

Income Classification Benchmark: From R (Academic Study) to Python (ML Pipeline)

Mehmet Ali Erkan
Middle East Technical University
Ankara, Turkey
maerkan@metu.edu.tr

Abstract: The aim of this paper is to statistically analyze salary classification by observing people's work class, education, race, gender, and other characteristics and categorizing salary prediction based on these characteristics. To get at the results, techniques like data visualization, statistical analysis and machine learning techniques were applied. The regressions, support vector machine, artificial neural network, random forest, and Xgboost machine learning algorithms are used to classify salary classification. Research questions are developed and analyzed prior to prediction in order to better understand relationships between variables in the data. Data cleaning techniques are used to create clean, appropriate data for the study. After the dataset has been cleaned, models and statistical tests are run. Sensitivity, accuracy and F1 score were used to assessed models due to the large number of the categorical variables. The analysis is conducted using R-studio.

Keywords—Salary Classification, Artificial Neural Network, Random Forest, Xgboost, Data Analytics

I. INTRODUCTION

Every employee has a salary. The payment of this salary depends on some parameters. This can vary from the person's education to age, from working hours to being single. In the study, people will be classified according to these variables as receiving less than \$50,000 annually and above. First, the general situation of the data will be examined, necessary arrangements will be made, models will be established and these models will be compared by looking at the sensitivity, F1 score.

II. LITERATURE REVIEW

Statisticians and researchers have done enormous amounts of research and analysis about predicting salary classification. Firstly, a regression model is suggested to predict salary classification [1]. Secondly, there is another piece of research about applying different machine learning techniques to the prediction of the classification salary, basically a comparison of unsupervised and supervised learning in research [2]. According to the results, the most accurate result is random forrest method.[2] Last research is about Salary Prediction Using Machine Learning [3]. Researchers concluded that the best outcome is produced by the decision tree. But if the featured attribute is small, KNN will perform better.[3]

III. METHODOLOGY

A. Dataset

Barry Becker extracted data from the 1994 Census database. The form of this data is taken from Kaggle. In addition, at first, the data has 32561 observations and 15 variables, 9 categoric 6 continuous. In this study, salary is used as a dependent variable. Since there was no information, the fnlwgt variable was removed. Moreover, workclass, native country, and occupation variables have NA values in the dataset. Occupation has the highest NA value. Imputation was made to fill these NA values. Since 32561 observations will keep operations and iterations long, 1000 samples were randomly taken from the data. The variables utilized in the analysis are listed below.

- “age”: age of the workers – Continuous variable
- “workclass”: sector of the workers – Categoric Variable
- “fnlwgt”: no information - Continuous variable
- “education”: education level - Categoric Variable
- “education-num”: number notation of education level- Continuous Variable
- “marital-status”: marital status of worker - Categoric variable
- “occupation”: occupation of worker - Categoric variable
- “relationship”: relationship status, wife or husband exc- Categoric Variable
- “race”: race of workers - Categoric Variable
- “sex”: gender of workers Categoric Variable
- “capital-gain”: The profit on earns on the sale of an assets - Continuous Variable
- “capital-loss”: The loss on sell the assets for less than adjusted basis - Continuous Variable
- “hours-per-week”: weekly working hours of workers- Continuous Variable
- “native-country”: native country of workers - Categoric Variable
- “salary”: earning less or more than 50k - Categoric Variable

B. Descriptive Statistics

Table 1 indicates descriptive statistics of 2 continuous variable of the data.

	age	Hours per week
Minimum	17	1.00
1 st Quartile	28	40.00
Median	37	40.00
Mean	38	40.44
3 rd Quartile	48	45.00
Maximum	90	99.00

Table 1 Descriptive Statistical Summary of Some Variables

According to table 1, the average of people's age is 38. The minimum age is 17 and the maximum age is 90.

Half of the people are above or below 37. 25 % of the people are below 28 and above 48. On the other hand, the average working hours per week are 40.44. Minimum working hours is 1 and maximum is 99. 25 of the working hours are below 40 and above 45. Since there is a considerable difference between the 3rd quartile and the maximum value, there might be outliers. Also this shows that the distribution of working hours may seem right-skewed. Lastly, the response variable consists of 24720 people who earned less than 50 thousand and 7841 people who earned more than \$50,000. This shows us that the data is imbalanced.

