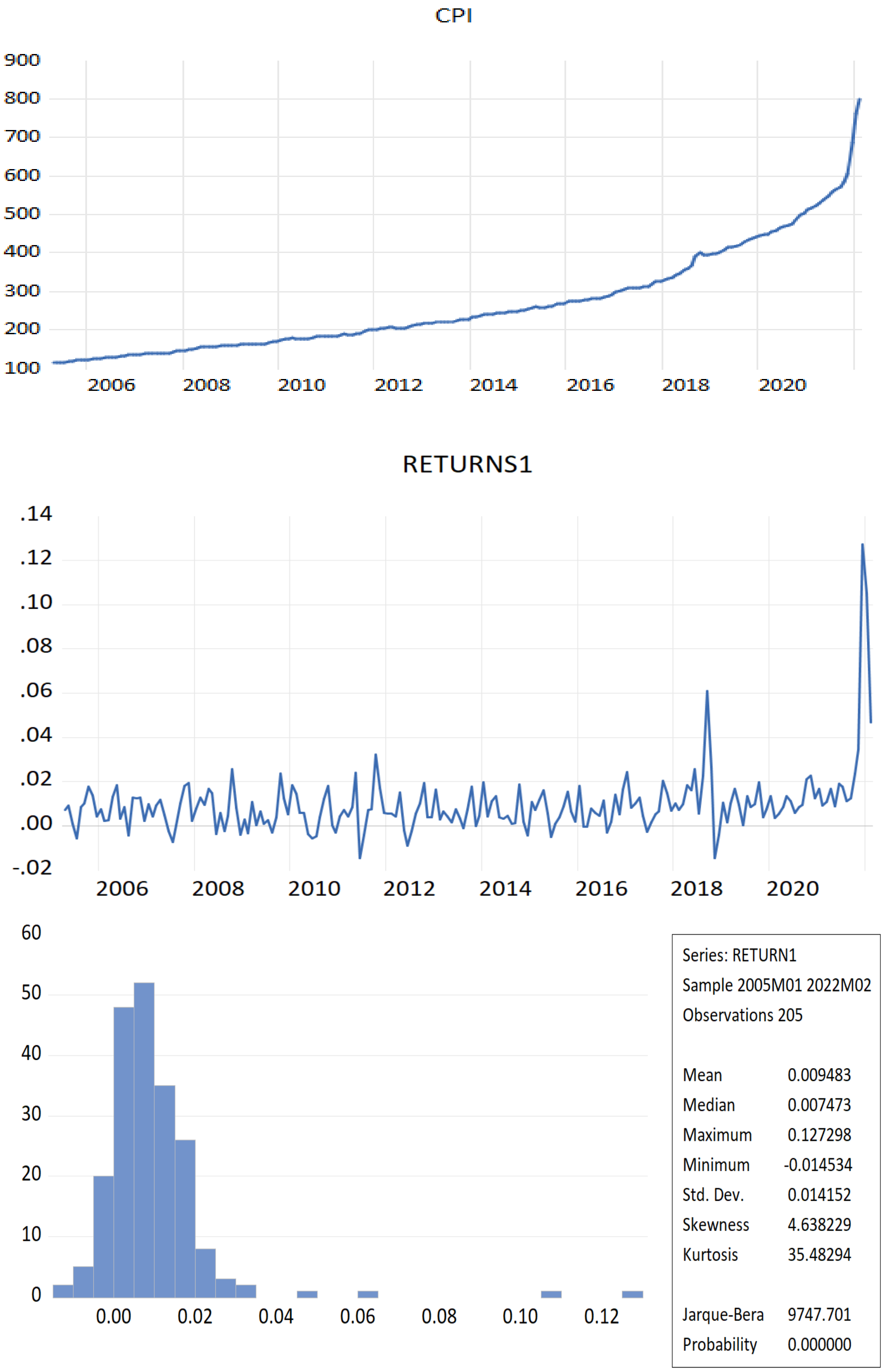
S.E of regression	0.0115
Sum squared resid.	0.0259
Log likelihood	614.6860
F-statistic	22.2680
Prob(F-statistics)	0.0000
Mean dependent var	0.0002
S.D. dependent var	0.0131
Akaike info criterion	-6.0765
Schwarz criterion	-6.0107

For the model specification, ARMA(1,2) is constructed as follows:

$$CPI_t = 0.00769 + 0.3647 CPI_{t-1} - 0.3637 \varepsilon_{t-2} + \varepsilon_t, \tag{C.8}$$

where ε is the error term and coefficients are statistically significant. For the variance equation below, $\hat{h}$ represents the variance and $\hat{u}_{t-1})^2$ shows the ARCH effect. ARCH effect is statistically significant since its p-value is less than 0.05:

$$\hat{h}_t = 4.43E - 05 + 0.7955(\hat{u}_{t-1})^2. \tag{C.9}$$



Series: RETURN1
Sample 2005M01 2022M02
Observations 205

Mean	0.009483
Median	0.007473
Maximum	0.127298
Minimum	-0.014534
Std. Dev.	0.014152
Skewness	4.638229
Kurtosis	35.48294
Jarque-Bera	9747.701
Probability	0.000000

Figure C.4: The graph of raw consumer price data, first difference taken and histogram of consumer price data.

Table C.10: The statistical results of lags and ARCH=1 of consumer price index variable.

Variable	Coefficient	Std.Error	z-statistic	Prob.
C	0.0077	0.0006	12.8440	0.0000
AR(1)	0.3647	0.0089	4.0634	0.0000
MA(2)	-0.3637	0.0628	-5.7892	0.0000

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C	4.43E-05	6.47E-06	6.8495	0.0000
$Resid(-1)^2$	0.7955	0.0893	8.9136	0.0000

S.E of regression	0.0129
Sum squared resid.	0.0332
Log likelihood	665.8644
Mean dependent var	0.0095
S.D. dependent var	0.0142
Akaike info criterion	-6.5432
Schwarz criterion	-6.4613

Afterwards, we check for the GARCH and EGARCH models. Details of the statistical results are given below. For the model specification, ARMA(1,2) is constructed as

Table C.11: The statistical results of lags and GARCH=1 of consumer price index variable.

Variable	Coefficient	Std.Error	z-statistic	Prob.
C	0.0077	0.0005	15.2323	0.0000
AR(1)	0.3496	0.1014	3.4458	0.0006
MA(2)	-0.4312	0.0601	-7.1706	0.000

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C	2.44E-05	9.10E-06	2.6838	0.0073
$Resid(-1)^2$	0.7309	0.1115	6.5576	0.0000
Garch(-1)	0.2397	0.1146	2.0920	0.0036

S.E of regression	0.0133
Sum squared resid.	0.0354
Log likelihood	675.3687
Durbin-Watson stat	1.0305
Mean dependent var	0.00953
S.D. dependent var	0.0142
Akaike info criterion	-6.5624
Schwarz criterion	-6.4648

follows:

$$CPI_t = 0.0077 + 0.3496CPI_{t-1} - 0.4312\varepsilon_{t-2} + \varepsilon_t. \tag{C.10}$$

where ε is the error term and coefficients are statistically significant. For the variance equation below, $\hat{h}$ defines the variance and $(\hat{u}_{t-1})^2$ represents ARCH effect. Here, ARCH and GARCH effect is statistically significant since its p-value is less than 0.05:

$$\hat{h}_t = 2.44E - 05 + 0.2397(\hat{h}_{t-1}) + 0.7309(\hat{u}_{t-1})^2. \tag{C.11}$$

Lastly, we control for the EGARCH model. Details of the statistical results are given below. EGARCH effect is statistically significant since its p-value is less than 0.05. On the other hand, in our variance equation C.12, there are both ARCH and GARCH terms. They are all statistically significant. Furthermore, according to the coefficients of each term, we can say that there is positive relation between the past variance and the recent variance in absolute value from the ARCH term 0.3676. For the leverage effect size, we see that the term is 0.5699, since this is positive, we can say that the good news tend to increase volatility more than bad news. For the GARCH term, it is also positive and statistically significant, past volatility helps to forecast future volatility.

$$\log(h_t) = -3.8997 + 0.3676\left|\frac{h_{t-1}}{\sqrt{h_{t-1}}}\right| + 0.5699\frac{h_{t-1}}{\sqrt{h_{t-1}}} + 0.6234\log(h_{t-1}). \tag{C.12}$$

C.1.5 Statistical Results of USD/TRY Currency Index for the Volatility Models

Finally, we control the volatility case of USD/TRY Currency Index. This data is taken from the Turkish Central Bank EVDS Data Central [1]. We also take this dataset from the beginning of 2005 to second month of 2022. We begin by checking the data from the statistical point of view. For that reason, we begin to control the stationarity of the USD/TRY dataset. When we first search for the raw data by using unit-root test, we obtain p-values higher than 0.05 as a result our dataset is not stationary. Hence, we take the first difference of USD/TRY data, defined as *rusdtry* and check again the unit-root test values. This time we observe p-values less than 0.05 as can be seen from Table C.13. Afterwards, we obtain USD/TRY data as stationary. For the

Table C.12: The statistical results of lags and EGARCH=1 of consumer price index variable.

Variable	Coefficient	Std.Error	z-statistic	Prob.
C	0.0089	0.0009	9.3338	0.0000
AR(1)	0.5363	0.0768	6.9772	0.0000
MA(2)	-0.4469	0.0567	-7.8746	0.0000

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C(4)	-3.8997	1.2297	-3.1142	0.0018
C(5)	0.3676	0.1737	2.1187	0.00341
C(6)	0.5699	0.1137	5.0117	0.0000
C(7)	0.6234	0.1273	4.8984	0.0000

S.E of regression	0.0120
Sum squared resid.	0.0292
Log likelihood	679.2630
Mean dependent var	0.00953
S.D. dependent var	0.0142
Akaike info criterion	-6.5908
Schwarz criterion	-6.4769

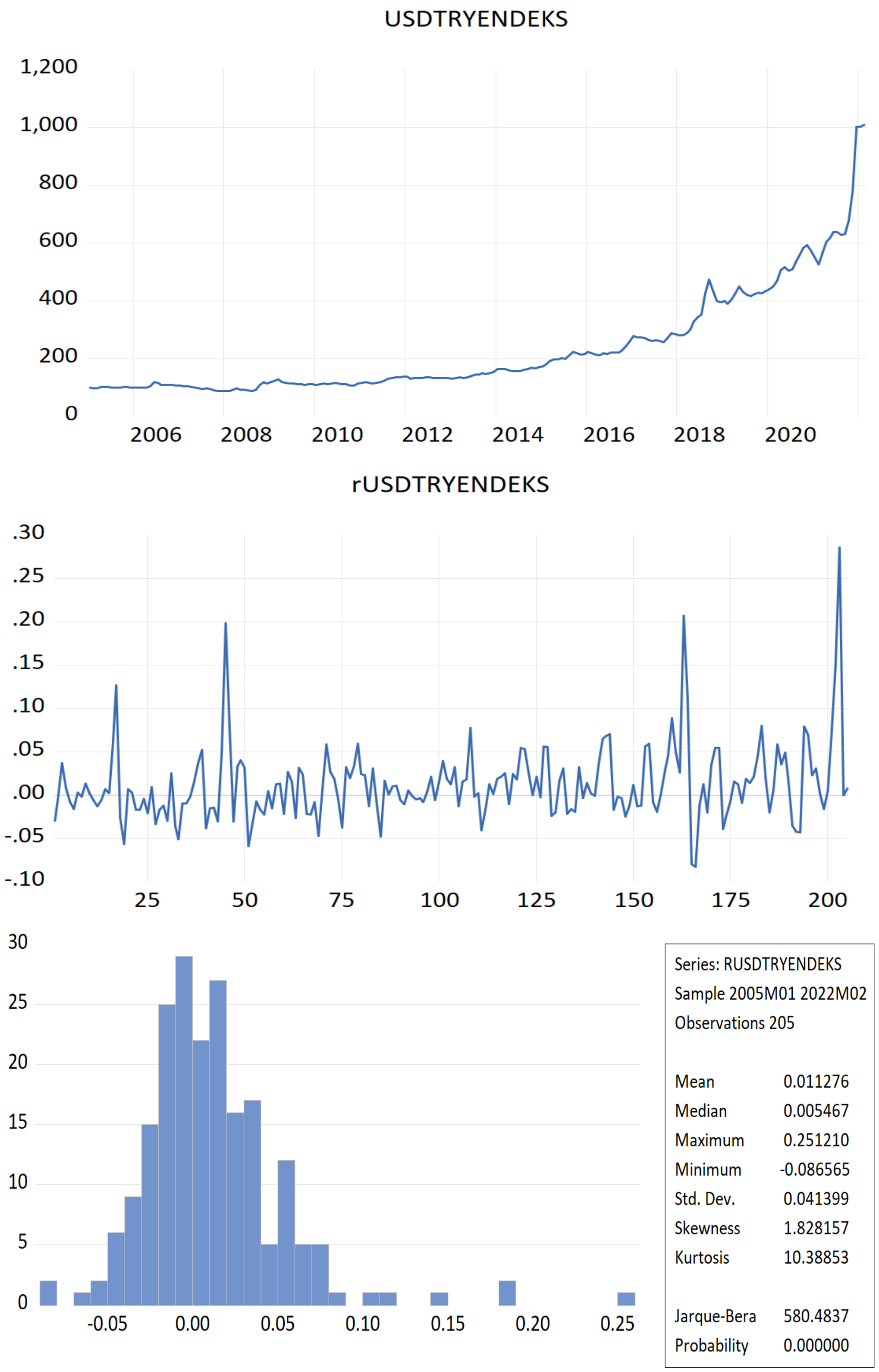
model specification, ARMA(1,1) is constructed as follows:

$$USDTRY_t = 0.0041 + 0.3565USDTRY_{t-1} - 0.2994\varepsilon_{t-1} + \varepsilon_t. \tag{C.13}$$

where ε is the error term and coefficients are statistically significant. For the variance equation below, $\hat{h}$ represents the variance and $(\hat{u}_{t-1})^2$ ARCH effect. Here, ARCH effect is statistically significant since its p-value is less than 0.05.

$$\hat{h}_t = 0.00061 + 0.7745(\hat{u}_{t-1})^2. \tag{C.14}$$

Later, we check for the GARCH and EGARCH models. Details of the statistical results are given below. EGARCH effect is statistically significant since its p-value is less than 0.05. On the other hand, in our variance equation C.15, there are both ARCH and GARCH terms. They are all statistically significant. Furthermore, according to the coefficients of each term, we can say that there is positive relation between the past variance and the recent variance in absolute value from the ARCH term 0.4474. For the leverage effect size, we see that the term is 0.3666, since this is positive, we can say that the good news tend to increase volatility more than bad news. For the GARCH term, it is also positive and statistically significant, past volatility



Series: RUSDTRYENDEKS
Sample 2005M01 2022M02
Observations 205

Mean	0.011276
Median	0.005467
Maximum	0.251210
Minimum	-0.086565
Std. Dev.	0.041399
Skewness	1.828157
Kurtosis	10.38853
Jarque-Bera	580.4837
Probability	0.000000

Figure C.5: The graph of raw usd/try index data, first difference taken and histogram of usd/try index data.

Table C.13: The statistical performance of usd/try index variable.

Null Hypothesis: d(usdtryindex) has a unit root
Exogenous: constant
Lag Lenth: 1 (Automatic-based on SIC, maxlag=14)

		t-statistic	Prob*
Augmented Dicky-Fuller test statistic		-10.2861	0.0000
Test critical values:	1% level	-3.4626	
	5% level	-2.8756	
	10% level	-2.5744	

Variable	Coefficient	Std.Error	t-statistic	Prob.
usdtryindex(-1)	-0.7884	0.0766	-10.2861	0.0000
d(usdtryindex(-1))	0.2524	0.0683	3.6969	0.0003
c	0.0901	0.0027	3.2773	0.0012

S.E of regression	0.0037
Sum squared resid.	0.2808
Log likelihood	380.1549
F-statistic	55.9826
Prob(F-statistics)	0.0000
Mean dependent var	5.66E-05
S.D. dependent var	0.0466
Akaike info criterion	-3.7158
Schwarz criterion	-3.6668

Table C.14: The statistical results of lags and ARCH=1 of usd/try index variable.

Variable	Coefficient	Std.Error	z-statistic	Prob.
C	0.0041	0.0025	1.6258	0.1040
AR(1)	0.3565	0.0933	3.8209	0.0001
MA(2)	-0.2994	0.0704	-4.2521	0.0000

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C	0.0006	9.48E-05	6.4481	0.0000
$Resid(-1)^2$	0.7745	0.1689	4.5848	0.0000

S.E of regression	0.0383
Sum squared resid.	0.2943
Log likelihood	403.2480
Mean dependent var	0.0114
S.D. dependent var	0.0414
Akaike info criterion	-3.9044
Schwarz criterion	-3.8231

Table C.15: The statistical results of lags and EGARCH=1 of consumer price index variable.

Variable	Coefficient	Std.Error	z-statistic	Prob.
C	0.0094	0.0033	3.0993	0.0019
AR(1)	0.3979	0.1003	3.9669	0.0001
MA(2)	-0.2589	0.0950	-2.7288	0.0064

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C(4)	-2.9571	0.7815	-3.7841	0.0002
C(5)	0.4474	0.1095	4.0850	0.0000
C(6)	0.3666	0.0778	4.6483	0.0000
C(7)	0.6195	0.1135	5.459	0.0000

S.E of regression	0.0375
Sum squared resid.	0.2821
Log likelihood	410.8978
Mean dependent var	0.00114
S.D. dependent var	0.0041
Akaike info criterion	-3.9598
Schwarz criterion	-3.8459

helps to forecast future volatility.

$$\log(h_t) = -2.9571 + 0.4474\left|\frac{h_{t-1}}{\sqrt{h_{t-1}}}\right| + 0.3666\frac{h_{t-1}}{\sqrt{h_{t-1}}} + 0.6195\log(h_{t-1}). \quad \text{(C.15)}$$

C.2 Statistical Results of the Indexes used for the Multivariate Case in Volatility Models

In this part, we evaluate both investor sentiment and consumer confidence indexes as an output and other variables as in inputs.

C.2.1 Statistical Results of Investor Sentiment Index with Multiple Input Variables

We prefer to show investor sentiment index as an output and other proxies which are NYSE share turnover, the closed-end fund discount, the number, and average first-day returns on IPOs (initial public offerings), the dividend premium, the equity which shares new issues, indpro (industrial production index), consserv (nominal services consumption), consdur (nominal durables consumption), consnon (nominal nondurables consumption), cpi (consumer price index), and employ (employment) [11, 12, 47]. All these variables have taken from the academic study of two researchers Baker and Wurgler in 2006 [11]. They update this index generally year by year. We use the latest updated version of this index which can be found in their academical website [98]. This index is up to 6th month of 2022. We employ this dataset from the beginning of 2000 to sixth month of 2022. For the model specifi-

Table C.16: The statistical results of lags and ARCH=1 of sentiment index variable as an output and other variables as an inputs with Normal Distribution.

Variable	Coefficient	Std.Error	z-statistic	Prob.
cefd	0.0527	0.0119	-4.4289	0.0000
consdur	-10.1079	1.0278	-9.8337	0.0000
consnon	6.2342	2.1945	2.8407	0.0045
consserv	27.2229	5.5940	4.8664	0.0000
cpi	-85.7082	4.1626	-20.5899	0.0000
employ	-3.9964	7.6052	-0.5255	0.5992
indpro	-10.2707	2.8422	-3.6137	0.0003
nipo	-0.0874	0.0246	-3.5426	0.0004
pdnd	0.0042	0.0109	0.3759	0.7070
ripo	0.0008	0.0023	0.3721	0.7098
s	-3.6873	0.1682	-21.9235	0.0000
C	0.1657	0.0230	7.2261	0.0000

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C	0.1046	0.0228	4.5780	0.0000
$Resid(-1)^2$	15.1417	0.7506	20.1735	0.0000

S.E of regression	3.9508
Sum squared resid.	4011.4170
Log likelihood	-524.0060
Mean dependent var	0.1956
S.D. dependent var	3.8809
Akaike info criterion	4.0000
Schwarz criterion	4.1871

cation, mean equation is constructed as follows:

$$\begin{aligned}SENT_t &= 0.1657 - 0.0527CEFD_t - 10.108CONSDUR_t + 6.234CONSNON_t \\ &+ 27.223CONSSERV_t - 85.708CPI_t - 3.996EMPLOY_t - 10.27 \\ &INDPRO_t - 0.0874NIPO_t + 0.004PDND_t + 0.0008RIPO_t - 3.687S_t.\end{aligned} \tag{C.16}$$

Here, except for $employ$, $pdnd$ and $ripo$, the rest of the variables are statistically statistically significant. For the variance equation below, $\hat{h}$ represents the variance and $(\hat{u}_{t-1})^2$ ARCH effect. Here, ARCH effect is statistically significant since its p-value is less than 0.05.

$$\hat{h}_t = 0.1046 + 15.1417(\hat{u}_{t-1})^2. \tag{C.17}$$

Later, we check for the GARCH and EGARCH models. We obtain results for EGARCH model, details of the statistical results are given below. For the model specification, mean equation is constructed as follows:

$$\begin{aligned}SENT_t &= -0.0367 + 0.00827CEFD_t + 0.4584CONSDUR_t - 20.0377CONS- \\ &NON_t + 19.3469CONSSERV_t + 34.6150CPI_t - 65.0713EMPLOY_t - 37. \\ &568INDPRO_t + 0.1352NIPO_t + 0.0246PDND_t - 0.006RIPO_t + 1.949S_t.\end{aligned} \tag{C.18}$$

Here, except for $consdur$ and $ripo$, rest of the variables are statistically statistically significant. EGARCH effect is statistically significant since its p-value is less than 0.05. On the other hand, in our variance Equation C.19, there are both ARCH and GARCH terms. ARCH term is statistically significant, however, GARCH term is not statistically significant because of the p-value. Furthermore, according to the coefficients of each term, we can say that there is positive relation between the past variance and the current variance in absolute value from the ARCH term 3.1612. For the leverage effect size, we see that the term is -1.5175, since this is negative, we can say that the bad news tend to increase volatility more than good news. For the GARCH term, it is positive, but not statistically significant, past volatility cannot help to forecast future volatility.

$$\log(h_t) = -1.7393 + 3.1612\left|\frac{h_{t-1}}{\sqrt{h_{t-1}}}\right| - 1.5175\frac{h_{t-1}}{\sqrt{h_{t-1}}} + 0.0013\log(h_{t-1}). \tag{C.19}$$

Table C.17: The statistical results of lags and EGARCH=1 of sentiment index variable as an output and other variables as an inputs.

