metin, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**Classification of Credit Score**

Eren Duralı  
*Middle East Technical University*  
e236124@metu.edu.tr

Abstract – This report aims to investigate factors which effects credit score and classification of the credit score using some machine learning techniques such as multinominal, neural network, random forest, SVM, Gradient Boost. In this paper we have 5 exploratory data analysis questions for getting know the data better and data cleaning process with missing values and how to impute, scaling with scale function, cross validation and modelling with “caret” package.

1. Introduction

Credit score has 3 levels Good, Standard and Poor. Credit score helps the banks with deciding that is it safe to give the loan or credit card to the person or not. So, these models are very useful for financial institutions. We use machine learning algorithms for helping to decision-making.

The data we took from Kaggle has some missing values and so many outliers since many of the machine learning models can deal with outliers that is not a huge issue for us but the missing values need imputations. After the imputations, we build our models and compare them with accuracy.

1. Methodology
2. Dataset

The data is taken from Kaggle it uploaded to Kaggle in 2022. Data consists of 100000 rows with 28 column some columns are unnecessary for the analysis such as ID, Customer id, ssn etc. So, we just extract that columns and we continue with 22 columns. With 21 columns we just try to classify credit score.

1. Descriptive Statistics

Table 1: Summary statistics all data

This is the summary statistics with all the data we get after deleting the columns which not useful for our modelling or making the complex the modelling process and with several filter for outliers and wrong values our summary statistics is given below:

metin, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturulduTable 2: Summary statistics with used data

After some filtering and extract the columns (id, customer\_id, ssn, occupation, name) we were ready for dealing with missing values.

1. metin, ekran görüntüsü, tasarım, sanat içeren bir resim

   Açıklama otomatik olarak oluşturulduMissing Imputation

Figure 1: Missing plot

We have 34113 missing values and most of the missing values is in monthly in hand salary so we check that column first. We find that if we omit all the missing values, we will have 45656 row so we think that is a huge waste of data so we decide the impute that. In the Missing value rates table (Table 3) you can see that monthly in hand salary column consist of 15% missing values. So we plot these missing values by using “mice” package (Figure 1).

ekran görüntüsü, metin içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturulduWe use the imputation by “na.approx” function which is in the zoo library after the imputation we still have 6647 missing values and all of them is in credit history age it is reasonable because of that “na.approx” function only used for continuous variable and credit history age was a character column. This column consists of year and months so we write a function to convert this columns just months and make this column continuous after that we use “na.approx” for this column too. After that process we do not have any missing values and data was ready for the EDA.

**Table 3: Missing value rates**

1. Explanatory Data Analysis

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturulduQ1) How does annual income effect the number of bank accounts?

Figure 2: Annual income vs Number of Bank Accounts

Annual income is a continuous variable which can think of richness. And this bar plot shows that when annual income doesn’t have a linear relationship between number of bank accounts as you see in the graph people who has lower annual incomes has extreme number of bank accounts it is either too low or too much. But people who has greater annual income is in the middle of this graph mostly.

Figure 3: Linear regression model

After we mention that linear relationship, I also run a linear model between these two variables. And p-value is .069 is not significant for alpha = 0.05 and also you can see that nor annual income neither num of bank accounts is significant for this model so we can conclude that these two variables have no linear relationship between them.

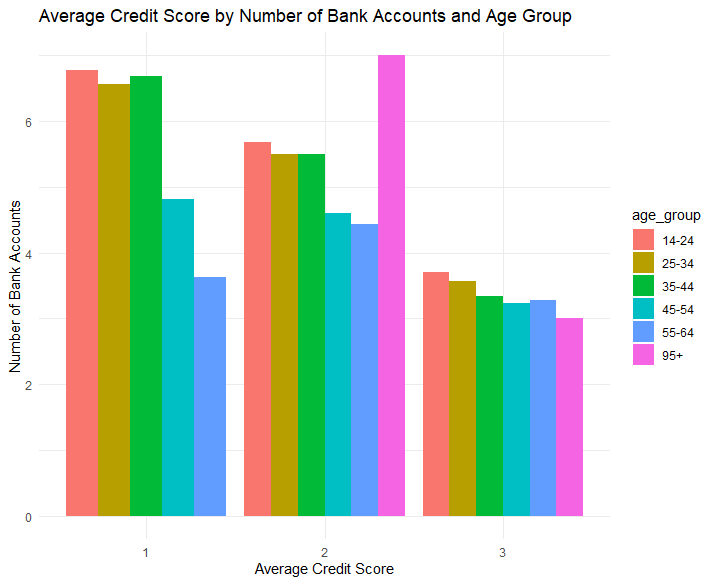
Q2) How does the number of bank accounts influence the credit score across different age groups?

