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.

I. 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.

II. Methodology

A. 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.

B. Descriptive Statistics

```
      ID          Customer_ID           Month               Name                 Age              SSN              Occupation
Min.   :  5634   Length:100000      Length:100000      Length:100000      Min.   :-500.0   Length:100000      Length:100000
1st Qu.: 43133   Class :character   Class :character   Class :character   1st Qu.:  24.0   Class :character   Class :character
Median : 80632   Mode  :character   Mode  :character   Mode  :character   Median :  33.0   Mode  :character   Mode  :character
Mean   : 80632                                                            Mean   : 110.9
3rd Qu.:118130                                                            3rd Qu.:  42.0
Max.   :155629                                                            Max.   :8698.0
                                                                          NA's   :4939
Annual_Income       Monthly_Inhand_Salary Num_Bank_Accounts Num_Credit_Card   Interest_Rate      Num_of_Loan       Type_of_Loan
Min.   :     7006   Min.   :  303.6       Min.   :  -1.00   Min.   :   0.00   Min.   :   1.00   Min.   :-100.00   Length:100000
1st Qu.:    19436   1st Qu.: 1625.6       1st Qu.:   3.00   1st Qu.:   4.00   1st Qu.:   8.00   1st Qu.:   1.00   Class :character
Median :    37551   Median : 3093.7       Median :   6.00   Median :   5.00   Median :  13.00   Median :   3.00   Mode  :character
Mean   :   178579   Mean   : 4194.2       Mean   :  17.09   Mean   :  22.47   Mean   :  72.47   Mean   :   2.78
3rd Qu.:    72843   3rd Qu.: 5957.4       3rd Qu.:   7.00   3rd Qu.:   7.00   3rd Qu.:  20.00   3rd Qu.:   5.00
Max.   :24198062   Max.   :15204.6       Max.   :1798.00   Max.   :1499.00   Max.   :5797.00   Max.   :1496.00
NA's   :6980       NA's   :15002                                                                 NA's   :4785
Delay_from_due_date Num_of_Delayed_Payment Changed_Credit_Limit Num_Credit_Inquiries  Credit_Mix         Outstanding_Debt
Min.   :-5.00       Min.   :  -3.00        Min.   :-6.49        Min.   :   0.00      Length:100000      Min.   :   0.23
1st Qu.:10.00       1st Qu.:   9.00        1st Qu.: 5.32        1st Qu.:   3.00      Class :character   1st Qu.: 566.08
Median :18.00       Median :  14.00        Median : 9.40        Median :   6.00      Mode  :character   Median :1166.37
Mean   :21.07       Mean   :  31.03        Mean   :10.39        Mean   :  27.75                         Mean   :1426.50
3rd Qu.:28.00       3rd Qu.:  18.00        3rd Qu.:14.87        3rd Qu.:   9.00                         3rd Qu.:1948.20
Max.   :67.00       Max.   :4397.00        Max.   :36.97        Max.   :2597.00                         Max.   :4998.07
                    NA's   :9746           NA's   :2091         NA's   :1965                            NA's   :1009
Credit_Utilization_Ratio Credit_History_Age Payment_of_Min_Amount Total_EMI_per_month Amount_invested_monthly Payment_Behaviour
Min.   :20.00            Length:100000      Length:100000         Min.   :    0.00    Min.   :   0.00         Length:100000
1st Qu.:28.05            Class :character   Class :character      1st Qu.:   30.31    1st Qu.:  72.24         Class :character
Median :32.31            Mode  :character   Mode  :character      Median :   69.25    Median : 128.96         Mode  :character
Mean   :32.29                                                     Mean   : 1403.12    Mean   : 195.54
3rd Qu.:36.50                                                     3rd Qu.:  161.22    3rd Qu.: 236.82
Max.   :50.00                                                     Max.   :82331.00    Max.   :1977.33
                                                                                      NA's   :8784

Monthly_Balance     Credit_Score
Min.   :   0.0078   Length:100000
1st Qu.: 270.1066   Class :character
Median : 336.7312   Mode  :character
Mean   : 402.5513
3rd Qu.: 470.2629
Max.   :1602.0405
NA's   :1209
```