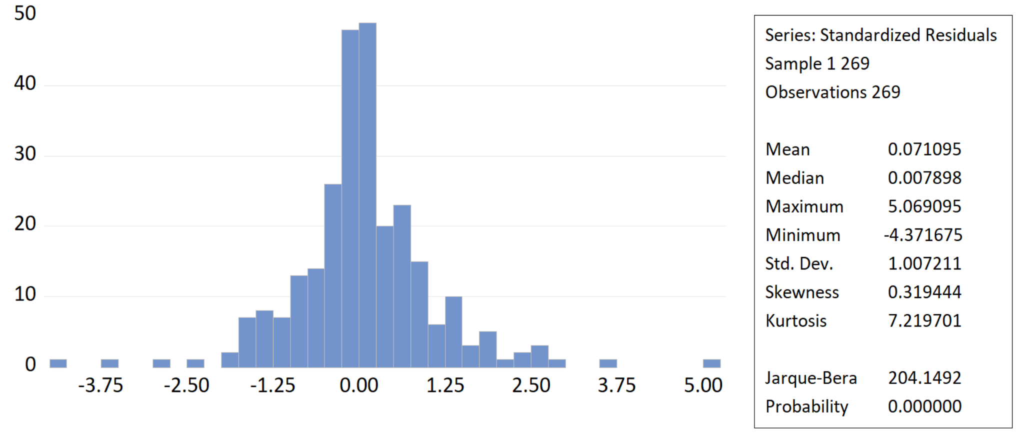
Variable	Coefficient	Std.Error	z-statistic	Prob.
cefd	0.0082	0.0020	4.0252	0.0001
consdur	0.4584	0.6236	0.7349	0.4624
consnon	-20.0377	1.7852	-11.2248	0.0000
consserv	19.3469	1.9821	9.7104	0.0000
cpi	34.6150	3.8826	8.9155	0.0000
employ	-65.0713	2.7864	-23.3533	0.0000
indpro	-37.5687	1.5361	-24.4565	0.0000
nipo	0.1352	0.0183	7.3778	0.0000
pdnd	0.0246	0.0064	3.8688	0.0001
ripo	-0.0063	0.0039	-1.5887	0.1121
s	1.9499	0.1378	14.1486	0.0000
C	-0.0367	0.0111	-3.3099	0.0009

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C(13)	-1.7393	0.1029	-16.8945	0.0000
C(14)	3.1612	0.1098	28.7861	0.0000
C(15)	-1.5175	0.0882	-17.2223	0.0000
C(16)	0.0013	0.0082	0.1648	0.8691

S.E of regression	4.1230
Sum squared resid.	4368.895
Log likelihood	-425.5763
Mean dependent var	0.1956
S.D. dependent var	3.8809
Akaike info criterion	3.2831
Schwarz criterion	3.4969

As shown in the above graphs C.9, C.10, C.11 and C.12, each of the labels which starts with an C define proxies. List of these is indicated in the following table:

We compute all these statistical results of sentiment index with multiple inputs by employing Normal distribution. Now, we estimate all these results also with Student-t distribution. According to the results, all of the variables have p-value is higher than 0.05. They are not statistically significant. Furthermore, ARCH effect is also not statistically significant since its p-value is higher than 0.05. Afterwards, we check for EGARCH effect as well. EGARCH effect is not statistically significant since its p-value is higher than 0.05. On the other hand, in our variance Equation C.19,



Series: Standardized Residuals
Sample 1 269
Observations 269

Mean	0.071095
Median	0.007898
Maximum	5.069095
Minimum	-4.371675
Std. Dev.	1.007211
Skewness	0.319444
Kurtosis	7.219701
Jarque-Bera	204.1492
Probability	0.000000

Figure C.6: Descriptive Statistics and Histogram of sentiment output variable and other input variables.

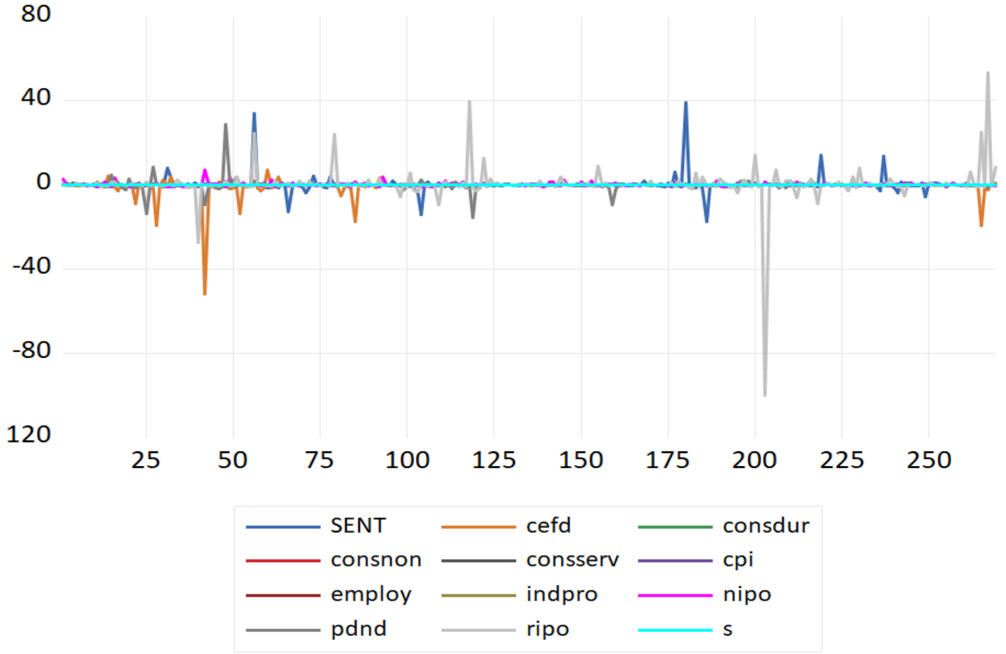


Figure C.7: The residuals of Sentiment variable and other proxies.

there are both ARCH and GARCH terms. ARCH term is not statistically significant, however, GARCH term is statistically significant because of the p-value. For the GARCH term, it is positive and past volatility helps to forecast future volatility. For the model specification, mean equation is constructed as follows:

$$\begin{aligned} SENT_t = -0.0040 - 0.0054CEFD_t + 0.2582CONSDUR_t + 1.7131CONSNON_t \\ - 4.4689CONSSERV_t - 6.8022CPI_t + 9.6905EMPLOY_t - 7.3516 \\ INDPRO_t - 0.0062NIPO_t - 0.0046PDND_t - 0.0010RIPO_t - 0.2462_t. \end{aligned} \tag{C.20}$$

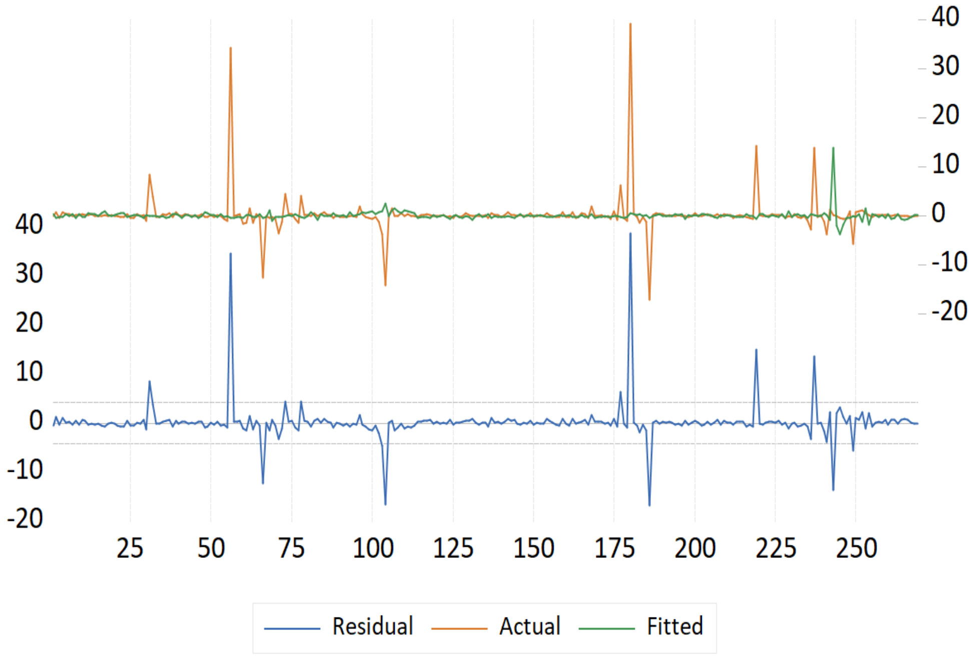


Figure C.8: The graph of actual and fitted residuals.

Table C.18: The Coefficient Labels.

Variable	Coefficient
cefd	C(1)
consdur	C(2)
consnon	C(3)
consserv	C(4)
cpi	C(5)
employ	C(6)
indpro	C(7)
nipo	C(8)
pdnd	C(9)
ripo	C(10)
s	C(11)
C	C(12)

In Equation C.20, sentiment has linear relation with nominal durables consumption (consdur) and nominal nondurables consumption (consnon) and inverse relation with rest of the other variables. In our variance equation C.21, there are both ARCH and GARCH terms. ARCH term is not statistically significant, however, GARCH term is statistically significant because of the p-value. For the GARCH term, it is positive

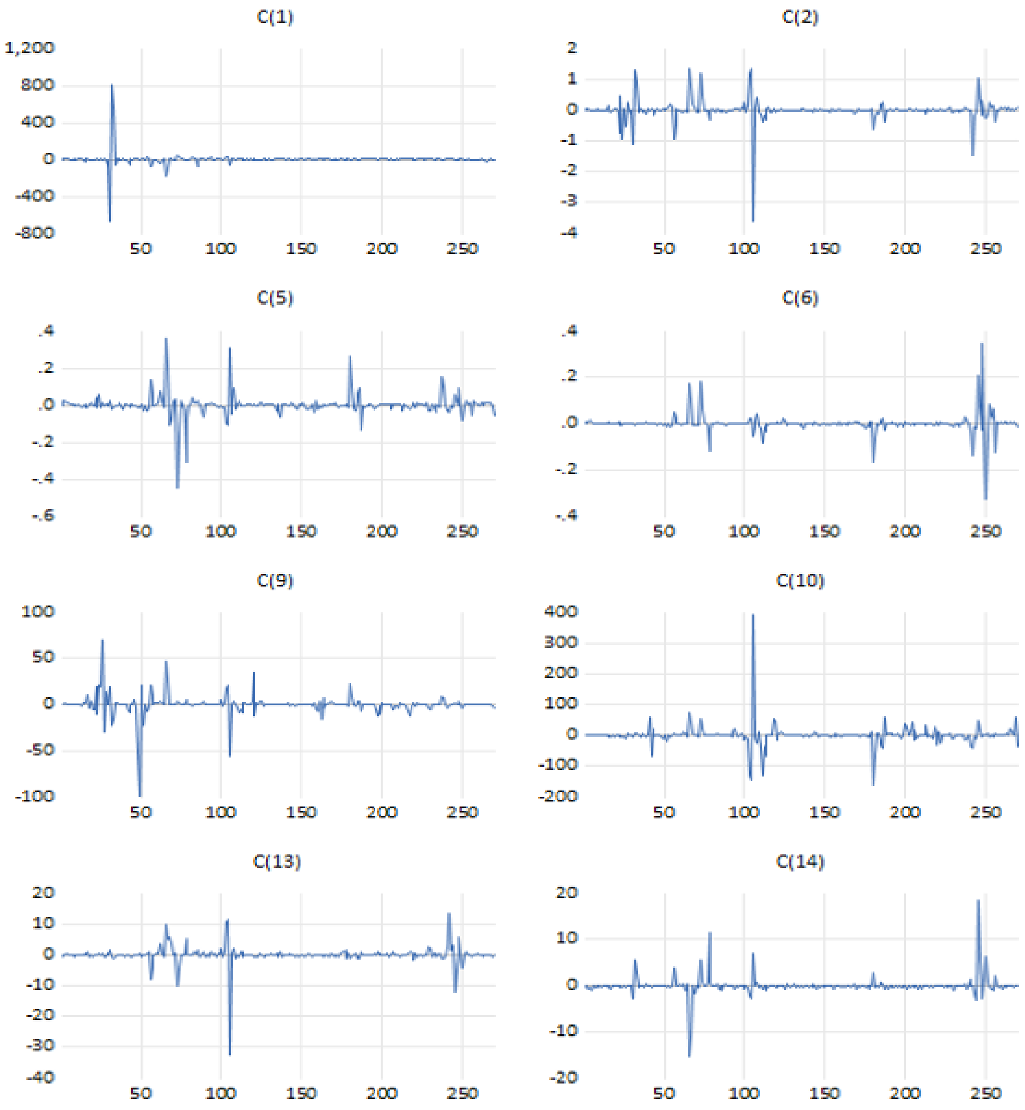


Figure C.9: The graph of gradients of objective functions.

and past volatility helps to forecast future volatility:

$$\log(h_t) = 2.1926 + 6.5796\left|\frac{h_{t-1}}{\sqrt{h_{t-1}}}\right| - 7.1589\frac{h_{t-1}}{\sqrt{h_{t-1}}} + 0.4493\log(h_{t-1}). \quad \text{(C.21)}$$