Figure 4: Average Credit Score by Number of Bank   
Accounts and Age Group

In this graph we can see the distribution of number of bank accounts between different age groups in 3 levels of credit score. “1” represents “Poor” credit score, “2” represents “Standard” and “3” represents "Good” credit score as you can see people who have good credit scores has less bank accounts in all age groups and we can see the correlation between number of bank accounts and credit score by looking at “Poor” credit score which has the highest number of bank accounts between 14-44. Lastly, we can say that younger people have more tendency for bank accounts than older peoples.

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturulduQ3) Is there a significant association between the type of loan (e.g., mortgage, car loan, personal loan) and the frequency of late payments?

**** Figure 5: Chi-squared test

We Build a contingency table for conducting a chi-squared test. Since p-value of the Chi-squared test < 0.05. We reject the null hypothesis which is the variables are independent of each other. This means that there is an association between the type of loan and the frequency of late payments.

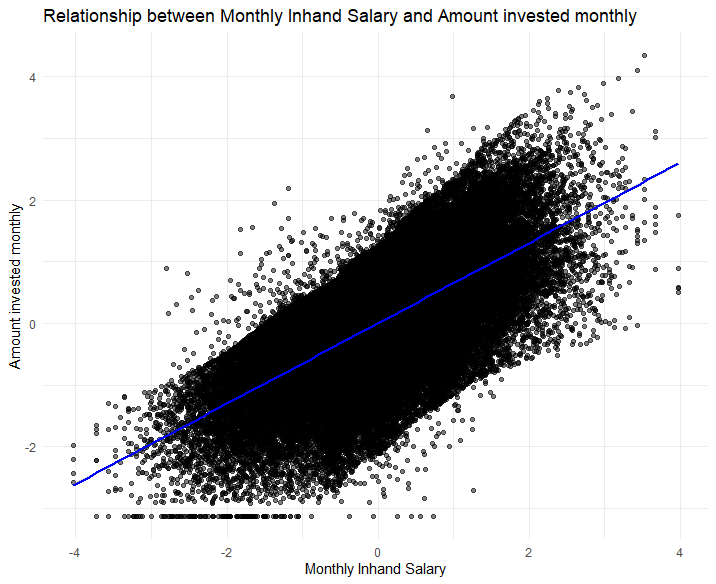
Q4) How does the monthly in-hand salary influence the amount invested monthly?

Figure 6: Linear relationship between Monthly in-hand salary and amount invested monthly.

We conduct a linear model with target variable amount invested monthly and response variable as monthly in-hand salary. Before the linear model we check for normality they were not follow normal distribution after the BestNormalize function process they check the assumption that follow normal distribution. In this plot we can clearly see that there is a positive linear relationship between monthly in-hand salary and amount invested monthly. This makes sense since if you have more money you can invest more.

1. Modelling

For modelling we use “caret” package. We encode all our categorical variables and normalize all continuous variables. Then separate the data as train and test by 80% to 20% respectively. We use the same cross-validation for all models which is 5-fold-cross-validation and also, we use sampling “up” because of that our credit score variables was imbalanced it is not perfectly balanced but it helps. We could not use SMOTE because DNwR package was not working.

1. metin, ekran görüntüsü, yazı tipi içeren bir resim

   Açıklama otomatik olarak oluşturulduMultinominal Model

Figure 7: Multinom Confusion Matrix

With a Kappa statistic of 0.3841 and an overall accuracy of 60.15% (95% CI: 0.5978 - 0.6051), the multinomial model for credit score prediction shows moderate agreement above random variation. It shows higher sensitivity for Classes 1 and 3, successfully identifying 65.71% of Class 1, 54.04% of Class 2, and 69.43% of Class 3 occurrences. For Classes 1, 2, and 3, the corresponding specificity values are 0.7950, 0.7876, and 0.8131, demonstrating strong negative identification. The values that are positive are 62.18%, 71.81%, and 52.65%, whereas the ones that are negative are 84.95%, 60.58%, and 92.65%. For Classes 1, 2, and 3, the balanced accuracy scores are 0.7261, 0.6640, and 0.7537, indicating the model's balanced performance. Class 2 predictions are useful, but they could be better.

