Prediction Model for German Credit Score data

By: Mehta Yash Piyush

Table of Contents

1. INTRODUCTION

1.1 Background and Context

Data science is a multidisciplinary field that involves the use of scientific methods, processes, and systems to extract knowledge and insights from data (Dhar, 2017).

It combines principles from computer science, statistics, and domain expertise to analyze and interpret complex data sets, and involves the use of various techniques and tools, such as machine learning and data visualization (Witten, et al., 2011). The field of statistics, on the other hand, involves the collection, analysis, interpretation, presentation, and organization of data (Agresti, 2013). It helps us understand the underlying patterns and trends in data, and make informed decisions based on statistical evidence (Pallant, 2011).

The application of data science and statistics can be seen in a variety of fields, including finance, healthcare, marketing, and education. For example, in finance, data science techniques can be used to analyze stock market trends and predict future performance (Hand, 2013) while in healthcare, statistical analysis can be used to identify risk factors for diseases and evaluate the effectiveness of treatments (Siddharthan, 2018). In recent years, the proliferation of big data and advances in computing have fueled the growth of data science and its applications (Manyika, et al., 2011)

This report will analyse the German credit scoring data using appropriate statistic techniques and data science algorithms to predict Credit Risk. The structure of this report is similar to the sequential process of data science: Data Exploration, Pre-processing, Analysis and Evaluation. We have used random forest and support vector machine to analyse our dataset. The performance of these models has been measured in terms of Accuracy, Precision, Recall, F-score and MCC.

1.2 Structure of the Report

This report will cover the methodology illustrating the procedures and measures taken to thoroughly analyse and evaluate this dataset. Thereafter, the results will be presented and a conclusion would be provided. Moreover, an evaluation of the different methods and possible further improvements will be described towards the end of the report.

2. METHODOLOGY

There are mainly six steps in data science viz: Framing the problem, Collection of data, Processing the data for analysis, Exploratory data analysis (EDA), Performing in-depth analysis and communicating the results of the analysis (Bansal, 2022). These steps have been accomplished and will be shown in the following sub-sections.

2.1 Framing the Problem

We have to predict the Credit Risk column, which is our target variable. This is a binary variable, with 0 representing the credit contract has been voided and 1 representing the credit contract has been adhered to. To predict Credit Risk, a categorical variable, we have used Random Forest, and Support Vector machines after conducting data pre-processing and exploratory data analysis.

2.2 Collection of the dataset

The data was in a .csv format which could be easily read by R using the read.csv function. There was no need to convert it to any other format.

2.3. Processing of the data for analysis

An essential step in a data science project is data pre-processing. One of the first tasks here is to obtain an appropriate grasp and understanding about the data. Therefore, firstly, data verification and validation is performed as follows

2.3.1 Data Verification

For verifying the data, the dimension of the loaded data was checked using the dim() command. We obtained 1,000 rows and 21 columns, which was stated in the assignment brief. Thereafter, using the names() command, the column headers were obtained. Then, in conjunction with str(), the View() command was utilized and the dataset was manually observed and the data type for each column was recorded in the table below, along with the column description

Table 1: Data Analysis for German Credit Score

SN	Column Name	Description	Data Classification	Initial Hypothesis (based on domain knowledge)	Importance of this variable for predicting Credit Risk
1	Status	Status of the debtor's checking account with the bank	Categorical - Ordinal	Higher the status, better it is for the bank.	3
2	Duration	Duration bank has provided credit to the debtor	Numerical	Lower the duration, better it is for the bank.	2
3	Credit History	Debtor's history of compliance with credit contracts	Categorical - Ordinal	Higher the credit history, better it is for the bank.	3
4	Purpose	Reason why debtor requires credit	Categorical – Nominal	Purpose is nominal variable so no preference.	1

5	Amount	Amount of credit debtor owes. This is in the Deutsche Mark currency.	Numerical	Lower the amount, better it is for the bank.	3
6	Savings	This is the total savings of the debtor.	Categorical - Ordinal	Higher the savings, better it is for the bank.	3
7	Employment Duration	How long the debtor is under employment with current employer	Categorical - Ordinal	Greater the employee duration, better it is for the bank.	2
8	Installment rate	Number of credit installments debtor has to pay as a % of debtor's disposable income.	Categorical - Ordinal	Higher the installment rate, the better it is for the bank.	2
9	Personal status and sex	Encoded variable of sex and marital status of debtor	Categorical - Nominal	Purpose is nominal variable so no preference.	In this case, 1 because it is hard to separate sex and marital status due to how it was encoded.
10	Other debtors	Answers the question: Are there any guarantors or debtors for the credit.	Categorical - Ordinal	If there are any guarantors or co-debtors, the better it is for the bank. (higher the better)	2
11	Present residence	Duration the debtor lives in current residence (in years)	Categorical - Ordinal	Greater the duration the debtor lives in current residence, the better it is for the bank. (higher the better)	2
12	Property	Debtor's most valuable property is highlighted	Categorical - Ordinal	Greater the value of property the debtor owns, the better it is for the bank. (higher the better)	3
13	Age	Age of the debtor (in years)	Numerical	Results here often vary for different age group by demographic characteristics, country and other factors and there are research backing the 2 diametric standpoints, making it inconclusive. We have to rely on the statistics for this data.	2

14	Other installment plans	Answers the question: Is the debtor on installment plans with any other bank?	Categorical - Ordinal	If the debtor is not on other installment plans, the better it is for the bank. (higher the better)	2
15	Housing	Type of housing debtor lives in	Categorical - Ordinal	If the debtor owns the housing, the better it is for the bank (higher the better).	2
16	Number of credits	Number of credits the debtor has with the credit-giving bank	Categorical - Ordinal	If the debtor has lower credits with current bank, the better it is for the bank (lower the better).	3
17	Job	Answers the question: Is the debtor on installment plans with any other bank?	Categorical - Ordinal	If the debtor has no installment plans with any other banks, the better it is for the bank (higher the better).	2
18	People Liable	Number of people that are financially dependent on the debtor	Categorical - Binary	Lesser the number of people financially dependent on debtor, the better it is for the bank (higher the better).	2
19	Telephone	Answers the question: Is there a telephone landline registered on the debtor's name	Categorical - Binary	If customer owns their telephone landline, the better it is for the bank (higher the better).	1
20	Foreign worker	Is the debtor a foreign worker?	Categorical – Binary	If debtor is a local worker, the better it is for the bank (higher the better).	2
21	Credit Risk	Has the debtor complied with the current credit contract with the credit-giving bank?	Categorical - Binary	If credit contract is complied with the bank, the better it is for the bank (higher the better).	