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:

```
   month               age          annual_income       monthly_inhand_salary num_bank_accounts num_credit_card   interest_rate
Length:74179       Min.   : 14.0   Min.   :     7006   Min.   :  303.6        Min.   : 0.000    Min.   :   0.00   Min.   :   1.00
Class :character   1st Qu.: 24.0   1st Qu.:    19436   1st Qu.: 1628.3        1st Qu.: 3.000    1st Qu.:   4.00   1st Qu.:   8.00
Mode  :character   Median : 33.0   Median :    37552   Median : 3096.0        Median : 6.000    Median :   5.00   Median :  14.00
                   Mean   : 33.3   Mean   :   177933   Mean   : 4199.5        Mean   : 5.372    Mean   :  22.72   Mean   :  73.67
                   3rd Qu.: 42.0   3rd Qu.:    72850   3rd Qu.: 5968.2        3rd Qu.: 7.000    3rd Qu.:   7.00   3rd Qu.:  20.00
                   Max.   :142.0   Max.   :24198062   Max.   :15204.6        Max.   :11.000    Max.   :1499.00   Max.   :5797.00
                                   NA's   :5195       NA's   :11150
 num_of_loan     type_of_loan       delay_from_due_date num_of_delayed_payment changed_credit_limit num_credit_inquiries
Min.   : 0.000   Length:74179       Min.   : 0.00       Min.   : 0.00          Min.   : 0.00        Min.   :   0.00
1st Qu.: 2.000   Class :character   1st Qu.:10.00       1st Qu.: 9.00          1st Qu.: 5.36        1st Qu.:   3.00
Median : 3.000   Mode  :character   Median :18.00       Median :14.00          Median : 9.42        Median :   6.00
Mean   : 3.537                      Mean   :21.11       Mean   :13.34          Mean   :10.47        Mean   :  27.56
3rd Qu.: 5.000                      3rd Qu.:28.00       3rd Qu.:18.00          3rd Qu.:14.90        3rd Qu.:   9.00
Max.   :23.000                      Max.   :67.00       Max.   :98.00          Max.   :36.49        Max.   :2597.00
                                                                               NA's   :1558         NA's   :1445
 credit_mix         outstanding_debt  credit_utilization_ratio credit_history_age payment_of_min_amount total_emi_per_month
Length:74179       Min.   :   0.23   Min.   :20.00            Length:74179       Length:74179          Min.   :    0.00
Class :character   1st Qu.: 568.81   1st Qu.:28.06            Class :character   Class :character      1st Qu.:   30.44
Mode  :character   Median :1167.20   Median :32.31            Mode  :character   Mode  :character      Median :   69.28
                   Mean   :1426.94   Mean   :32.30                                                     Mean   : 1409.63
                   3rd Qu.:1950.21   3rd Qu.:36.51                                                     3rd Qu.:  161.37
                   Max.   :4998.07   Max.   :50.00                                                     Max.   :82331.00
                   NA's   :754
amount_invested_monthly payment_behaviour  monthly_balance     credit_score
Min.   :   0.00         Length:74179       Min.   :   0.0078   Length:74179
1st Qu.:  72.17         Class :character   1st Qu.: 269.9876   Class :character
Median : 128.97         Mode  :character   Median : 336.8337   Mode  :character
Mean   : 195.48                            Mean   : 402.5666
3rd Qu.: 236.58                            3rd Qu.: 469.9375
Max.   :1977.33                            Max.   :1602.0405
NA's   :6473                               NA's   :892
```

Table 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.