C.2.2 Statistical Results of Consumer Confidence Index with Multiple Input Variables

In this part, similarly to sentiment index, we assign consumer confidence index (CCI) as an output and other variables unemployment index (UN), consumer price index (CPI) and USD/TRY index as an input variables. As explained in the previous part C.1.2, CCI, UN and CPI are taken from Turkish Statistical Institute (TURKSTAT) [2]

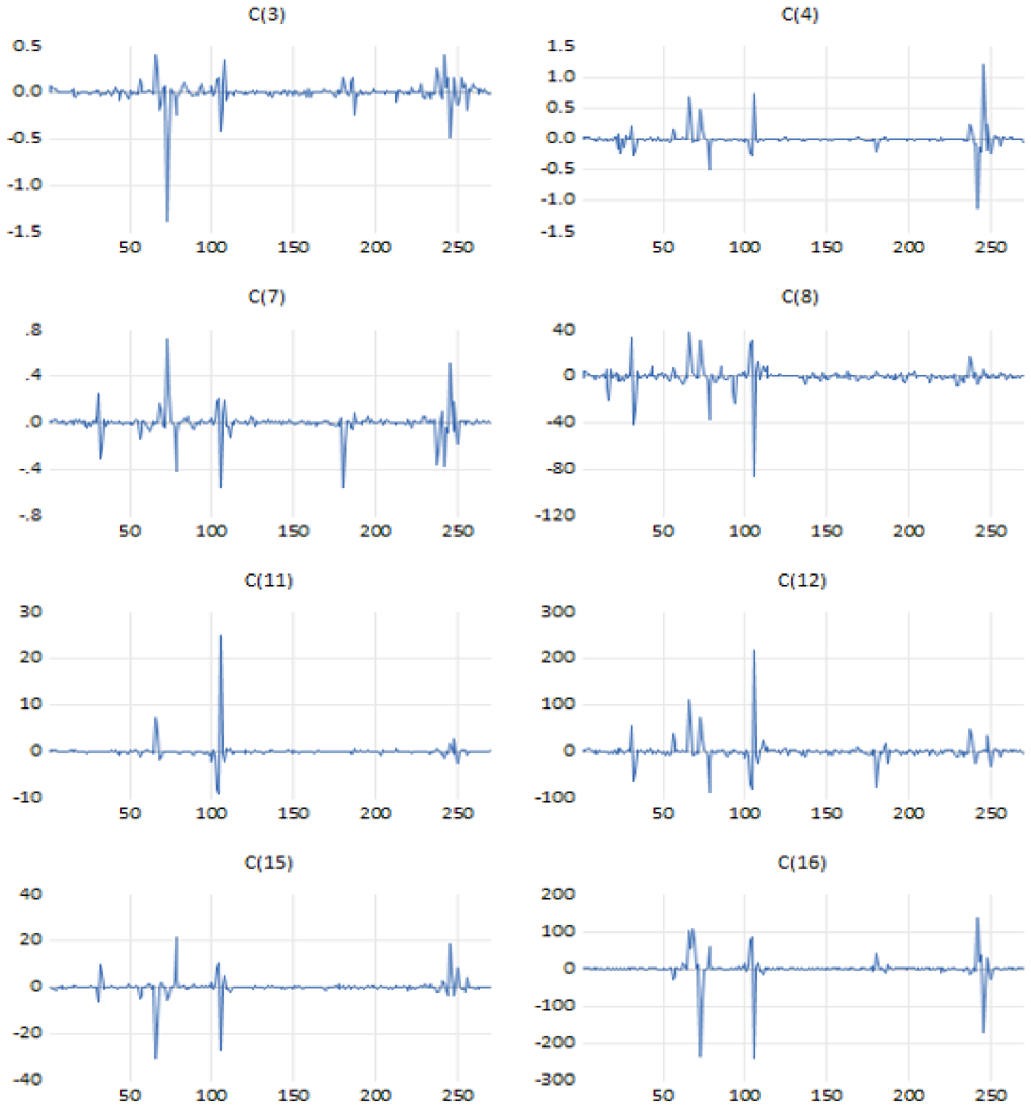


Figure C.10: The graph of gradients of objective functions.

and USD/TRY Index is from Central Bank of Türkiye EVDS Data Central [1]. For the model specification, mean equation is constructed as follows:

$$CCI_t = 0.0052 - 0.00106UN_t - 0.4608CPI_t - 0.2006USDTRY_t. \tag{C.22}$$

here, un is not statistically significant since the p-value is higher than 0.05. Furthermore, consumer confidence index has an inverse relation with consumer price index and currency index. For the variance equation below, $\hat{h}$ represents the variance and $(\hat{u}_{t-1})^2$ ARCH effect. Here, ARCH effect is statistically significant since its p-value is less than 0.05:

$$\hat{h}_t = 0.0003 + 0.5014(\hat{u}_{t-1})^2. \tag{C.23}$$

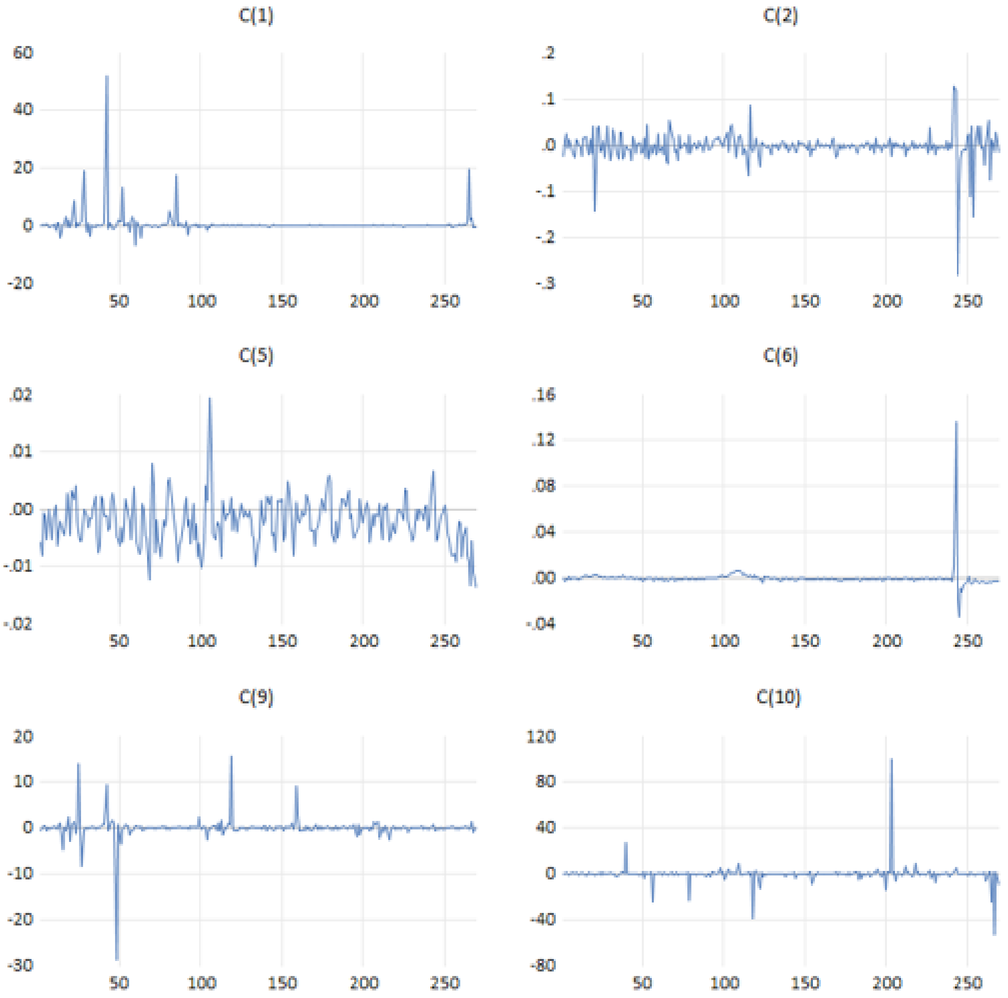


Figure C.11: The graph of derivatives of the equation specification.

For the GARCH and EGARCH model results are not obtained. On the other hand, apart from normal distribution, we apply student-t distribution as well. For the model specification, mean equation is constructed as follows:

$$CCI_t = 0.0052 - 0.0107UN_t - 0.4608CPI_t - 0.2006USDTRY_t. \tag{C.24}$$

According to the student-t distribution, UN is again not statistically significant since the p-value is higher than 0.05. Furthermore, consumer confidence index has an inverse relation with consumer price index and currency index. For the variance equation below, $\hat{h}$ represents the variance and $(\hat{u}_{t-1})^2$ ARCH effect. Here, ARCH effect is statistically significant since its p-value is less than 0.05:

$$\hat{h}_t = 0.0003 + 0.5011(\hat{u}_{t-1})^2. \tag{C.25}$$

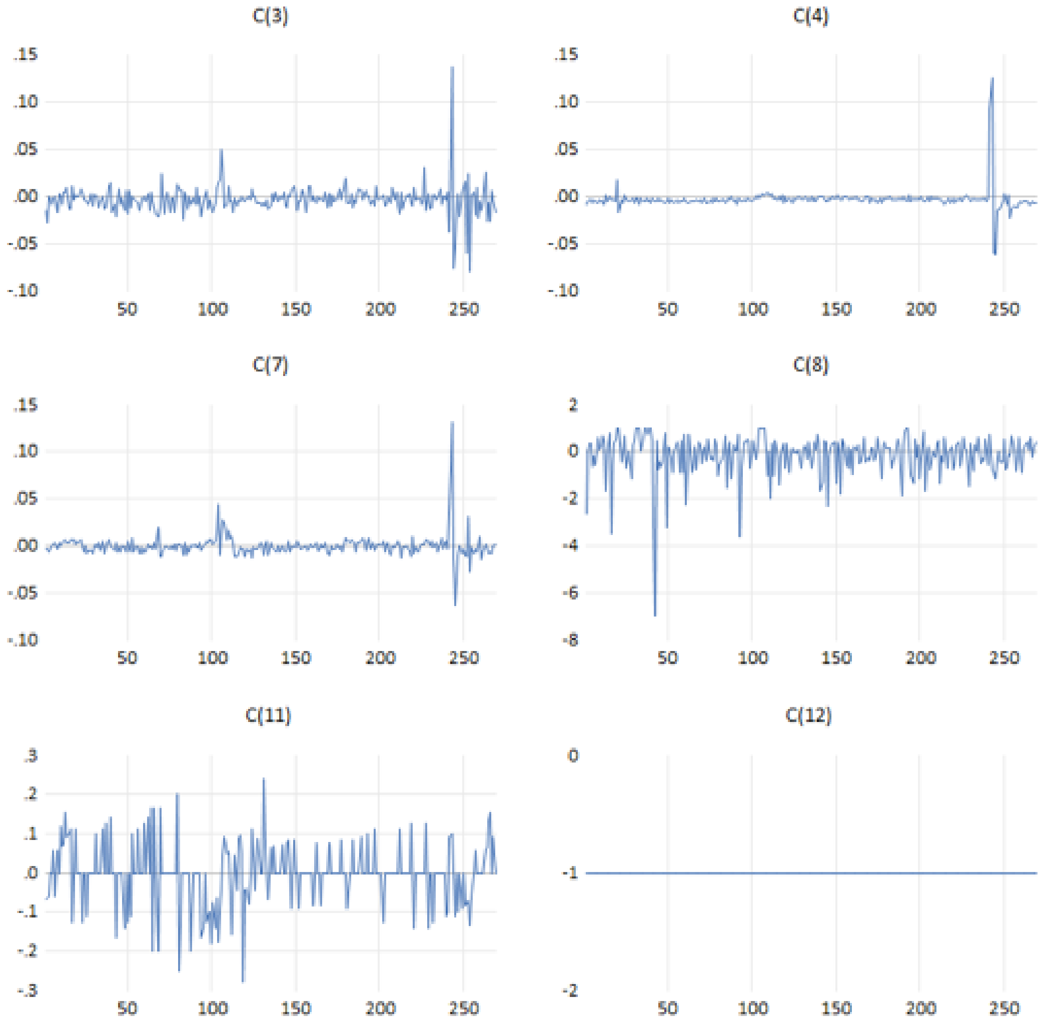
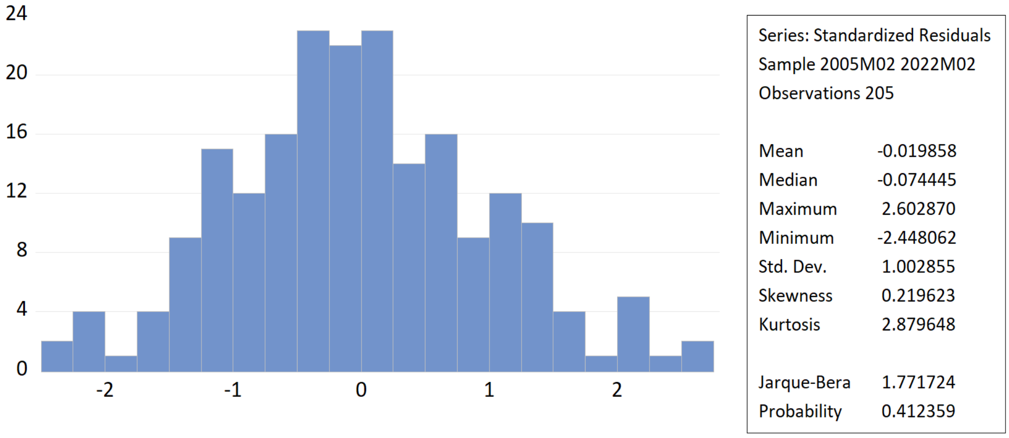