In the second column of the table, we have the column name and its explanation in the third column. In the fourth column of this table, we have identified the data type as it will help in understanding which analysis model, we can apply based on our end-goal.

The last two columns are where domain knowledge and real-life common sense has been applied. In the fifth column, we have tried to hypothesize how the direction of the certain variable will move with credit risk (if better for bank, that means there will be a higher credit risk score – 1 and vice versa). The last column ranks the importance of this variable on credit risk from a scale of 1 to 3, 1 being the lowest and 3 being the highest. As we will see later on, this would be useful for the feature engineering and feature extraction process.

2.3.2. Data Validation

The data was then checked for both NA values and null values. Firstly, the data was skimmed through manually and then the is.na() and is.null() functions were applied and only "FALSE" was noticed. This result was confirmed once again by a graphical test using the vis_miss() function. Code snippet is as follows.

```
#Reading csv file
cred_raw <- read.csv(file = "German_Credit_data.csv", header = TRUE)

##UNDERSTANDING DATA

#Plotting the 5-number statistical summary of our data
summary(cred_raw)

#Understanding structure of our data
str(cred_raw)

#Understanding dimensions of data (number of rows and number of columns)
dim(cred_raw)

#Obtaining the column headers of our data
names(cred_raw)

##DATA CLEANING

#Checking for NA values (Missing values) by graphical method
vis_miss(cred_raw)

#Confirming the above and checking for NA or NULL values by eye test
is.na(cred_raw)
is.null(cred_raw)
```

Figure 1: Data Check for missingness

If there were some missing data, the next step would be checking whether it was MCAR or not and then applying the appropriate imputation technique by mean or median, or conducting a listwise deletion technique. Now the data was ready for further analysis.

2.4 Exploratory data analysis

EDA was conducted in two parts: Graphically and non-graphically. This was done both for the numerical features present in the dataset, via a correlation matrix and heatmap and for the categorical features through a contingency table.

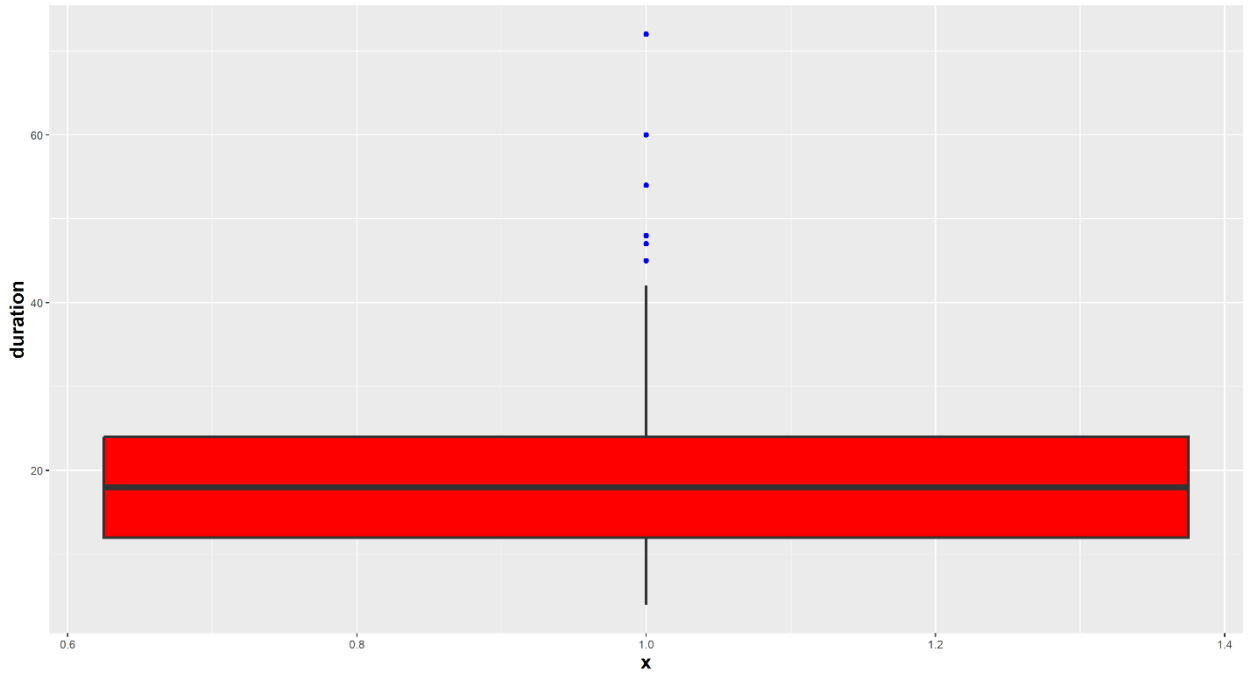
This process of conducting EDA is important to better understand the data, uncover patterns and relationships and check assumptions of the method we are using for our analysis. The first step in this process would be to utilize the summary() function and get the 5 number summary for our variables.