C. Missing Imputation

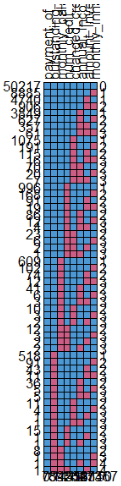

Figure 1: Missing plot

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).

We 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.

age	annual_income	monthly_inhand_salary	num_bank_accounts	num_credit_card
0.00000000	0.07003330	0.15031208	0.00000000	0.00000000
interest_rate	num_of_loan	delay_from_due_date	num_of_delayed_payment	changed_credit_limit
0.00000000	0.00000000	0.00000000	0.00000000	0.02100325
num_credit_inquiries	outstanding_debt	credit_utilization_ratio	total_emi_per_month	amount_invested_monthly
0.01947991	0.01016460	0.00000000	0.00000000	0.08726189
monthly_balance	month	type_of_loan	credit_mix	credit_history_age
0.01202497	0.00000000	0.00000000	0.00000000	0.08959409
payment_of_min_amount	payment_behaviour	credit_score		
0.00000000	0.00000000	0.00000000		

Table 3: Missing value rates

D. Explanatory Data Analysis

Q1) How does annual income effect the number of bank accounts?

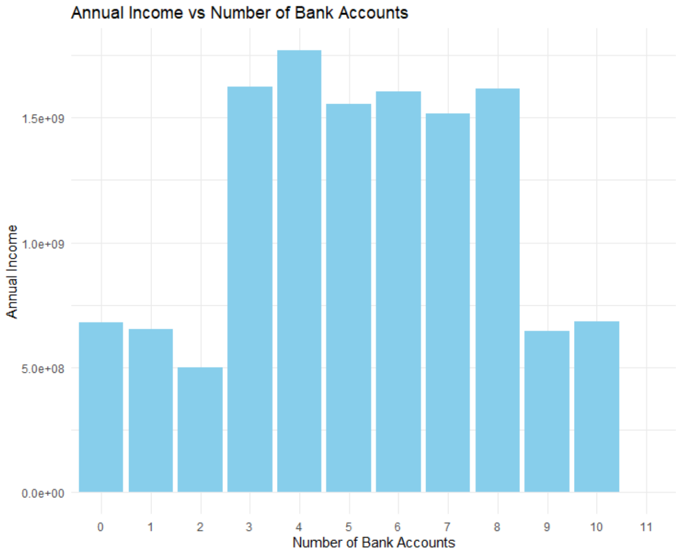

Figure 2: Annual income vs Number of Bank Accounts

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.

```
Call:
lm(formula = annual_income ~ num_bank_accounts, data = credit5_norm_num_sc)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.1313 -0.1100 -0.0986 -0.0736 16.8019 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)  
(Intercept)       -1.337e-16  3.741e-03   0.000   1.0000  
num_bank_accounts -6.783e-03  3.741e-03  -1.813   0.0699 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1 on 71438 degrees of freedom
Multiple R-squared:  4.6e-05,	Adjusted R-squared:  3.201e-05 
F-statistic: 3.287 on 1 and 71438 DF,  p-value: 0.06985
```

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?

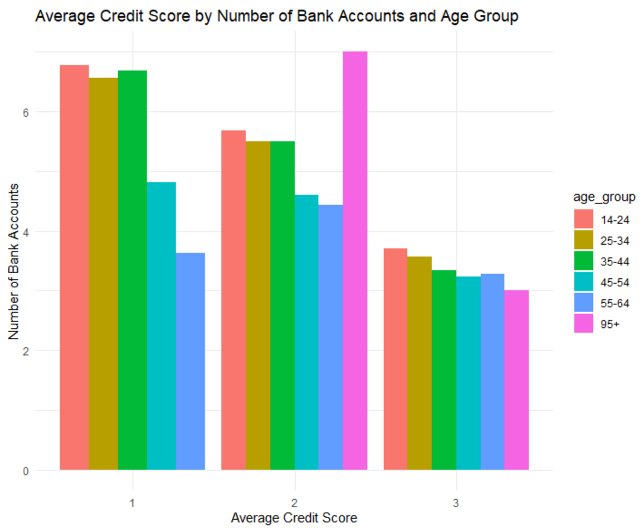

Figure 4: Average Credit Score by Number of Bank
Accounts and Age Group

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.