Figure C.12: The graph of derivatives of the equation specification.



```
Series: Standardized Residuals
Sample 2005M02 2022M02
Observations 205

Mean          -0.019858
Median        -0.074445
Maximum        2.602870
Minimum       -2.448062
Std. Dev.      1.002855
Skewness       0.219623
Kurtosis       2.879648

Jarque-Bera    1.771724
Probability    0.412359
```

Figure C.13: Descriptive Statistics and Histogram of consumer confidence output variable and other input variables.

Table C.19: The statistical results of lags and ARCH=1 of sentiment index variable as an output and other variables as an inputs with Student-t Distribution.

Variable	Coefficient	Std.Error	z-statistic	Prob.
cefd	-0.0040	0.0076	-0.5851	0.5585
consdur	0.5850	0.9010	0.6492	0.5162
consnon	2.0779	2.1104	0.9846	0.3248
consserv	-8.8231	4.1362	-2.1331	0.0329
cpi	-10.1135	6.1264	-1.6508	0.0098
employ	12.3650	7.9241	1.5604	0.1187
indpro	-6.2397	3.2612	-1.9133	0.0557
nipo	-0.0014	0.0239	-0.059	0.9522
pdnd	-0.0058	0.0106	-0.5527	0.5804
ripo	-0.0013	0.0025	-0.5339	0.5934
s	-0.4017	0.2572	-1.5615	0.1184
C	0.0417	0.0231	1.7996	0.0719

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C	157.1400	82493.7300	0.0019	0.9985
$Resid(-1)^2$	2972.8600	15603.0000	0.0019	0.9985

S.E of regression	3.9652
Sum squared resid.	4040.8700
Log likelihood	-266.8240
Mean dependent var	0.1956
S.D. dependent var	3.8809
Akaike info criterion	2.8953
Schwarz criterion	2.2958

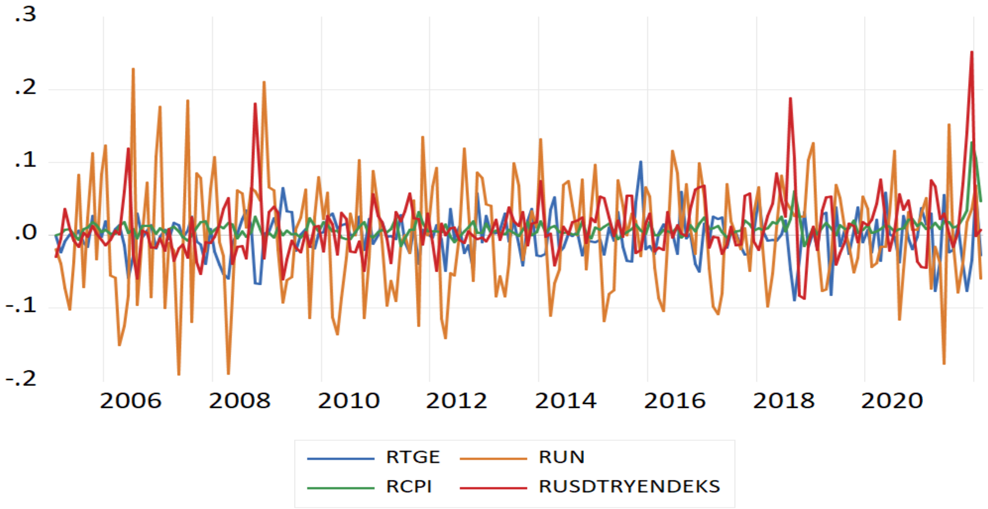


Figure C.14: The residuals of consumer confidence index variable and other proxies.

Table C.20: The statistical results of lags and EGARCH=1 of sentiment index variable as an output and other variables as an inputs with Student-t Distribution.

Variable	Coefficient	Std.Error	z-statistic	Prob.
cefd	-0.0054	0.0080	-0.6742	0.5002
consdur	0.2582	1.0982	0.2352	0.8141
consnon	1.7131	2.6488	-0.6467	0.5178
consserv	-4.4689	6.9328	-0.6446	0.5192
cpi	-6.8022	7.2209	-0.9420	0.3462
employ	9.6905	12.0080	0.8069	0.4197
indpro	-7.3516	3.9708	-1.8514	0.0641
nipo	-0.0062	0.0308	-0.2032	0.8390
pdnd	-0.0046	0.0129	-0.3554	0.7223
ripo	-0.0010	0.0042	-0.2384	0.8116
s	-0.2462	0.3150	-0.7814	0.4346
C	-0.0040	0.0320	-0.1267	0.8992

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C(13)	2.1926	2.1585	1.0159	0.3097
C(14)	6.5796	12.1498	0.5415	0.5881
C(15)	-7.1589	13.2178	-0.5416	0.5881
C(16)	0.4493	0.0644	6.9765	0.0000

S.E of regression	3.9685
Sum squared resid.	4047.5960
Log likelihood	-272.3464
Mean dependent var	0.1956
S.D. dependent var	3.8809
Akaike info criterion	2.1512
Schwarz criterion	2.3785

As shown in the above graphs C.9, C.10, C.11 and C.12, each of the labels which starts with an C define proxies. List of these is indicated in the following table:

C.2.2.1 Evaluation of Application of the Consumer Confidence Index with One-Lag

Consumer Confidence Index is explained by the Turkish Statistical Institute (TÜİK) approximately two weeks later according to the other indexes (Consumer Price Index and Unemployment Index). Therefore, we also want to try these applications

Table C.21: The statistical results of lags and ARCH=1 of consumer confidence index variable as an output and other variables as an inputs with Normal Distribution.

Variable	Coefficient	Std.Error	z-statistic	Prob.
UN	-0.0010	0.0220	-0.4854	0.6274
CPI	-0.4608	0.1136	-4.0579	0.0000
USD/TRYINDEX	-0.2006	0.0361	-5.5488	0.0000
C	0.0052	0.0019	2.6978	0.0070

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C	0.0003	6.08E-05	6.0639	0.0000
$Resid(-1)^2$	0.5014	0.1413	2.6978	0.0070

S.E of regression	0.0262
Sum squared resid.	0.1385
Log likelihood	471.0680
Mean dependent var	-0.0016
S.D. dependent var	0.0028
Akaike info criterion	-4.5372
Schwarz criterion	-4.4399

Table C.22: The statistical results of lags and ARCH=1 of consumer confidence index variable as an output and other variables as an inputs with Student-t Distribution.

Variable	Coefficient	Std.Error	z-statistic	Prob.
UN	-0.0107	0.0220	-0.4864	0.6267
CPI	-0.4608	0.1156	-3.9850	0.0001
USD/TRYINDEX	-0.2006	0.0361	-5.4260	0.0000
C	0.0052	0.00194	2.6715	0.0076

		Variance Equation		
Variable	Coefficient	Std.Error	z-statistic	Prob.
C	0.0003	6.65E-05	5.5509	0.0000
$Resid(-1)^2$	0.5011	0.1432	3.4995	0.0005

S.E of regression	0.0262
Sum squared resid.	0.1385
Log likelihood	471.0680
Mean dependent var	-0.0016
S.D. dependent var	0.0028
Akaike info criterion	-4.5275
Schwarz criterion	-4.4140

with one-lag of CCI. The results of these with one-lag CCI has almost same values according to AIC and BIC values for univariate case. On the other hand, for the application of multiple input variables, results which is used with CCI instead of lagged

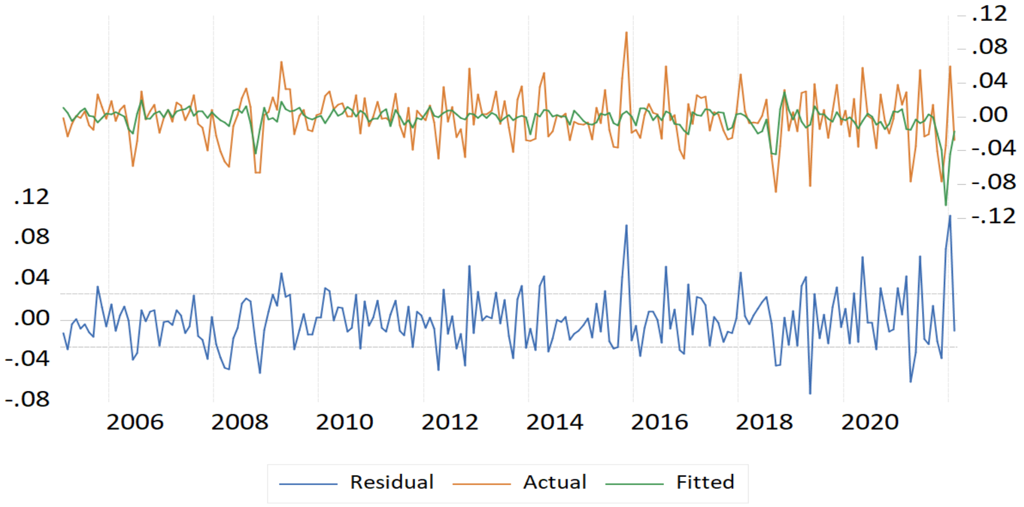


Figure C.15: The graph of actual and fitted residuals.