C. Exploratory Data Analysis

Six research questions were established and addressed in this section of the data analysis. The answers to these research questions have improved our understanding of the facts.

C.1 How does level of education relate to salary?

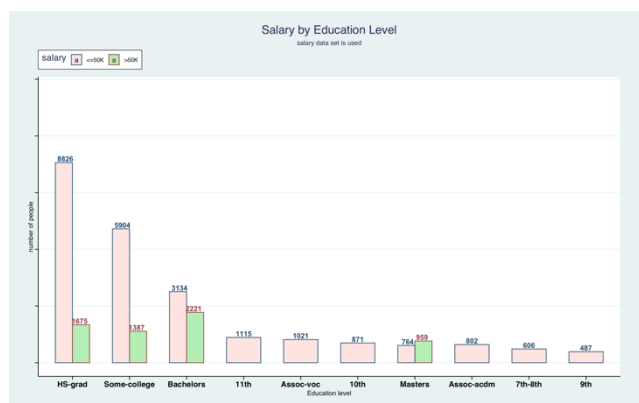


Figure 1 Plot for salary by education level

As it can be seen from the bar plot given in Figure 1, it seems there is a differences in salary between education level. The majority of high school graduates earn less than \$50,000, whereas master's degrees have the greatest earnings rate over \$50,000 among all educational specialties. (Pearson's Chi-squared test, p -value $< 2.2e-16$).

C.2 How worker's age distribute over workclass type?

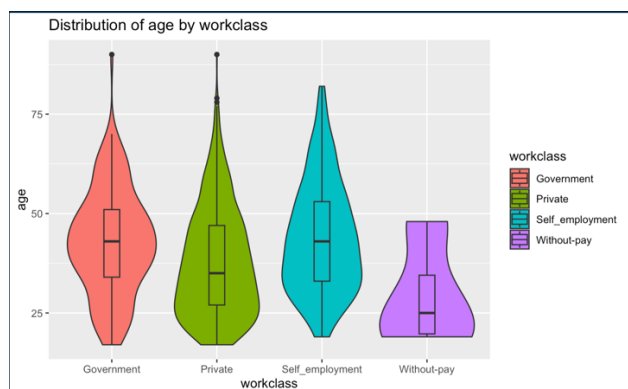


Figure 2 Violin plot for age by workclass

	Df	Sum Sq	Mean Sq	F value	P value
workclass	3	6286	2095.3	11.44	2.24e-07
Res.	996	182462	183.2		
Total	999				

Table 2 The Results of ANOVA

Since workclass variable has more than two levels, ANOVA can be conducted. In this way, it would be learned whether the levels of workclass variable have at least one different effect on people's age. Some of the workclasses are significantly different. After that normality checked residuals are not normal. Box-cox transformation was applied but transformed data are not normal again. Since normality assumption is not satisfied, Kruskal-Wallis was applied which is a non-parametric version of the one-way ANOVA (p -value $< 2.2e-16$). Thus, since the p -value is less than 0.05, it is clear that the work-class statuses differ significantly. Following that, pairwise was used. The matching pairings of procedures are significantly different when the p -values are less than 0.05, according to the data. Between self-employed individuals and private workers, as well as between government and private employees, there is a significant difference in age.

D. Missingness

Since there are so many missing observations in real-life data, it is always crucial to identify the missingness process. Three variables in this data have a significant percentage of missing values.