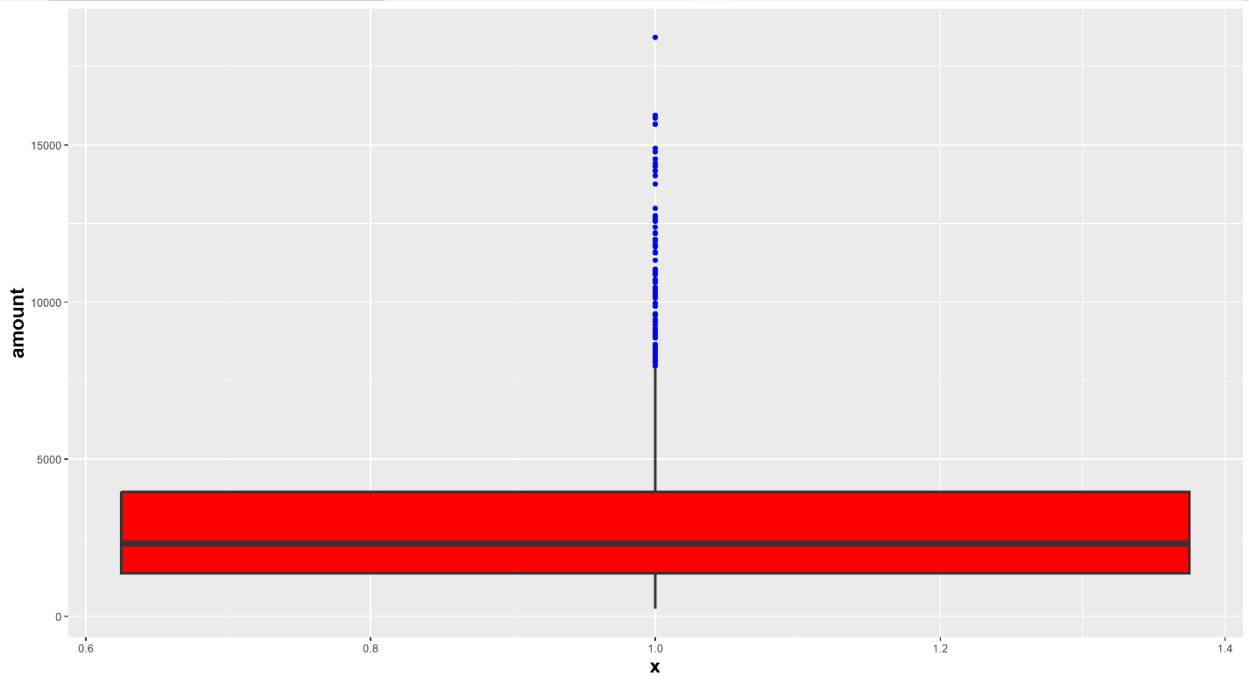
```
> #Plotting the 5-number statistical summary of our data
> summary(before_cred)
     status         duration     credit_history     purpose
 Min.   :1.000   Min.   : 4.0   Min.   :0.000   Min.   : 0.000
 1st Qu.:1.000   1st Qu.:12.0   1st Qu.:2.000   1st Qu.: 1.000
 Median :2.000   Median :18.0   Median :2.000   Median : 2.000
 Mean   :2.577   Mean   :20.9   Mean   :2.545   Mean   : 2.828
 3rd Qu.:4.000   3rd Qu.:24.0   3rd Qu.:4.000   3rd Qu.: 3.000
 Max.   :4.000   Max.   :72.0   Max.   :4.000   Max.   :10.000
     amount         savings      employment_duration installment_rate
 Min.   :  250   Min.   :1.000   Min.   :1.000       Min.   :1.000
 1st Qu.: 1366   1st Qu.:1.000   1st Qu.:3.000       1st Qu.:2.000
 Median : 2320   Median :1.000   Median :3.000       Median :3.000
 Mean   : 3271   Mean   :2.105   Mean   :3.384       Mean   :2.973
 3rd Qu.: 3972   3rd Qu.:3.000   3rd Qu.:5.000       3rd Qu.:4.000
 Max.   :18424   Max.   :5.000   Max.   :5.000       Max.   :4.000
 personal_status_sex other_debtors   present_residence    property
 Min.   :1.000       Min.   :1.000   Min.   :1.000     Min.   :1.000
 1st Qu.:2.000       1st Qu.:1.000   1st Qu.:2.000     1st Qu.:1.000
 Median :3.000       Median :1.000   Median :3.000     Median :2.000
 Mean   :2.682       Mean   :1.145   Mean   :2.845     Mean   :2.358
 3rd Qu.:3.000       3rd Qu.:1.000   3rd Qu.:4.000     3rd Qu.:3.000
 Max.   :4.000       Max.   :3.000   Max.   :4.000     Max.   :4.000
      age        other_installment_plans    housing      number_credits
 Min.   :19.00   Min.   :1.000           Min.   :1.000   Min.   :1.000
 1st Qu.:27.00   1st Qu.:3.000           1st Qu.:2.000   1st Qu.:1.000
 Median :33.00   Median :3.000           Median :2.000   Median :1.000
 Mean   :35.54   Mean   :2.675           Mean   :1.928   Mean   :1.407
 3rd Qu.:42.00   3rd Qu.:3.000           3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :75.00   Max.   :3.000           Max.   :3.000   Max.   :4.000
      job        people_liable     telephone      foreign_worker
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000
 Median :3.000   Median :2.000   Median :1.000   Median :2.000
 Mean   :2.904   Mean   :1.845   Mean   :1.404   Mean   :1.963
 3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :4.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
```

Figure 2: Summary Statistics for German Credit Data

2.4.1. Box plot to detect outliers

An outlier in a sample survey is an observation that differs significantly from the majority or all the other observations (Ghosh & Vogt, 2012). It is standard practice to display outliers for numerical variables to understand possible skewness of data and decipher its quality. Boxplots provide a five-number summary like the boxplot() function, but in a graphical manner. The box plot for the numerical variables in the dataset (duration, amount and age) are shown below:


duration
60
40
20
0.6
0.8
1.0
1.2
1.4
x


amount
15000
10000
5000
0
0.6
0.8
1.0
1.2
1.4
x

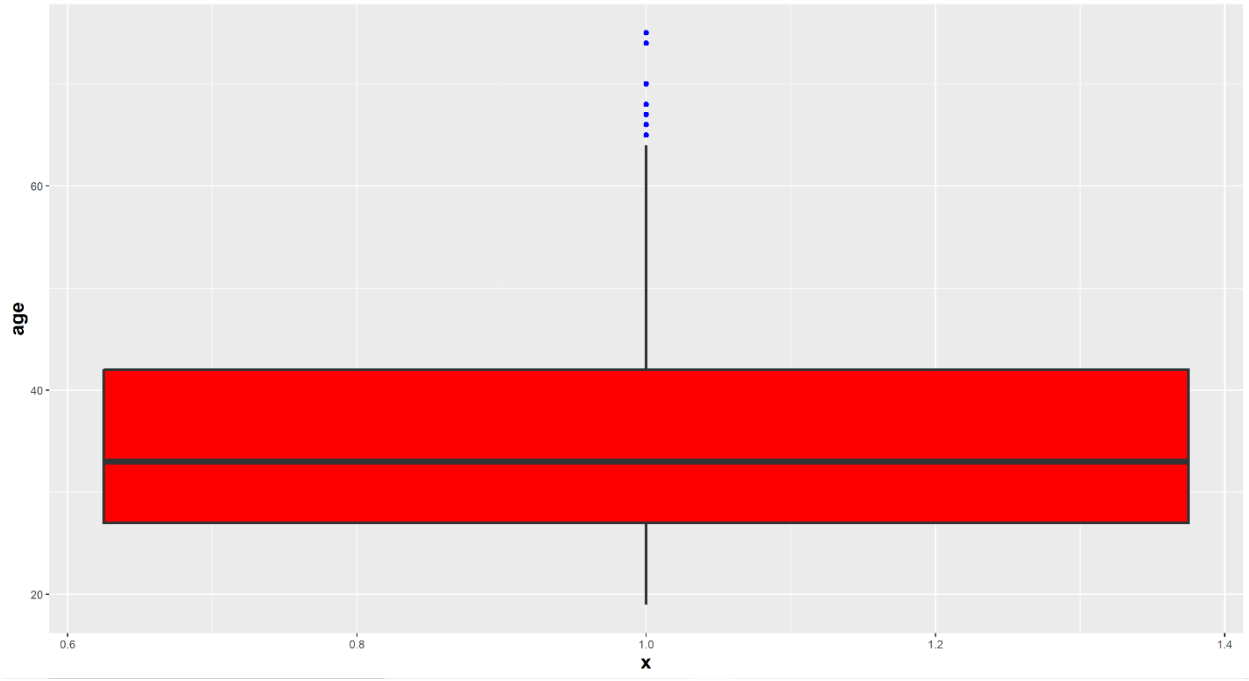


Figure 3: Boxplot of different columns of the dataset

The box plot shows that there are few outliers in all three feature variables. However, it does not make sense to remove these outliers logically. For example, outliers in "duration" column would be those with large numbers of Deutsche Mark currency on credit and it would especially be important for banks to assess the credit risk on these amounts. Likewise, age is a number which cannot be constricted and the range of values in duration of loans would also be essential in the analysis of credit risk. Hence, no outliers were removed.

2.4.2. Histogram of Categorical variables

A histogram is a graphical representation of the distribution of a dataset. It is a graph with a horizontal axis that represents the range of values in the dataset, and a vertical axis that represents the frequency of those values (Gonick & Smith, 1993). The shape of the histogram can give us a sense of the underlying distribution of the data, such as whether it is symmetrical or skewed. Histograms are commonly used in statistical analysis to visualize the distribution of a dataset and identify patterns or trends. They are also useful for comparing the distribution of multiple datasets, or for identifying outliers or other anomalies in the data (McGill, et al., 1978)

The histogram of the count and categorical columns were plotted below using the code

```
kde <- ggplot(gather(credit_data, cols, value), aes(x = value)) +
  geom_histogram(aes(y = ..density..), fill="red") +
  geom_density(col = "#06038D", size = 2) +
  facet_wrap(.~cols, scales = "free") +
  labs(x = "Feature Value", y = "Frequency", size = 20) +
  theme_dark() + theme(axis.title = element_text(size = 15, face =
"bold"), plot.title = element_text(face = "bold",hjust = 0.5))+
  ggtitle("Histograms of Features")

kde
```

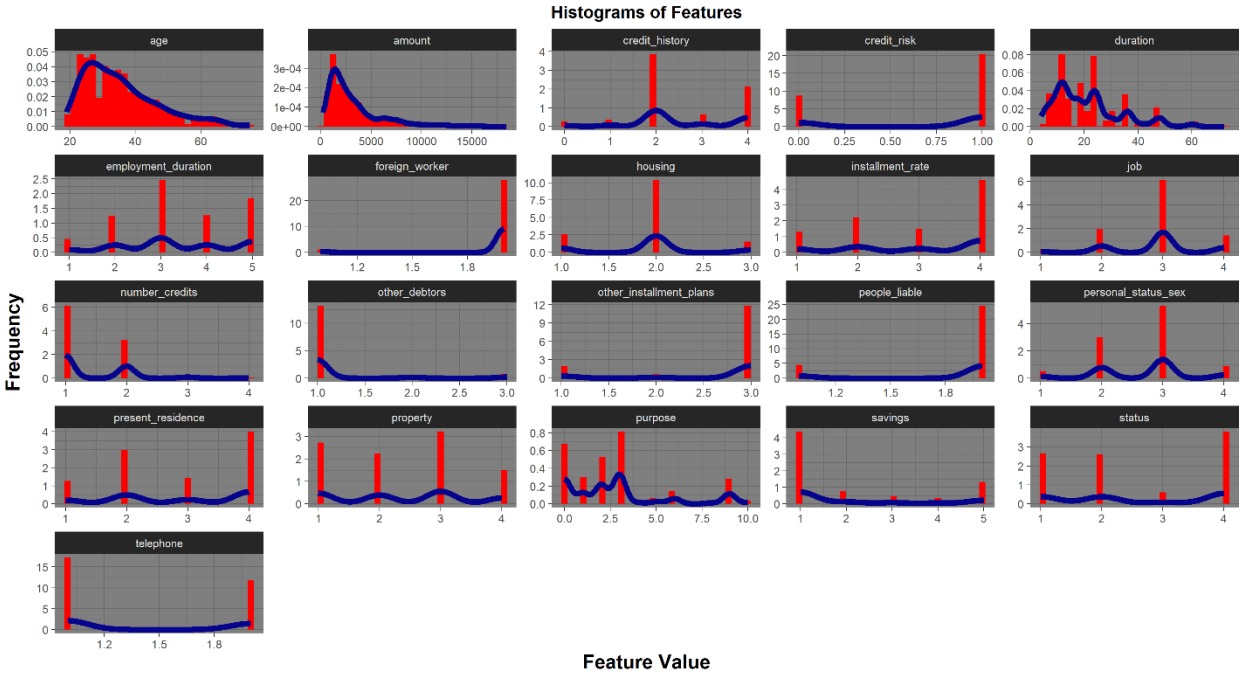


Figure 4: Histogram of count and categorical features.