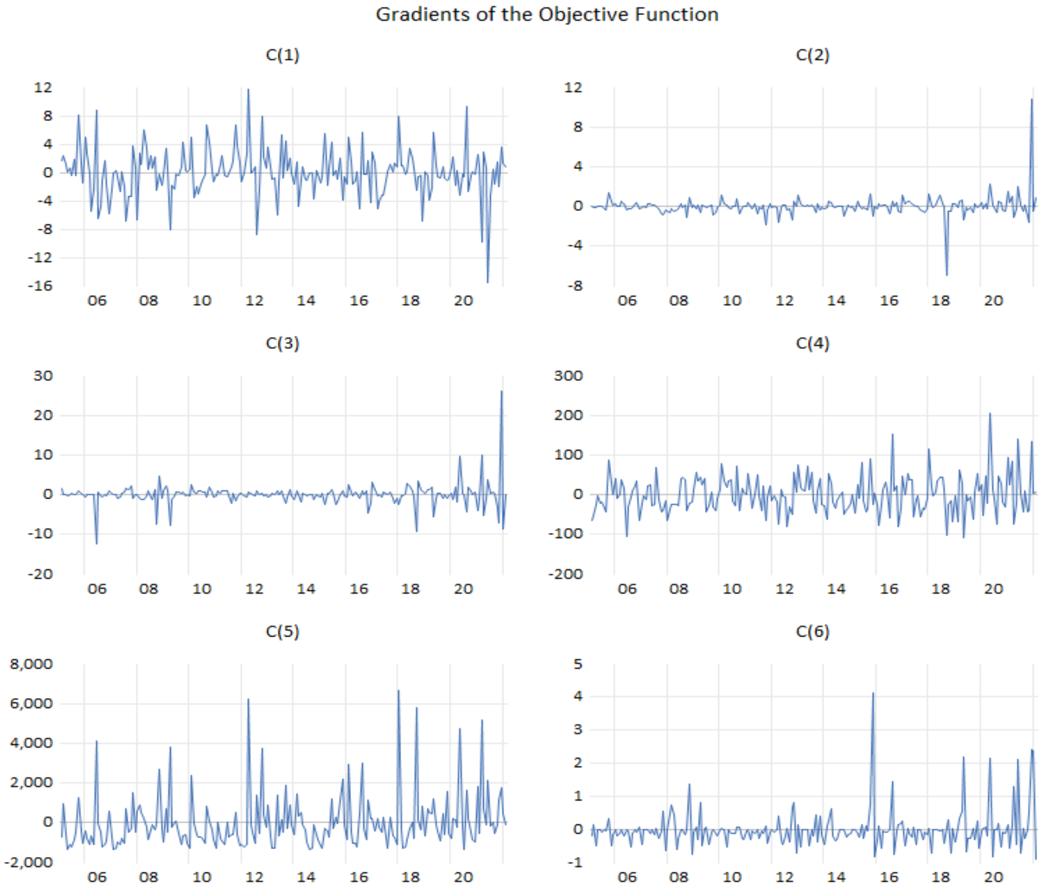


Figure C.16: The graph of gradients of objective functions.

CCI are obtained much better results. From this point of view, we can come to terms that the strength of dependency structure of variables does not create a issue by not taking the lag of CCI.

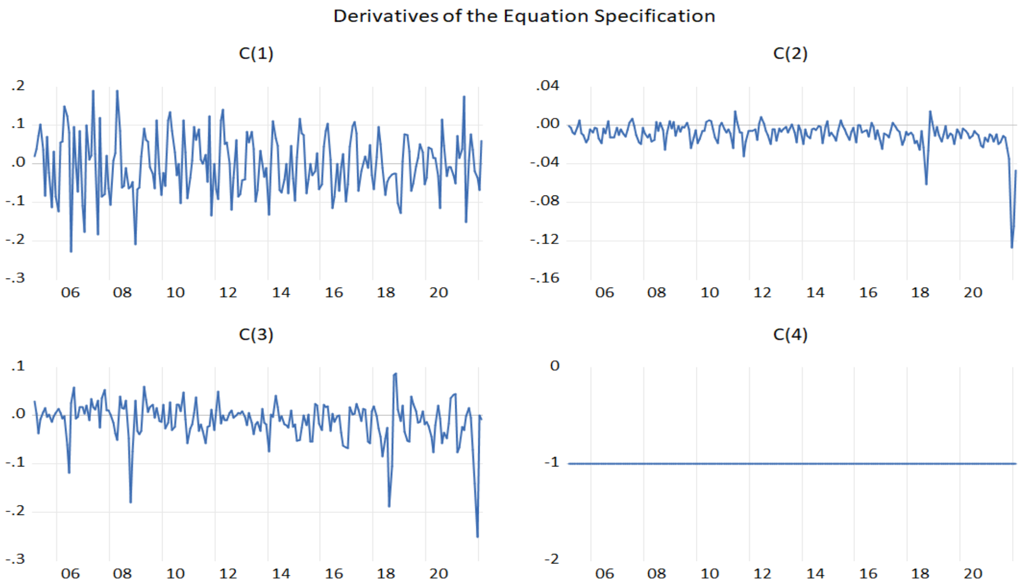


Figure C.17: The graph of derivatives of the equation specification.

Table C.23: The Coefficient Labels.

Variable	Coefficient
rUN	C(1)
rCPI	C(2)
RUSDTRY	C(3)
C	C(4)
C	C(5) (For the variance equation)
$Resid(-1)^2$	C(6) (ARCH term)

C.2.2.2 Evaluation of Sub-Indexes and Biasness of the Consumer Confidence Index

The consumer confidence index is an influential gauge indicative of consumers' overall impression as a result of their judgments and expectations on numerous subjects. The indicator measures consumer confidence in economic activity. Currently, TURKSTAT uses the TRAMO-SEATS approach, which is based on the ARIMA model established by the Bank of Spain and recommended by Eurostat, to seasonally adjust Consumer Confidence Indices (CCI) [2]. By the European Commission's Directorate-General for Economic and Financial Affairs (DG ECFIN) in 2018, to better evaluate and understand customer evaluations and expectations, alternative index research were conducted in order to establish an index that would represent, improve the quality of the consumer confidence index, and be more predictive. To maintain DG ECFIN compliance and to better evaluate consumer evaluations and expectations, In order to guarantee that the index findings more accurately represent consumer sentiment and to boost consumer confidence To enhance the quality of the consumer confidence index, it was decided to modify two of the four sub-indices utilized in the computation as of September 2020 by TURKSTAT. From January 2004 until August 2020, the modification procedure was carried out. As a result, from the sub-indexes expectation for the number of unemployed in the next 12 months and likelihood of saving in the next 12 months are switched with the sub-indexes which are expenditure on durable goods during the next 12 months compared to the prior 12-month period and the present financial status of the household in comparison to the prior 12-month period [2].

On the other hand, the consistency of standard least squares estimators in regression models is based on the assumption that the explanatory variables have no correlation with the error term. This assumption is easily broken, especially when crucial explanatory factors are left out of the model. Such exclusions are unavoidable due to the model's incapacity to collect necessary variables. As a result, not only is it possible to estimate the effects of essential variables, but estimates for additional effects in the model may be biased and hence deceptive. This is commonly referred to as an omitted variable bias [36, 52]. In our model, in case of any bias since the changes occur in the sub-indexes of CCI, we check biasness. In order to test for existency of omitted variable bias, we employ *Ramsey Reset Test*.
Here, the hypothesis are:
H_0: the model has no omitted variables
H_1: the model has omitted variables. By using Ramsey Reset test, we obtain the following results:

	Value	df	Prob.
t-statistic	0.3237	199	0.7464
F-statistic	0.1048	(0.1990)	0.7464
Likelihood ratio	0.1074	1	0.7431

According to the p-value of statistics, they are all higher than 0.05, therefore we reject

H_1 hypothesis. Thus, CCI is not biased. At this point, we can say that even though CCI sub-index computation method is updated in 2020, this has not cause biasness.

APPENDIX D

APPENDIX

D.1 Hidden Markov Model

A hidden Markov model (HMM) is a statistical model designed to represent the hidden states of an arrangement and how they evolve as a result of a Markov process [92]. HMM has been studied on a large scale and widely employed in statistics and machine learning. Nowadays, it has become a general statistical tool to model sequences or time series where the observations based on some underlying states. HMM is one of the most well-known approaches in the field of time series modeling due to its ability to remove patterns and perform appropriate calculations [44]. HMMs have been applied in numerous fields. They are extensively employed in speech recognition, bioinformatics, semiconductor malfunction and also have arised in engineering, image processing, and the areas of physical and biological sciences [60, 92]. HMM was first devised in speech recognition, but is extensively applied to forecast stock market data. There has been a lot of work done with techniques and algorithms for training models for forecasting the next day close value of the stock market, for which randomly produced transition probability matrices, emission probability matrices, and previous probability matrices have been considered. In order to improve the prediction accuracy and overcome overfitting problem, Hassan et. al (2005) employed HMM to success better optimization. They attempted to design a model that combined HMM and neural networks for stock market forecasting, and they also combined HMM with fuzzy logic rules to improve accuracy in forecasting on non-stationary stock data sets [43, 60]. Lee et al. (2009) modeled the stock return as a mixture of Gaussian and discrete Markov Chain in order to enhance the predictability of the stock model. They introduced another economic data to show Double HMM which works with the Markov Chain of the economic states separately, which gives model more degree of freedom. As a consequence, they affirm that the Double HMM forecast better than the Single HMM [56].

The broad HMM approach framework is an unsupervised learning strategy that allows us to investigate novel patterns without imposing a template throughout the learning process.

The financial time series is generated by an underlying stochastic process that is most likely related to market conditions and investment decisions that are unknown to the general public. As a result, there is a good fit between sequential data and HMM in which forecasting for the next state is based solely on the current state rather than the entire history of the previous process [60]. Li et al. (2016) use the Hidden Markov Expert Model to forecast the underlying state change of the S&P 500 Index on a monthly basis. They distinguish between two regimes: bull market and bear market. The underlying states can indirectly guide market price direction and are thus essential at the same level. For example, based on their study, the bear/bull states suggest a more distinct temporal pattern than the original market price [60].

HMM has a sequence of observations and sequence of states which produces them. Observations sequence is denoted as $O = (O_1, O_2, ...O_T)$ where each observation is an object from the set $o = \{o_1, o_2, ..., o_M\}$. Here, T shows the length of the observation and state sequences, and M denotes the number of possible observations. The hidden states which produce these observations are shown by $S = (S_1, S_2, ..., S_T)$ where each state is an object from a set of states $s = \{s_1, s_2, ..., s_N\}$.Here, N shows the number of possible states [29]. There are also conditional independence assumptions shown below.

- $P(O_k|S_1, ..., S_T, O_1, ..., O_T) = P(O_k|S_k)$ for any $1 \leq k \leq T$.
- $P(O_i, O_j|S_i, S_j) = P(O_i|S_i, S_j)P(O_j|S_i, S_j) = P(O_i|S_i)P(O_j|S_j)$ for $1 \leq i, j \leq T$.
- $P(S_k|S_1, ..., S_k - 1) = P(S_k|S_k - 1)$ for any $2 \leq k \leq T$, i.e., states from a Markov chain [29].

Because of these assumptions, the joint probability of the system can be written as:

$$\begin{aligned} P(O_1,...,O_T,S_1,...,S_T) &= P(O_1,...,O_T|S_1,...,S_T)P(S_1,...,S_T) \\ &= P(O_1|S_1)P(O_2|S_2)...P(O_T|S_T)P(S_1)... \\ & P(S_2|S_1)P(S_3|S_2)...P(S_T|S_T-1). \\ &= (\textstyle\prod_{i=1}^{T} P(O_i|S_i))P(S_1)(\prod_{i=2}^{T} P(S_i|S_i-1)). \end{aligned} \quad [29]. \tag{D.1}$$

For that reason, we require the probabilities below in order to define an HMM [29].

- *Transition probabilities*: $a_{ij} = P(S_t = j|S_t - 1 = i)$ for $1 \leq i, j \leq N$.
- *Emission probabilities*: $a_{ij} = P(O_t = j|S_t - 1 = i)$ for $1 \leq i \leq N$, $1 \leq j \leq M$.
- *Initial probabilities*: $\pi = P(S_1|s_i)$ for $1 \leq i \leq N$.

Transition and emission probabilities can also written in a matrix form. If we say, A, B and initial probabilities asa vector π. Here, it is denoted for HMM as $\lambda = (A, B, \pi)$. By modeling via HMM, we are focusing fundamentally to solve three basic problems [29]:

- Finding likelihood of an observation sequence given a HMM with parameters with λ.
- Finding the most probable state sequence given the model parameters and the observation sequence.
- Estimating the model parameters given sequences of states and observations [29].

D.1.0.1 Calculation of Likelihood

While the state sequence is hidden, find the likelihood of an observation sequence given the model parameters. There are three basic approaches to this calculation: naive and quicker approaches, forward and backward algorithms.

Naive Approach

First, we find the likelihood given a specific state as shown previously and then, we sum over all possible states as below:

$$P(O|\lambda) = \sum_{S} P(O, S|\lambda), \tag{D.2}$$

where λ is the model parameter. There are N^T possible states. For that reason, if N and T are large, this approach becomes computationally demanding in order of $\mathcal{O}(N^T)$ to calculate the likelihood [29].