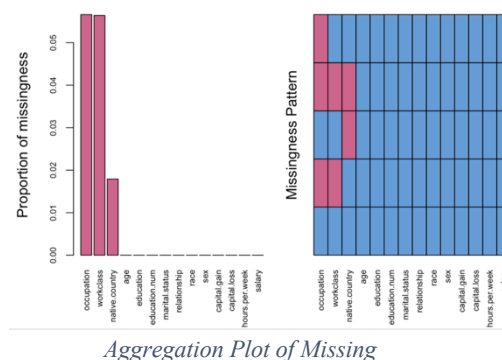


Figure 3

Aggregation Plot of Missing

The ratio of missingness and the pattern of missingness have been used to comprehend NA structure. Workclass, native country, and occupation variables in the dataset have NA values. The NA value that is highest is occupation. Margin plots shows that two-plots have the same attitude. Thus, MAR was applied for the process.

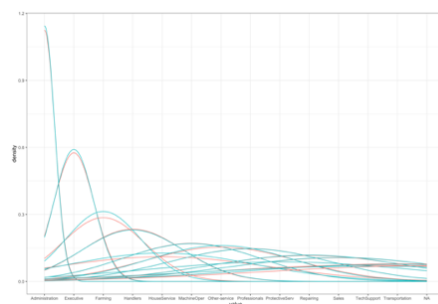


Figure 4 Density Plots of Missing

Imputation was successful since the densities are near to the density line. Additionally, the dataset does not contain any NA values.

E. One Hot Encoding – Feature Selection with Boruta

Before the model, one hot encoding was applied to categorical variables with three levels or more. After that, the Boruta method was used to create the final data set.

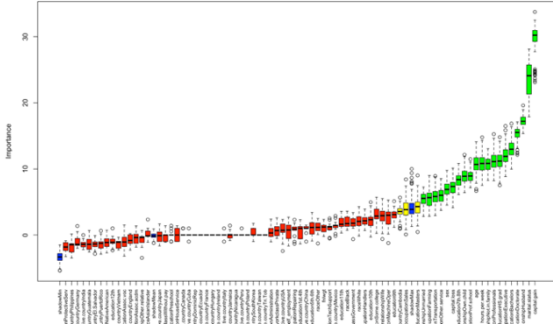


Figure 5 Feature Selection with Boruta

F. Imbalanced Data Problem

When we have unequal instances for various classes in classification problems, this is referred to as unbalanced data. In our case, it can be seen our dependent variable. This situation can be overcome by trying methods such as smote, up, down, rose and choosing the most appropriate method and applying sampling. When the accuracy is compared according to the methods in Figure 6, it is seen that the smote method is ahead.

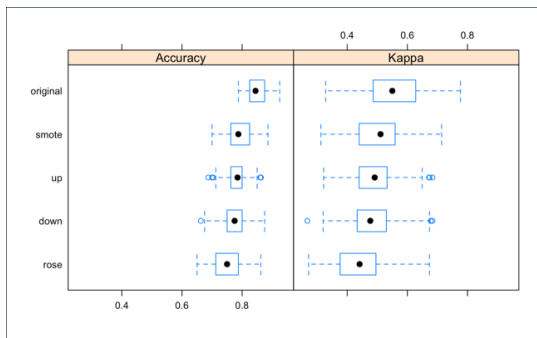


Figure 6 Accuracy Comparison of Imbalanced Data

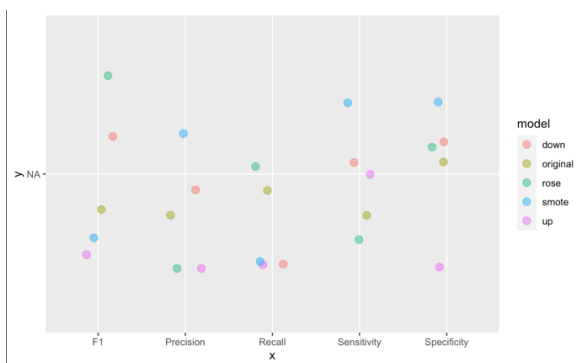


Figure 7 Methods Comparison of Imbalanced Data

In addition, when it is compared all methods in Figure 7, in terms of values such as specificity, accuracy, and F score, it was decided to use the smote method sampling because it was the closest to the original data as F score parameter and also because it had high sensitivity.

G. Modelling

New data is produced following missing imputation, feautere selection with boruta and choosing right sampling method to imbalanced data. The Salary Prediction can be classified using this new data. The data is split into train data and test data before the models are built (Cross Validation).