Looking at the histogram of our numerical variables (age, amount and duration), it seems both age and amount have a positively skewed normal distribution, while duration could has a different distribution due to data points achieving high values intermittently.

On the other hand, for the binary variables (People liable, telephone and foreign worker), we can notice a Bernoulli distribution and we can identify which of the two variables are highest for that particular dataset. Example: for foreign workers, we see that most of the debtors in the given sample are local workers. This also raises important question about the data collection process:

Was this sample inclusive of all population and were there biases built in, or other confounding variables that led to this distribution of data?

Apart from the density plot of numerical variables which conveyed a great deal of information about the distribution, it is good practice to generate Q-Q plots of our numerical variable to check whether it is normally distributed. If the data is normally distributed, it will follow the straight line. Otherwise, it wouldn't adhere to the straight line.

```
gather(numerical[-4], condition, measurement) %>%
  ggplot(aes(sample = measurement)) +
  facet_wrap(~condition, scales = "free") +
  labs(title = "Q-Q plot of numerical variables", x = "Quantities of data", y
= "Quantiles") +
  stat_qq() +
  stat_qq_line()
```

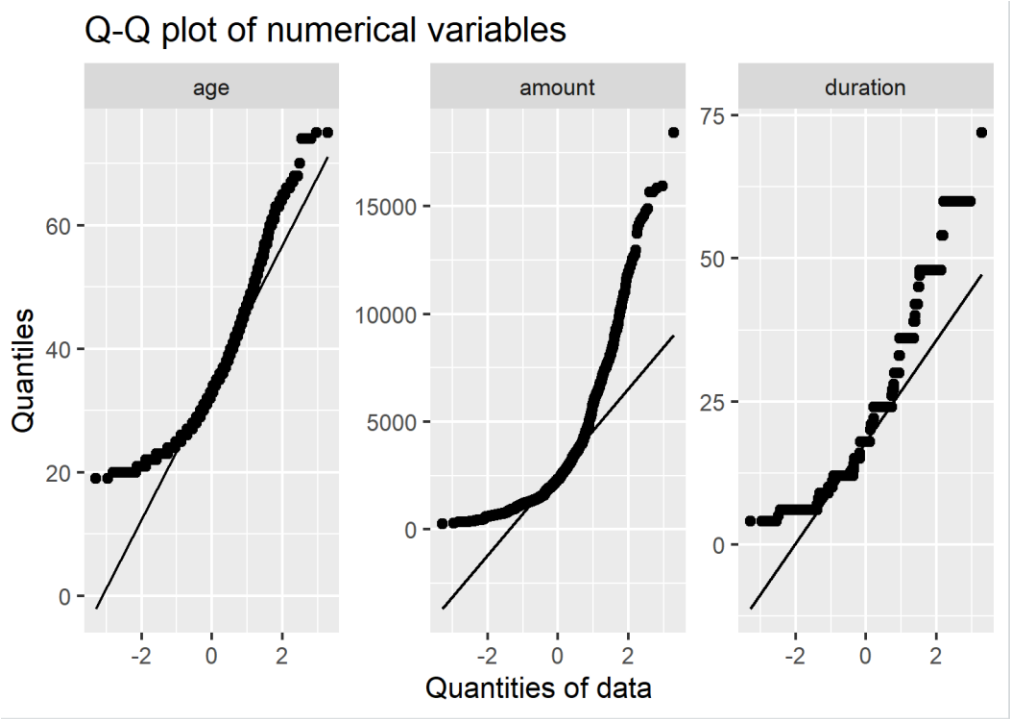


Figure 5: Q-Q Plot for the numerical variables

2.4.4 Finding Correlation between the numerical variables

Correlation is a statistical measure that describes the relationship between two variables. It is used to quantify the degree to which two variables are associated or related to each other (Agresti & Finlay, 2009). Correlation is often used to identify patterns or trends in data and to understand the

relationships between different variables (Bowerman, et al., 2018). Note: We cannot calculate these correlation coefficients for nominal variables (purpose, personal_status_sex) so those have been removed.

Firstly, the pearson correlation matrix will be generated as follows:

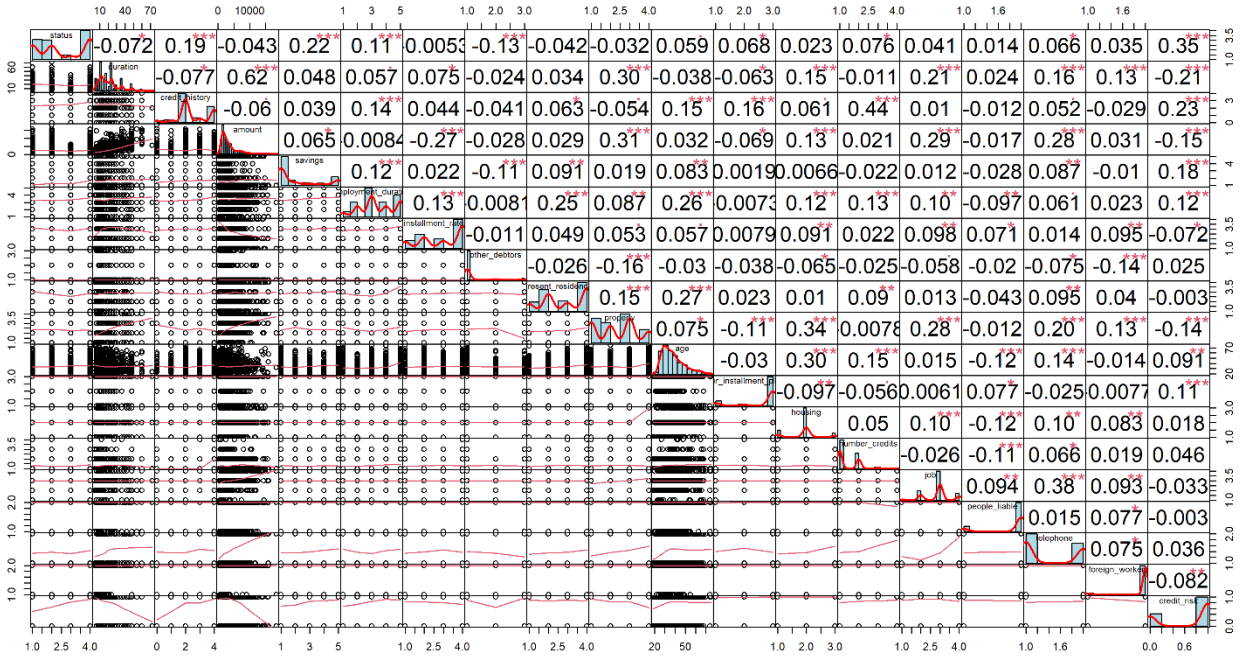


Figure 6: Pearson Correlation Matrix

Name and the feature are shown in the diagonal line of the plot. The top half of the plot shows the correlation coefficients. The bottom half shows the correlation graph between the features.

A rule of thumb here is that any value greater than 0.6 means strong positive correlation. Therefore, this matrix gives us an idea of the relationship between any two variables. Looking at the last column, we don't notice any strong positive or negative correlation of any variable with credit_risk. Besides this, a sanity check was performed against the predictions made by domain knowledge in Table 1. Moreover, we can see here that age has a very slight positive correlation (+0.09) with credit risk, meaning that greater the age, better the credit risk score.

This can be better visualized using a visual correlation matrix, marked with dots as shown below:

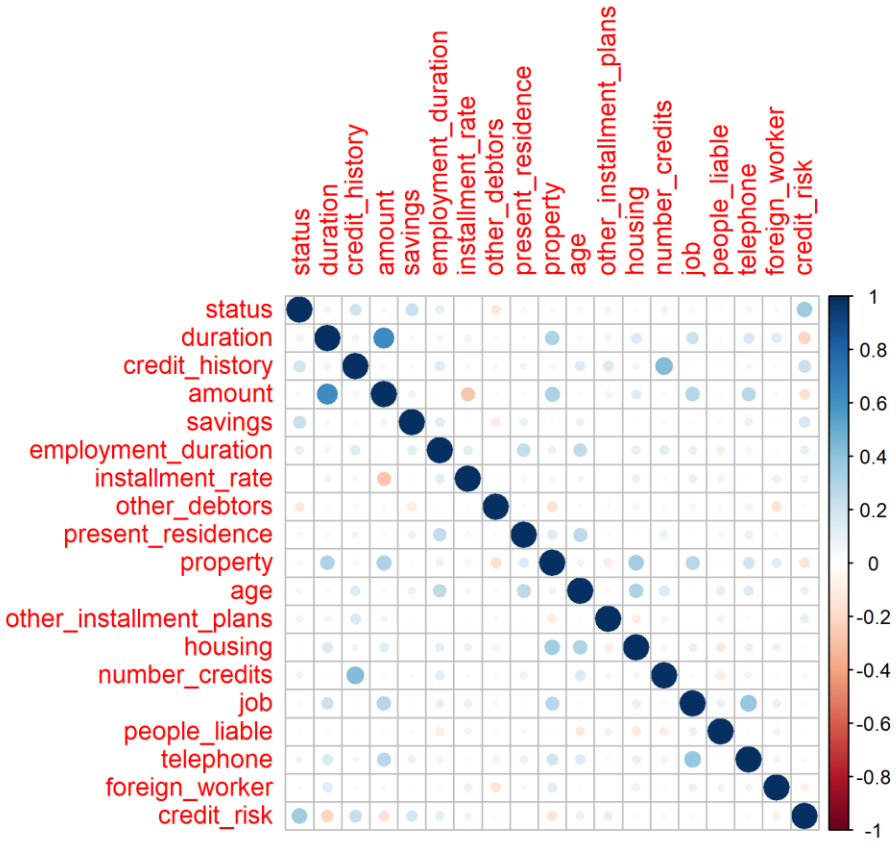


Figure 7: Visual Correlation Matrix (Pearson)

Another matrix was generated with data values obtained by the Spearman coefficient. This is because Spearman coefficient is better suited to categorical variables, which proliferates our dataset. Moreover, we do not know whether there is a monotonic relationship between our variables. Below is the spearman correlation matrix:

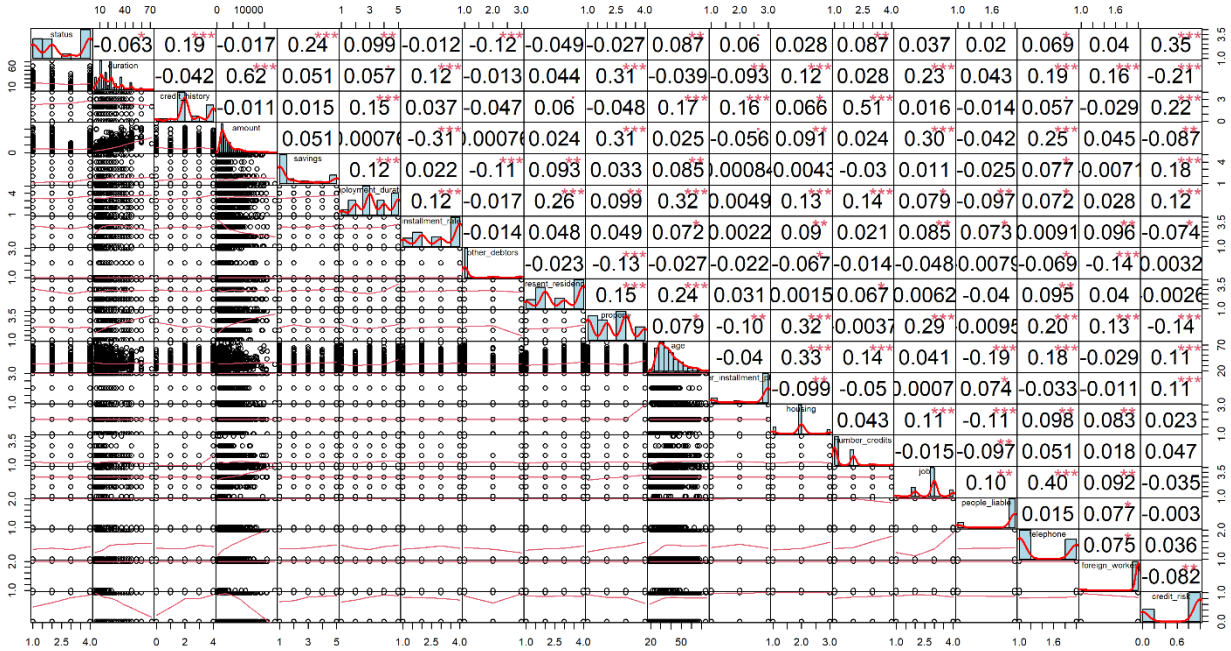


Figure 8: Spearman Correlation Matrix

We notice slight changes for some of the categorical variables and numerical variables. For example: the relationship between age with credit_risk has increased to 0.11 from the previous 0.091, with an increase in significance level as well. Below is the visual correlation matrix, marked with dots as shown below:

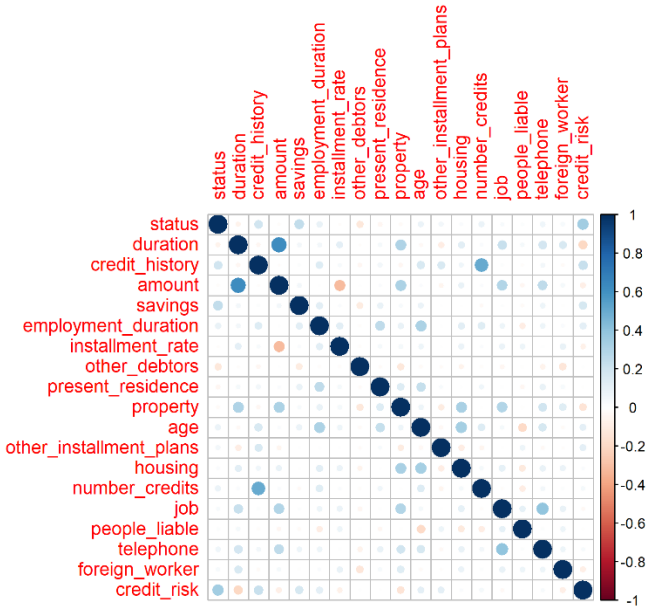


Figure 9: Visual Correlation Matrix (Spearman)

2.4.5 Analyzing categorical variables using contingency table

A contingency table is a statistical tool used to analyze the relationship between two or more categorical variables (Gelman & Hill, 2007) It is a tabular summary of the counts or proportions of the values of the variables in a dataset. Contingency tables are often used to investigate whether there is a significant association between the variables, and to identify patterns or trends in the data (Gelman & Hill, 2007).

Table 2: Contingency Table

	0	1	Sum
credit_history	650	1895	2545
purpose	871	1957	2828
savings	502	1603	2105
employment_duration	951	2433	3384
installment_rate	929	2044	2973
personal_status_sex	776	1906	2682
other_debtors	338	807	1145
present_residence	855	1990	2845
property	776	1582	2358
other_installment_plans	767	1908	2675
housing	574	1354	1928
number_credits	410	997	1407
job	881	2023	2904
people_liable	554	1291	1845
telephone	413	991	1404
foreign_worker	596	1367	1963
credit_risk	0	700	700
Sum	10843	26848	37691

The chi-squared test is a statistical test that is used to assess the independence of two categorical variables in a contingency table. It helps determine whether there is a significant association between the variables being studied.

```
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 366.93, df = 34, p-value < 2.2e-16
```

Because the p-value is so low, it would still pass our strict confidence interval. This suggests that there is a strong association between all these categorical variables.

2.4.6 Principal component analysis

Principal component analysis (PCA) is a statistical procedure that is used to reduce the dimensionality of a data set by projecting the data onto a lower-dimensional subspace (Jolliffe, 2002). This is done by identifying directions in the data (by plotting eigenvector) that have the highest variance and constructing new variables, called principal components. One benefit of PCA is its use in feature extraction and how it can increase the signal to noise ratio by removing unnecessary features in the data.

One way to visualize this is through a scree plot.