Forward Algorithm

The Forward Algorithm is a dynamic programming example where we seperate the problem into sub problems and use the earlier results in a recursion. By this way, it can be found the inference problem faster than the naive approach. The likelihood of the observation sequence and a specific state at the last position of the state sequence given model parameters summed over all possible states are given as below [29]:

$$P(O|\lambda) = P_\lambda(O) = \sum_{i=1}^{N} P(S_T = s_i, O|\lambda). \tag{D.3}$$

Therefore, to find the term in the summation conditional on the model parameters, λ, we define

$$\alpha_k(S_k) = P_\lambda(S_k = s_i, O_1, ..., O_k). \tag{D.4}$$

We complete the forward algorithm with the complexity $\mathcal{O}(N^2T)$. For large values of N and T, this complexity is lower than the complexity of the naive approach [29].

Backward Algorithm

The Backward algorithm is is similar to the forward algorithm, except the starting point of the calculation. Hereby, we find the likelihood by the expression as below:

$$\begin{aligned} P_\lambda(O) &= \textstyle\sum_{i=1}^{N} P_\lambda(S_1 = s_i, O) \\ &= \textstyle\sum_{i=1}^{N} P_\lambda(S_1 = s_i) P_\lambda(O_1|O_2, ..., O_T, S_1 = s_i) P_\lambda(O_2, ..., O_T|S_1 = s_i) \\ &= \textstyle\sum_{i=1}^{N} P_\lambda(S_1 = s_i) P_\lambda(O_1|S_1 = s_i) P_\lambda(O_2, ..., O_T|S_1 = s_i). \\ &= \textstyle\sum_{i=1}^{N} \pi(s_i) b_{s_i}, o_1 P_\lambda(O_2, ...O_T|S_1 = s_i). \end{aligned} \tag{D.5}$$

To obtain the solution, the last term in the sum needed to be obtained. It is calculated by the following expression.

$$\begin{aligned}\beta_k(S_k) &= P_\lambda(O_{k+1}, ..., O_N|S_k)\\ &= \textstyle\sum_{S_{k+1}=s_1}^{S_N} P_\lambda(O_{k+1}, ..., O_T, S_{k+1}|S_k)\\ &= \textstyle\sum_{S_{k+1}=s_1}^{S_N} P_\lambda(O_{k+2}, ..., 0_T|S_{k+1}, S_k, O_{k+1})P_\lambda(O_{k+1}|S_{k+1}, S_k)P_\lambda(S_{k+1}|S_k)\\ &= \textstyle\sum_{S_{k+1}=s_1}^{S_N} P_\lambda(O_{k+2}, ..., O_T|S_{k+1})P_\lambda(O_{k+1}|S_{k+1})P_\lambda(S_{k+1}|S_k)\\ &= \textstyle\sum_{S_{k+1}=s_1}^{S_N} \beta_{k+1}(S_{k+1})b(S_{k+1})O_{k+1}a(S_k, S_{k+1})\end{aligned} \tag{D.6}$$

for $1 \leq k \leq N-1$. For $\beta_T(S_T)$, we cannot use the definition which is shown above, since it involves $O_(N+1)$ which does not exist. Thus, if we use recursion formula for $k = T-1$,

$$\begin{aligned}\beta_T - 1(S_T - 1) &= \textstyle\sum_{S_T=s_1}^{S_N} P_\lambda(O_T, S_T|S_{T-1})\\ &= \textstyle\sum_{S_T=s_1}^{s_N} \beta_T(S_T)P_\lambda(O_T|S_T)P_\lambda(S_T|S_{T-1}).\end{aligned} \tag{D.7}$$

However, $P_\lambda(O_T, S_T|S_{T-1})$ can be also written as,

$$\begin{aligned}P_\lambda(O_T, S_T|S_{T-1}) &= P_\lambda(O_T, S_T|S_{T-1})P_\lambda(S_T|S_{T-1})\\ &= P_\lambda(O_T|S_T)P_\lambda(S_T|S_{T-1}).\end{aligned} \tag{D.8}$$

For that reason, for Equation D.8 to hold, $\beta_T(S_T) = 1$. By using Equation D.1 we obtained

$$P_\lambda(O) = \sum_{i=1}^{N} \pi(s_i)b_{(s_i),o_1}\beta_1(S_1 = s_i). \tag{D.9}$$

Viterbi Algorithm: Inference of the most probable path

This recursive technique is used to discover the most likely sequence, also known as the path, given the observation sequence and parameters. After initializing the state, the previous pathways determined are employed in the calculation at each step. The goal is to obtain:

$$S^* = \text{argmax}_S\, P(S|O) \tag{D.10}$$

If $f(a) \geq 0$ for all a and $g(a,b) \geq 0$ for all a, b, we have

$$\max_{a,b} f(a)g(a,b) = \max_a\{f(a)\max_b g(a,b)\}, \tag{D.11}$$

we have

$$\arg max_S P(S|O) = \arg max_S P(S|O) \tag{D.12}$$

since $P(O)$ does not include any element from hidden states [29].

Expectation-Maximization (EM) Algorithm

When data is partial or contains missing values, the EM approach is commonly used to derive the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set [17].

EM Algorithm consists of two main steps. The first step, which is called *Expectation*, happens when some missing values or latent variables are obtained in the data, because of the problems or limitations that occur in the observation process [17]. In order to tackle this problem, the anticipated value for each of these hidden variables is estimated in this stage. The second phase occurs when optimizing the likelihood function is computationally difficult. The likelihood function, on the other hand, can be simplified by assuming the existence and values of hidden factors [17]. The complete data is used which is obtained in the previous (expectation) step and the parameters are updated in this step.

We have $p(x|\theta)$ which is density function where θ is set of parameters, our distribution is $\mathcal{X} = \{X_1, \ldots, X_N\}$ with data size of N. Here, we assume that data $\mathcal{X}$ is incomplete. It is observed and generated by distribution. On the other hand, complete dataset is defined as $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ and joint density function is specified as [17]:

$$p(z|\theta) = p(x, y|\theta) = p(y|x, \theta)p(x|\theta). \tag{D.13}$$

This joint density stems from the marginal density function $p(x|\theta)$ and the assumption of latent variables and parameter value estimates [17]. By using this new defined density function, a new likelihood function which is named as complete-data likelihood, can be generated as $\mathcal{L}(\theta|\mathcal{Z}) = \mathcal{L}(\theta|\mathcal{X}, \mathcal{Y}) = p(\mathcal{X}, \mathcal{Y}|\theta)$. Here, this new likelihood function is a random variable because of the missing information $\mathcal{Y}$ is unknown and random. Thus, the likelihood function $\mathcal{L}(\theta|\mathcal{X}, \mathcal{Y}) = h_{\mathcal{X},\theta}(\mathcal{Y})$ for some function where $\mathcal{X}$ and θ are constant and $\mathcal{Y}$ is a random variable, here the incomplete-data likelihood function referred as the original likelihood $\mathcal{L}(\theta|\mathcal{X})$ [17].

At first, the EM Algorithm computes the expected value of the complete-data log-likelihood $\log p(x, y|\theta)$ with respect to the unknown data $\mathcal{Y}$ given the observed data $\mathcal{X}$ and the current parameter estimates, which is defined as [17]:

$$Q(\theta, \theta^{(i-1)}) = E[\log p(\mathcal{X}, \mathcal{Y}|\theta)|\mathcal{X}, \theta^{(i-1)}]. \tag{D.14}$$

Here, $\theta^{(i-1)}$ are the present parameters estimates that we used to interpret the expectation and θ are the new parameters that we optimize to raise Q. At this point, to clarify this statement, $\mathcal{X}$ and $\theta^{(i-1)}$ are constants, θ is a normal variable which needs to be adjusted and $\mathcal{Y}$ is a random variable conducted by the distribution $f(y|\mathcal{X}, \theta^{(i-1)})$. For that reason, the right side of the Equation D.14 can be written as [17]:

$$E(\log p(\mathcal{X}, \mathcal{Y}|\theta)|\mathcal{X}, \theta^{(i-1)}) = \int_{y \in \mathcal{Y}} \log p(\mathcal{X}, y|\theta) f(y|\mathcal{X}, \theta^{(i-1)}) dy. \tag{D.15}$$

Here, $f(y|\mathcal{X}, \theta^{(i-1)})$ is the marginal distribution of the unobserved data and is dependent on both the observed data $\mathcal{X}$ and on the present parameters,and $\mathcal{Y}$ is the space of the values y can take on. Suppose there is a function $h(.,.)$ of two variables and $h(\theta, Y)$ where Y is a random variable runned by some distribution $f_Y(y)$ and θ is constant. Then $q(\theta) = E_Y[h(\theta, Y)] = \int_y h(\theta, Y) f_Y(y) dy$ is now a deterministic function that could be maximized if preferred [17].

The interpretation of this expectation is defined as *E-step* of the algorithm. If we realize the meaning of two arguments in the function $Q(\theta, \theta')$, the first argument θ relates to the parameters that eventually will be optimized to maximize the likelihood and the second argument θ' relates to the parameters which is used to utilize the expectation. On the other side, the second step, *M-step*, of the EM Algorithm is to maximize the expectation that is found in the first step, which is [17]:

$$\theta^{(i)} = \underset{\theta}{\mathrm{argmax}} Q(\theta, \theta^{(i-1)}). \tag{D.16}$$

These steps are repeated as needed. Each repetition ensures that the loglikelihood increases, and the method attempts to converge to a local maximum of the likelihood function [17].

Baum-Welch Algorithm: Estimating the Model Parameters

The Baum-Welch technique generates optimal HMM parameters for the beginning HMM and an established sequence of observations. Because the Baum-Welch algorithm is a subset of the EM algorithm, it converges to a local solution that may or may not be the global optimum [78].

The Baum-Welch technique is a repetition algorithm and an EM approach version. Here, the computation begins with an initial estimation of the parameters λ, which is iterated until λ converges. The following phrase is used in these calculations for the predictor of the transition probability between ith and jth variables [29]:

$$\hat{a}_{ij} = \frac{\text{Expected number of transitions from } i \text{ to } j}{\text{Expected number of transitions from } i}. \tag{D.17}$$

In order to obtain these expected values, the below equation is applied:

$$\begin{aligned} \xi(i,j) \quad &= P(S_t = i, S_{t+1} = j|O,\lambda), \\ &= \tfrac{P(S_t=i,S_{t+1}=j,O|\lambda)}{P(O|\lambda)}, \\ &= \tfrac{\alpha_t(S_t=s_i)a_{ij}b_{S_{t+1}=s_j,o_{t-1}}\beta_{t+1}(S_{t+1}=s_j)}{\sum_{j=1}^{N}\alpha_t(S_t=s_j)\beta_t(S_t=s_j)}. \end{aligned} \tag{D.18}$$

To predict the emission probability matrix B, similar to Equation D.18

$$\hat{b}_j(o_k) = \frac{\text{Expected number of times being in state } s_j \text{ observing } o_k}{\text{Expected number of times being in state } s_i}. \tag{D.19}$$

In the following, the meaning of γ_t is

$$\gamma_t(j) = P(S_t = j|O,\lambda) = \frac{P(S_t = j, O|\lambda)}{P(O|\lambda)} = \frac{\alpha_t(S_t = s_j)\beta_t(S_t = s_j)}{\sum_{j=1}^{N}\alpha_t(S_t = s_j)\beta_t(S_t = s_j)}. \tag{D.20}$$

Here, etimate for b_j can be written as follows:

$$\hat{b}_j(o_k) = \frac{\sum_{t=1 st O_k = o_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)}. \tag{D.21}$$

In addition, the estimate of the initial probability π can be expressed as

$$\hat{\pi}_i = \gamma_1(i). \tag{D.22}$$

D.1.1 Application of HMM

The consumer confidence index forecasts the future consumption and saving habits of households based on responses to questions about their predicted financial status, their feelings about the general economic situation, unemployment, and their ability to save. An index above 100 indicates an increase in consumer confidence in the future economic condition, as a result of which they are less likely to save and more willing to spend money on significant purchases in the next 12 months. Values fewer than 100 indicate a negative outlook on future economic developments, which may result in a preference to save more and consume less. The purpose of the Consumer Tendency Survey in Turkey is to measure consumers' present situation assessments and future period expectations on personal financial standing and general economic course, as well as to estimate consumers' expenditure and saving inclinations for the near future. The poll includes a randomly selected sample of all adults aged 15 and up who work in both urban and rural locations. The index is assigned a value between 0 and 200. Consumers who score above 100 are optimistic, while those who score below 100 are pessimistic. Application is done by using the same datasets as we used in previous part, worked economic variables are; The Consumer Confidence Index (CCI), Unemployment Index, Consumer Price Index, USD/TRY Index. The CCI, Unemployment Index and CPI are taken from Turkish Statistical Institute (TÜİK) and USD/TRY exchange rates are taken from the Central Bank of Turkey (TCMB, EVDS Data Central). All these datasets are monthly, starting from 2005 to up to now. For that reason, we have 206 variables at total.