Based on the compatibility of the train and test sets among many cross validation techniques, repeated k-fold cross validation with 5 repeates were decided by considering the sensitiviity and also F1 score.

1. Multiple Linear Regression

Under certain assumptions, multiple linear regression is specialized to forecast the outcome of the response variable using a number of independent variables. Below is a description of the multiple linear regression model.

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p + \varepsilon$$

In this study, backward elimination is used to build the final model. Insignificant variables were eliminated from the model. Since the data is unbalanced threshold value (0.38) was used with the help of InformationValue package. Moreover, interaction terms were added to the model, but no meaningful results were found, so interaction terms were taken from the model.

Additionally, train data is employed in the multiple linear model design. The output is not displayed in this section due to the final model's large number of variables. But there are certain significant interpretations that should be mentioned.

First of all, the entire model is significant because the F statistic's p-value is less than 0.05, also considering the VIF values, it is clear that there is no multicollinearity in the model. Secondly, AUC score is 0.895. It has a rather good capacity to accurately classify attributes from the two groups. It is clear from the model's summary output that every component that hasn't been deleted is statistically significant. Lastly, interpreting coefficient, if working hours in a week increase 1 hour, the salary can be 4% change to other salary category (>50k) and if a person's occupation is about executive, the salary can be 142.66% passes to the higher salary class from 50k.

2. Artificial Neural Networks

Another machine learning algorithm that can be applied to classification problems is the artificial neural network (ANN). In this work, ANN is carried out using R studio's "keras" and "tensorflow" packages. The ANN model makes use of every variable in the train data. Prior to modeling, categorical data are converted to dummy variables using the max-min scaling method.

An ANN model's linear activation function is used in construction. Smote sampling technique was used. Ten hidden units are present in the model. The tuning for learning rate is 0.01 and for dropout rate is 0.4.

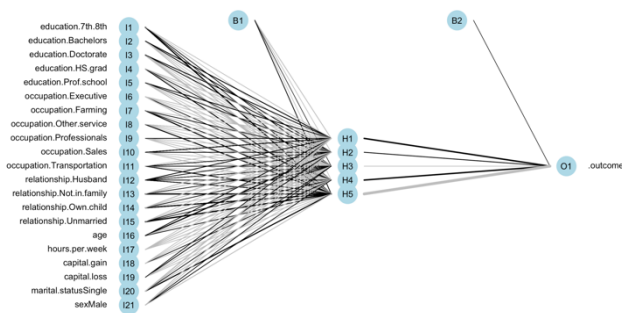


Figure 8 Plot for Artificial Neural Network

An ANN model's linear activation function is used in construction. Two layers NN model where the first layer has 21 neurons and the second one has 5 neurons. The network uses 116 weights to produce the final output.

ANN	Sensitivity	F-Score
train	0.8029	0.8647
test	0.8516	0.9010

Table 3 The Comparison of ANN

Overfitting does not appear in the model. However, since the test values are higher than the train, it might be underfitting.

3. Support Vector Machine

A machine learning approach called Support Vector Machine (SVM) can be applied to classification and regression problems. It is employed in this study to address the classification issue. Smote technique was utilized since the data sets were unbalanced. Modeling is done using a function kernel. The chosen SVM type is "svmRadial". Tuning parameter 'sigma' was held constant at a value of 0.04.

The ability to create a feature significance plot with SVM in Caret is a good feature. In SVM, the term "variable importance" refers to a metric that expresses how much each input variable (or feature) contributed to the model's ability to make decisions. It can be seen in the next plot.