Q3) Is there a significant association between the type of loan (e.g., mortgage, car loan, personal loan) and the frequency of late payments?

```
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 586953, df = 231583, p-value < 2.2e-16
X-squared
 17.11059
```

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.

III. 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.

A. Multinominal Model

```
> confusionMatrix(predictions_pca, test_data$credit_score)
Confusion Matrix and Statistics

          Reference
Prediction     1     2     3
         1 13210  9054   976
         2  4112 19874  2741
         3  2782  7850  8440

Overall Statistics

               Accuracy : 0.6015
                 95% CI : (0.5978, 0.6051)
    No Information Rate : 0.5327
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3841

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.6571   0.5404   0.6943
Specificity            0.7950   0.7876   0.8131
Pos Pred Value         0.5684   0.7436   0.4425
Neg Pred Value         0.8495   0.6005   0.9256
Prevalence             0.2912   0.5327   0.1761
Detection Rate         0.1913   0.2879   0.1222
Detection Prevalence   0.3366   0.3871   0.2762
Balanced Accuracy      0.7261   0.6640   0.7537
```

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.

B. Neural Network

```
> confusionMatrix(predictions_nn, test_data$credit_score)
Confusion Matrix and Statistics

          Reference
Prediction     1     2     3
         1 14864 10546  1052
         2  2296 17725  2397
         3  2944  8507  8708

Overall Statistics

               Accuracy : 0.5982
                 95% CI : (0.5945, 0.6018)
    No Information Rate : 0.5327
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3948

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.7394   0.4819   0.7163
Specificity            0.7630   0.8545   0.7987
Pos Pred Value         0.5617   0.7907   0.4320
Neg Pred Value         0.8769   0.5913   0.9294
Prevalence             0.2912   0.5327   0.1761
Detection Rate         0.2153   0.2567   0.1261
Detection Prevalence   0.3833   0.3247   0.2920
Balanced Accuracy      0.7512   0.6682   0.7575
```

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.

C. SVM

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2     3
         1 13613  9538  1129
         2  4182 20936  3757
         3  2309  6304  7271

Overall Statistics

               Accuracy : 0.6057
                 95% CI : (0.6021, 0.6094)
    No Information Rate : 0.5327
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3784

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.6771   0.5693   0.5981
Specificity            0.7820   0.7539   0.8486
Pos Pred Value         0.5607   0.7251   0.4578
Neg Pred Value         0.8550   0.6056   0.9081
Prevalence             0.2912   0.5327   0.1761
Detection Rate         0.1972   0.3032   0.1053
Detection Prevalence   0.3517   0.4182   0.2301
Balanced Accuracy      0.7296   0.6616   0.7233
```

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.

D. Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2     3
         1 14692  7541   313
         2  3151 23171  2872
         3  2261  6066  8972

Overall Statistics

               Accuracy : 0.6784
                 95% CI : (0.6749, 0.6819)
    No Information Rate : 0.5327
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4939

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.7308   0.6300   0.7380
Specificity            0.8395   0.8133   0.8536
Pos Pred Value         0.6516   0.7937   0.5186
Neg Pred Value         0.8836   0.6585   0.9384
Prevalence             0.2912   0.5327   0.1761
Detection Rate         0.2128   0.3356   0.1300
Detection Prevalence   0.3266   0.4229   0.2506
Balanced Accuracy      0.7852   0.7217   0.7958
```

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.

E. Model Comparison

1) Random Forest
2) SVM
3) Multinominal
4) Neural Network

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

IV. 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.

V. References

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