Markov Models are a probabilistic process that uses the present state to predict the likelihood of advancing to the next state. When the complexity is predicated on not knowing the probability of each regime change and how to explain these probabilities changing over time, HMM come into play. They can evaluate the transition probabilities for each regime and then produce the most likely regime depending on current conditions [60]. In this analysis, we apply Hidden Markov Model to forecast the underlying state change of the Consumer Confidence Index at a monthly interval. The regimes are defined as being three states, optimistic investor, pessimistic investor and neutral investor.

The goal is to observe and analyze the transition between various stages, as well as to define a path through these phases [60]. We have 206 variables at total. Unfortunately, only one value is higher than 100, which is seen in 2006. That is why, at this point we require a region so that we could make an analyses about the optimism and pessimism level of investors. To construct this region, first we take the mean of our CCI data, which is 88.9, afterwards we calculate the standard deviation of our dataset, which is

6.91. By using *Chebyshev's Inequality*, we determine our region:

$$P(|x-\mu| \geq k) \leq \frac{\sigma^2}{k^2}, \tag{D.23}$$

where μ is mean and σ represents standard deviation, besides $\sigma^2 > 0$ [85]. Here, to determine significant interval which provide us the most effective region, we set the constant $k = 0.5$. After calculation, our region is in the interval between 85.45 and 92.36. Thus, we evaluate the optimisim level of investors if the variables are higher than 92.36, and pessimisim level of investors if the variables are lower than 85.45. Between these interval are called as 'neural' or 'irresponsive' investors.

After all these determinations, we acquire probability matrices which are initial, transition and emission probabilities. According to the initial probability, there is 0.36 probability chance for investors to be optimistic, 0.41 chance to be pessimistic and 0.23 to be neutral. The transition matrix tells us the probability of moving from one state to each of the states. Here, the transition matrix shows us that there is a 74.3% chance that it stays in optimistic state, there is 16.2% chance it moves to pessimistic state and 9.5% chance to move the neutral state based on the current data set. On the other hand, there is a 15.3% chance for pessimistic investors to become optimistic and 84.7% chance to stay in the pessimistic state. Moreover, for the neutral investors, only 13% chance to move to optimistic state and 87% chance to stay in neutral state.

We calculated posterior odds over the full data set to see the posterior probability of being in each state at each time point for a particular sequence of observations and a given Hidden Markov Model. For the emission probability matrix we determine an observation. One observation spreads from a state at each time step. Observation symbols are: H (High), M (Medium) and L (Low). While constructing this observation, to observe reasonable results, we get the $k = 1$ in Chebyshev's inequality. That is, if the results are higher than 95.81, our observation is named as *H*, if the value is less than 81.99, our observation is called as *M*, and if the value is between these interval, then the observation is *L*. Our results show us that the probability of optimistic investors to stay in *M* interval is 100.0%. However, the probability for pessimistic investors to be observed at H level is 30.0% and to be observed at *M* level is 70.0%. Also, for the neutral investors, probability of staying in L interval is 95.0% and *M* level is 5.0%.

Table D.1: The result of HMM

States
optimistic
pessimistic
neutral

Table D.2: Observations.

Symbols
H
L
M

Table D.3: StartProbs

optimistic	0.3600
pessimistic	0.4100
neutral	0.2300

Table D.4: TransProbs

	to		
from	optimistic	pessimistic	neutral
optimistic	0.7400	0.1600	0.1000
pessimistic	0.1500	0.8500	0.000
neutral	0.1300	0.0000	0.8700

Table D.5: EmissionProbs

	to		
from	H	L	M
optimistic	0.0000	0.0000	1.0000
pessimistic	0.2900	0.0000	0.7100
neutral	0.0000	0.9600	0.0400

Viterbi training requires substantially less computing work than Baum-Welch training, but results in the same or slightly lower performance. As a result, it is commonly used by designers of speech recognition systems. The Baum-Welch algorithm, on the other hand, exhibits some unusual properties: in the case of discrete HMMs, it does not require any model initialization, only non-zero random values confirming the stochastic constraints; and it thoroughly uses all available data to generate robust and optimal estimates [80]. From the standpoint of Viterbi training, it is demonstrated that even in the situation of discrete HMMs, some suitable initialization is required, either by using models discovered for other databases or by training initial models on a hand-labeled subset of the training database [80].

In the following part, we see application of our model to observe the comparison and similarities between these algorithms and to analyse the probability of our states.

In order to make an inference based on observable data and a trained model. Based on our calculations from observed data, these are the most likely states, we use the *viterbi algorithm* [60]. According to viterbi algorithm, we obtain the same initial probabilities for optimistic, pessimistic and neutral investors. However, for the transition and emission probability matrices we see different results. For the results of viterbi algorithm, the transition matrix shows us that there is a 93.4% chance that it stays in optimistic state, there is 2.5% chance it moves to pessimistic state and 4.1% chance to move the neutral state based on the current data set. On the other hand, there is a 10.3% chance for pessimistic investors to become optimistic and 89.7% chance to stay in the pessimistic state. Moreover, for the neutral investors, only 8% chance to move to optimistic state and 92% chance to stay in neutral state. According to the emission probability matrix, the probability of optimistic investors to stay in *M* interval is 100.0%. However, the probability for pessimistic investors to be observed at *H* level is 64.1% and to be observed at *M* level is 35.9%. Also, for the neutral investors, probability of staying in *L* interval is 97.8% and *M* level is 2.2%.

Table D.6: StartProbs obtained with Viterbi Algorithm

optimistic	0.3600
pessimistic	0.4100
neutral	0.2300

Table D.7: Transmission Probabilities obtained with Viterbi Algorithm

	to		
from	optimistic	pessimistic	neutral
optimistic	0.9300	0.0300	0.0400
pessimistic	0.1000	0.9000	0.0000
neutral	0.0000	0.9700	0.0300

The Baum-Welch algorithm is the most trustworthy algorithm for training the HMM,

Table D.8: Emission Probabilities obtained with Viterbi Algorithm

	to		
from	H	L	M
optimistic	0.0000	0.0000	1.0000
pessimistic	0.6400	0.0000	0.3600
neutral	0.0000	0.9800	0.0200

according to the literature. The Baum-Welch algorithm can use an observation sequence to train the supplied HMM and generate a new HMM for detection [103]. Therefore, apart from viterbi algorithm, we also see our probability matrix values via *Baum-Welch algorithm*. There are only a few differences between the results that we obtain with viterbi algorithm. Initial probabilities are same with the previous algorithms. For the transition probability matrix, there is a 92.3% chance that it stays in optimistic state, there is 3.6% chance it moves to pessimistic state and 4.1% chance to move the neutral state based on the current data set. On the other hand, there is a 12.8% chance for pessimistic investors to become optimistic and 87.2% chance to stay in the pessimistic state. Moreover, for the neutral investors, only 8.3% chance to move to optimistic state and 91.7% chance to stay in neutral state. For the emission probability matrix, the probability of optimistic investors to stay in *M* interval is 100.0%. However, the probability for pessimistic investors to be observed at *H* level is 60.6% and to be observed at *M* level is 39.4%. Also, for the neutral investors, probability of staying in *L* interval is 95.9% and *M* level is 4.1%.

The essential problem related with HMMs is to take a sequence of observations, which is defined as a set of hidden states, and fit the most potential HMM, in other words, specify the parameters that most possibly represent what happens in the scene. The Baum-Welch algorithm, which is an EM algorithm is used when the HMM parameters are not immediately (empirically) measurable, which is usually the case in real applications similar to our situation. The second step is the "decoding task", searching the most probable sequence of hidden states given some observations, namely, getting the hidden states that produced the monitored output. The sequence of states obtained by the Viterbi algorithm is then compared with the ground truth to measure the accuracy [76].

Table D.9: Initial Probabilities obtained with Baum-Welch Algorithm

optimistic	0.3600
pessimistic	0.4100
neutral	0.2300

Table D.10: Transmission Probabilities obtained with Baum-Welch Algorithm

	to		
from	optimistic	pessimistic	neutral
optimistic	0.9200	0.040	0.040
pessimistic	0.1300	0.8700	0.0000
neutral	0.080	0.0000	0.9200

Table D.11: Emission Probabilities obtained with Baum-Welch Algorithm

	to		
from	H	L	M
optimistic	0.0000	0.0000	1.0000
pessimistic	0.6000	0.0000	0.4000
neutral	0.0000	0.9600	0.0400

Long-term thinking is essential for investors' overall success. However, the majority of them are concerned about short-term portfolio adjustments. This concern stems from recent increases in volatility over the last few years. The stock market and the economy have historically moved in cycles that repeat themselves. As a result, understanding the various stages of the economy might aid in guiding investment decisions. When investors are extremely confident, they tend to expand their stock holdings. On the reverse side, safe-haven investments such as gold and bonds will fall out of favor. In a bear market, there is a lack of faith in the economy [60]

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Kalaycı, Betül
Nationality: Turkish (TC)

EDUCATION

Degree	Institution	Year of Graduation
M.S. in Financial Mathematics	Middle East Technical University	2017
B.S. in Mathematics	Ankara University	2014

PUBLICATIONS

Thesis

Kalaycı, B. Identification of coupled systems of stochastic differential equations in finance including investor sentiment by multivariate adaptive regression splines, Master's thesis, Middle East Technical University, 2017.

International Publications

Kalaycı, B., Özmen, A. and Weber, G.-W. Mutual relevance of investor sentiment and finance by modeling coupled stochastic systems with MARS, Annals of Operations Research, 295(173), pp. 1–24, 2020.

Kalaycı, B., Purutçuoğlu, V. and Weber, G.-W. Operation Research in Neuroscience:

A Recent Perspective of Operation Research Application in Finance, Operations Research 1, CRC Press, 2022. 170-190.

Kalaycı, B., Purutçuoğlu, V., Defterli, Ö. ,Uğur, Ö. and Weber, G.-W. Application of Various Modelling Techniques into Consumer Confi- dence Index: A Recent Perspective of Operation Research Application in Finance, Operations Research 2, CRC Press, 2023 (accepted).

Conference Talks

Kalaycı, B., Purutçuoğlu, V. and Weber, G.-W. Modelling the Mutual Interaction of Finance and Human Factor via Various Economic Indicators in: 41st Eurasia Business and Economics Society (41st EBES Conference), Berlin, Germany, 2022.

Kalaycı, B., Purutçuoğlu, V. and Weber, G.-W. Construction of Mutual Interaction Between Finance and Human Factor by Various Modeling Techniques, in: 8th International Conference on Economics (ICE-TEA 2022), Kapadokya University, Nevşehir, Türkiye, 2022.

Kalaycı, B., Özmen, A. and Weber, G.-W. Mutual Relevance of Investor Sentiment and Finance by Modeling Coupled Stochastic Systems by Using MARS in: 29th European Conference on Operational Research (EURO 2018), Valencia, Spain, 2018.

Kalaycı, B., Özmen, A. and Weber, G.-W., Identification of Systems of Stochastic Differential Equations for Generalized Model Classes in Financial Mathematics Including Investor Sentiment in: International Workshop on Mathematical Methods in Engineering (MME 2017), Çankaya University, Ankara, Türkiye, 2017.