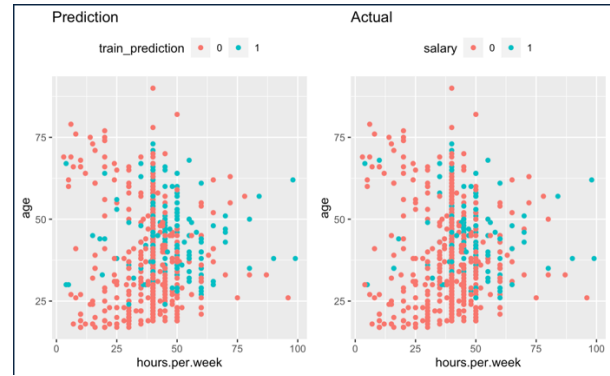


Figure 9 Plot Comparison to Age and hours per week

SVM	Sensitivity	F-Score
train	0.9031	0.9098
test	0.8387	0.8844

Table 4 The Comparison of SVM

There is neither an overfitting nor an underfitting issue when the test and train outcomes are compared.

4. Decision Tree

Regression and classification issues can both be resolved using decision trees. The algorithm can be shown as a graphical tree-like structure that predicts the outcomes using a variety of customized parameters. Smote technique was utilized since the data sets were unbalanced

ANN	Sensitivity	F-Score
train	0.9917898	0.9894
test	0.9032258	0.8833

Table 5 The Comparison of Decision Tree

There may be overfitting due to the sensitivity and F score being almost 1 in the train dataset. Also, the difference between train and the test dataset.

5. Random Forest

Classification problems can be used using the machine learning method random forest. The algorithm is based on trees. Smote technique was utilized since the data sets were unbalanced. The final model is built for this problem after the "ntree" and "mtry" have been tuned.

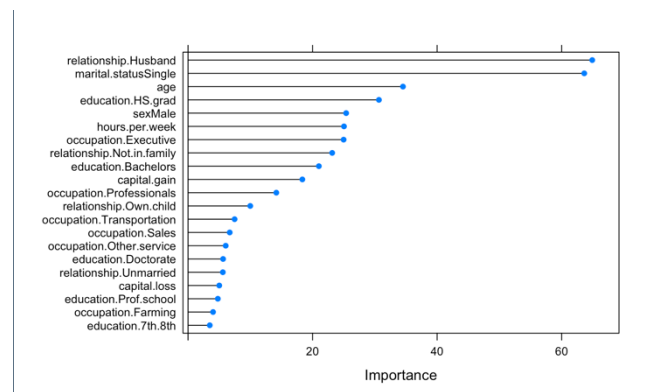


Figure 9 Plot for importance of RF

The significance of the random forest's variables is depicted in Figure 9. Important variables in the random forest model include being husband, single-marital status, age, education, sex, working hours per week and sector. The multiple linear regression model also considers the significance of the previously listed variables.

Random Forest	Sensitivity	F-Score
train	0.9031	0.9098
test	0.9032	0.9055

Table 6 The Comparison of RF

There is neither an overfitting nor an underfitting issue when the test and train outcomes are compared.

6. XGBoost

The last machine learning algorithm utilized in this work to classify salary is called Xgboost. Categorical variables were turned into dummy variables before to running the model. Smote technique was utilized since the data sets were unbalanced.

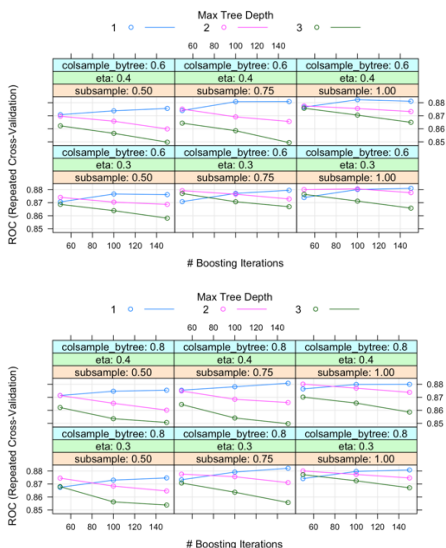


Figure 10 Plot for training process XGBoost

First, the parameter needs to be tuned. The model is built using the train data once the parameters have been adjusted and the ideal number of trees have been selected. In figure 10, It can be seen that training process for XGBoost.

XGBoost	Sensitivity	F-Score
train	0.8834	0.8997
test	0.9032	0.9121

Table 7 The Comparison of XGBoost

There is neither an overfitting nor an underfitting issue when the test and train outcomes are compared.