1. metin, ekran görüntüsü, yazı tipi içeren bir resim

   Açıklama otomatik olarak oluşturulduNeural Network

Figure 8: Neural Network Confusion Matrix

With a Kappa statistic of 0.3948 and an overall accuracy of 59.82% (95% CI: 0.5945 - 0.6018) for credit score prediction, the neural network model shows moderate agreement above random variation. It demonstrates superior sensitivity for Classes 1 and 3, properly identifying 73.94% of Class 1, 48.19% of Class 2, and 71.63% of Class 3 occurrences. The specificity values for Classes 1, 2, and 3 are 0.7630, 0.8545, and 0.7987, in that order. Positive predictive values are 76.10%, 75.07%, and 51.60%, whereas negative predictive values are 87.69%, 69.14%, and 93.79%. For Classes 1, 2, and 3, the balanced accuracy scores are 0.7512, 0.6682, and 0.7575. The model's overall effectiveness is mediocre, and its Class 2 forecasts might use some work.

1. metin, ekran görüntüsü, yazı tipi, menü içeren bir resim

   Açıklama otomatik olarak oluşturulduSVM

Figure 9: SVM Confusion Matrix

The SVM model's performance, as represented by the confusion matrix and accompanying statistics, shows a mixed level of accuracy in classifying the three classes. The overall accuracy is 60.57%, with a 95% confidence interval ranging from 60.21% to 60.94%. This indicates that the model's performance is significantly better than random guessing, as evidenced by the P-Value being less than 2.2e-16.

Class-wise analysis reveals varying degrees of effectiveness. Class 1 has the highest sensitivity (0.6771) and balanced accuracy (0.7296), indicating it is relatively well-detected. Class 2, while having a lower sensitivity (0.5693), benefits from a high specificity (0.8539), reflecting its fewer false positives. Class 3 shows a balanced performance with a sensitivity of 0.5981 and a balanced accuracy of 0.7233. The model's Kappa statistic of 0.3784 suggests moderate agreement beyond chance. Overall, the SVM model demonstrates competent classification abilities with room for improvement in certain areas, particularly in handling class imbalances and enhancing sensitivity for underrepresented classes.

1. metin, ekran görüntüsü, yazı tipi içeren bir resim

   Açıklama otomatik olarak oluşturulduRandom Forest

Figure 10: Random Forest Confusion Matrix

With a Kappa statistic of 0.4939 and an overall accuracy of 67.84% (95% CI: 0.6749 - 0.6819) for credit score prediction, the random forest model shows moderate to strong agreement that goes beyond chance. 73.08% of Class 1, 63.00% of Class 2, and 73.80% of Class 3 instances are accurately identified by it, demonstrating better sensitivity in all classes when compared to earlier models. The specificity values for Classes 1, 2, and 3 are, respectively, 0.8395, 0.8133, and 0.8356. 83.16%, 65.85%, and 93.84% are the negative predictive values, whereas 65.16%, 79.37%, and 51.86% are the positive values. For Classes 1, 2, and 3, the balanced accuracy scores are 0.7852, 0.7217, and 0.7958. All things considered, the model performs better, especially when it comes to Class 2 forecasts.

1. Model Comparison
2. Random Forest
3. SVM
4. Multinominal
5. Neural Network

Models ranks by accuracy is this. We can conclude that Random Forest is best for credit score classification.

1. Conclusion

We conclude that Random Forest is the best classification for credit score for this dataset. But in other and further studies this might change. We have so many news machine learning algorithms which might give better results from those algorithms. Also, data gathering is the important part of these studies and I know that financial institutions are better with this. As a result, Machine learning algorithms might be nearly perfect for classification of credit score with the better data gathering, data pre-processing and modeling.

1. References

[1] Credit score classification. (2022, June 22). Kaggle. https://www.kaggle.com/datasets/parisrohan/credit-score-classification/data?select=train.csv