H. Performance Comparison on Train and Test Dataset

It's crucial to look into train performance to comprehend how well the model fits the data. It's crucial to look at test results to comprehend how the model functions with the new variables. In this study, sensitivity, F-score, accuracy and specificity are utilized to compare the performances of the test finding the minority is more important than finding the exact accuracy in such classification problems, sensitivity and F score were especially used in the performance measure for the final model. Sensitivity and F score calculations are shown in below.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F score} = 2 * \frac{(\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$$

IV. RESULTS

In this part, the results are showed for the following models;

- i. Multiple Linear Regression
- ii. Artificial Neural Networks
- iii. Support Vector Machine
- iv. Decision Tree
- v. Random Forest
- vi. XGBoost

Performance Comparison for Train Data

	Accuracy	Sensitivity	Specificity	F Score
Multiple Linear R.	0.8000	0.8571	0.6178	0.8671
Artificial Neural N.	0.8088	0.8030	0.8272	0.8647
Support Vector M.	0.8637	0.8998	0.7382	0.9073
Decision Tree	0.9837	0.9917	0.95811	0.9894
Random Forest	0.8637	0.8998	0.7329	0.9073
XGBoost	0.8500	0.8834	0.74345	0.8997

Table 8 Performance Comparison for Train Data

Performance Comparison for Test Data

	Accuracy	Sensitivity	Specificity	F Score
Multiple Linear R.	0.8350	0.8838	0.6666	0.8925
Artificial Neural N.	0.8550	0.8516	0.8666	0.9010
Support Vector M.	0.8300	0.8387	0.8000	0.8844
Decision Tree	0.8150	0.9032	0.5111	0.8833
Random Forest	0.8600	0.9032	0.7111	0.9091
XGBoost	0.8650	0.9032	0.7333	0.9121

Table 9 Performance Comparison for Test Data

From Table 8, it can be seen accuracy, sensitivity, specificity and F score for each method. All the models are constructed with train data. Moreover, it can be seen the test result from Table 9. Since, it is a classification problem and finding the minority is more crucial, so firstly sensitivity has been taken account then F score as well. Lastly, less than 50k salary (0) is a reference category for evaluation.

Initially, random forest and Xgboost method gave similar results for both train and test data. Moreover, these models do not have overfitting and underfitting problems. The best prediction for the test data is provided by XGBoost but also Random Forest gave the same result and the specificity and F score values are much closer to the train data in Random Forest. Lastly, although the Decision tree model in train dataset has the most values, the test values do not confirm this and maybe there is an overfitting problem.

V. CONCLUSION

First, exploratory data analysis is done in this piece of work. Research questions are formulated, and various graphical techniques are used to resolve them. The processing of missing data and data cleaning procedures are used then. One hot encoding and feature selection with Boruta are applied. Less characteristics are chosen with the help of Boruta and also stepwise regression. The creation of fresh data is aided by its new features, normalized response, and significant categorical factors. Then, different methods are applied to regression and decided touse "repeatedcv" with 5 repeats. Since the response which is salary variable is imbalanced data, some methods are applied and decided to use "smote" method in sampling. Salary prediction is classified using a variety of techniques using this arranged data. These techniques show that Random Forest results better sensitivity and F score, compatible with train set. Additionally, XGBoost model produce almost same encouraging results as well.

VI. REFERENCE

[1] Gopal, Krishna, et al. "Salary Prediction Using Machine Learning." *International Journal of Innovative Research in Technology*, IJIRT(Www.Ijirt.Org), 4 June 2021, ijirt.org/Article?manuscript=151548.

[2] Srivastava, Suyash, et al. "Comparing Various Machine Learning Techniques for Predicting the Salary Status." *EasyChair Home Page*, EasyChair, 10 Feb. 2020, easychair.org/publications/preprint/sRSZ.

[3] Matbouli, Yasser T., and Suliman M. Alghamdi. "Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations." *MDPI*, Multidisciplinary Digital Publishing Institute, 12 Oct. 2022, www.mdpi.com/2078-2489/13/10/495.