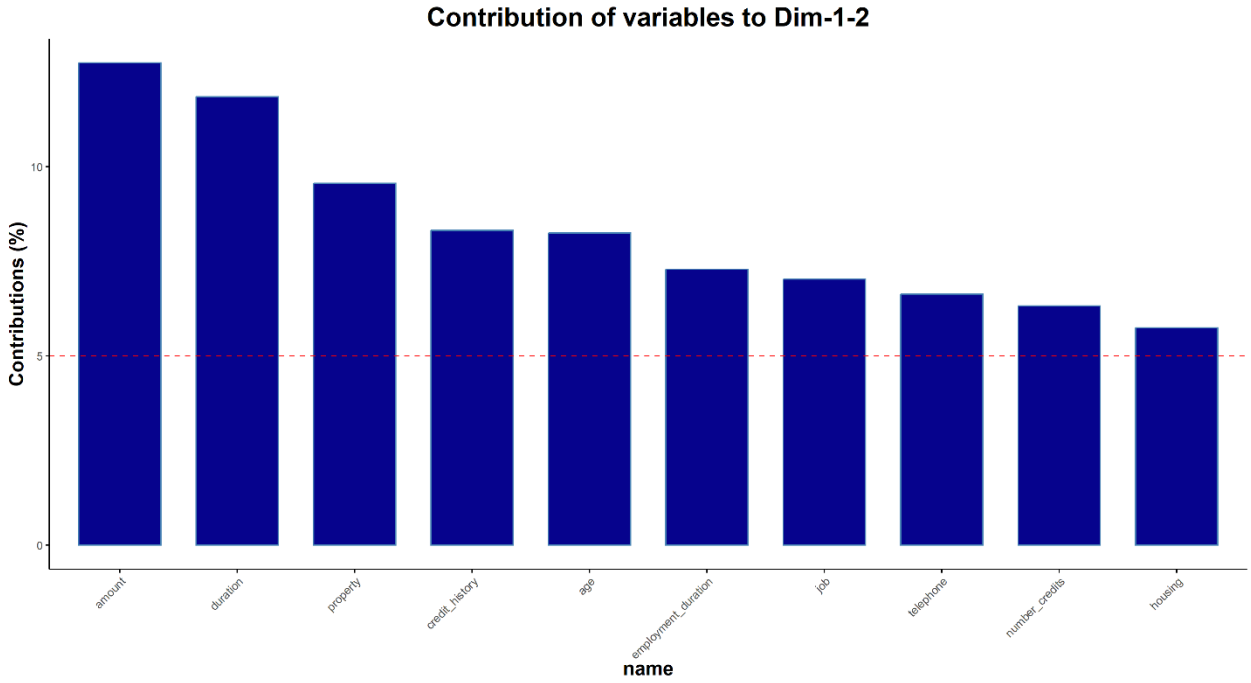


Figure 10: Scree Plot for PCA

It can be seen that around 13% of the variance can be explained by the first component, another 9% by the second component. This scree plot reveals that no one variable is extremely contributing to the variance. This can also be shown by calculating the expected average contribution and viewing that as the cut-off line (in red above). Here, we see that almost all the features contribute above the average in term of variance.

Lastly, we plot a PCA graph, color coded based on the most important feature. This will help in deciding which features to be removed:

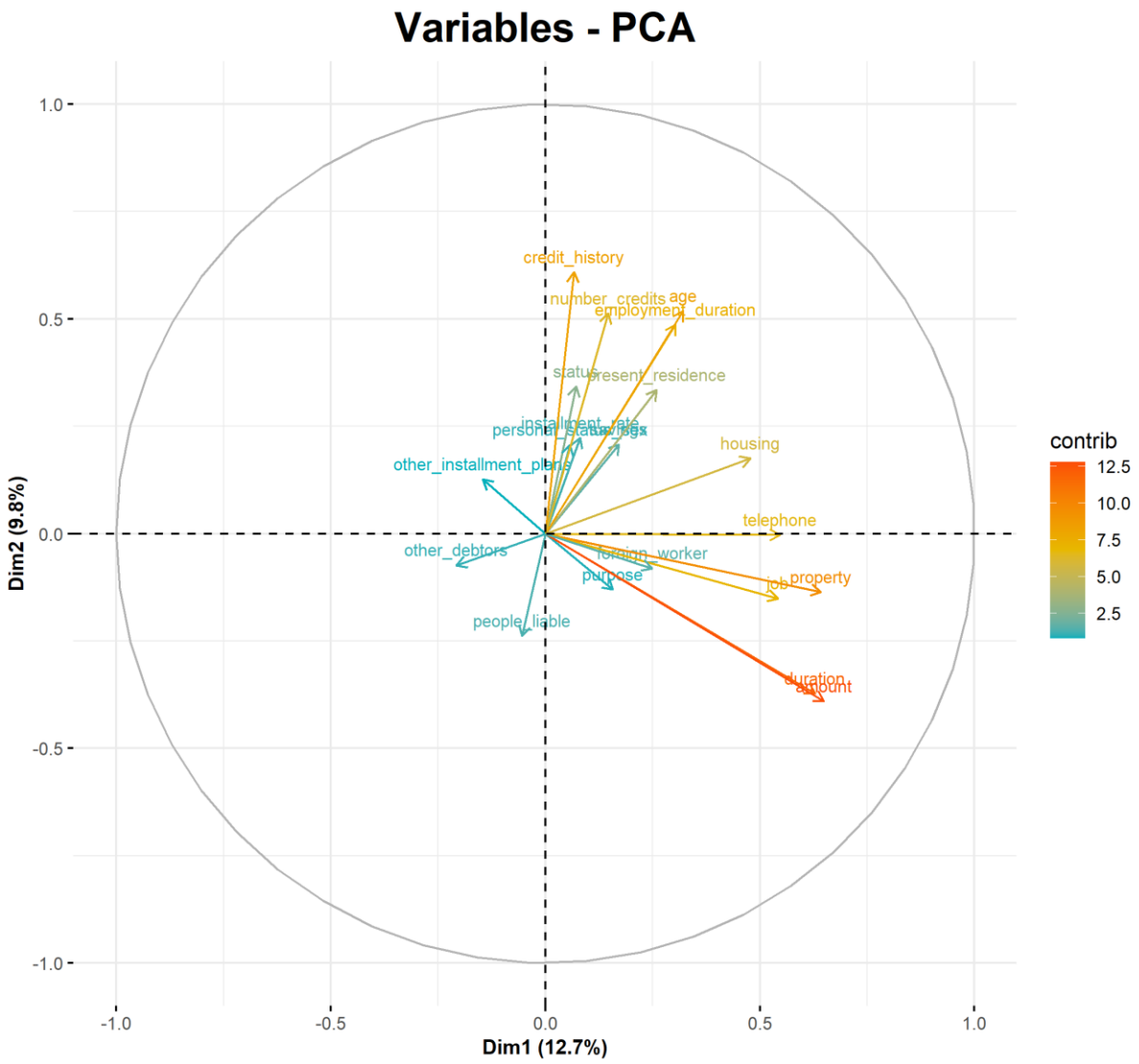


Figure 11: PCA Plot for German Credit Score Data

If we were to remove certain features, our PCA graph hints to remove the features, "other_debtors" and "people_liable", which surprisingly were considerably important features for detecting credit_score according to domain knowledge.

For further visualization, the data objects will be mapped the data objects onto a biplot as well, to observe their strengths and weaknesses in terms of the extracted components.

```
pca_biplot <- fviz_pca_ind(new_pca_credit, col.ind = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE) +
  theme(plot.title = element_text(face = "bold",hjust = 0.5, size =
20),axis.title.x = element_text(face="bold"), axis.title.y =
element_text(face="bold")) +
  theme(axis.title = element_text(size = 15))

pca_biplot
```

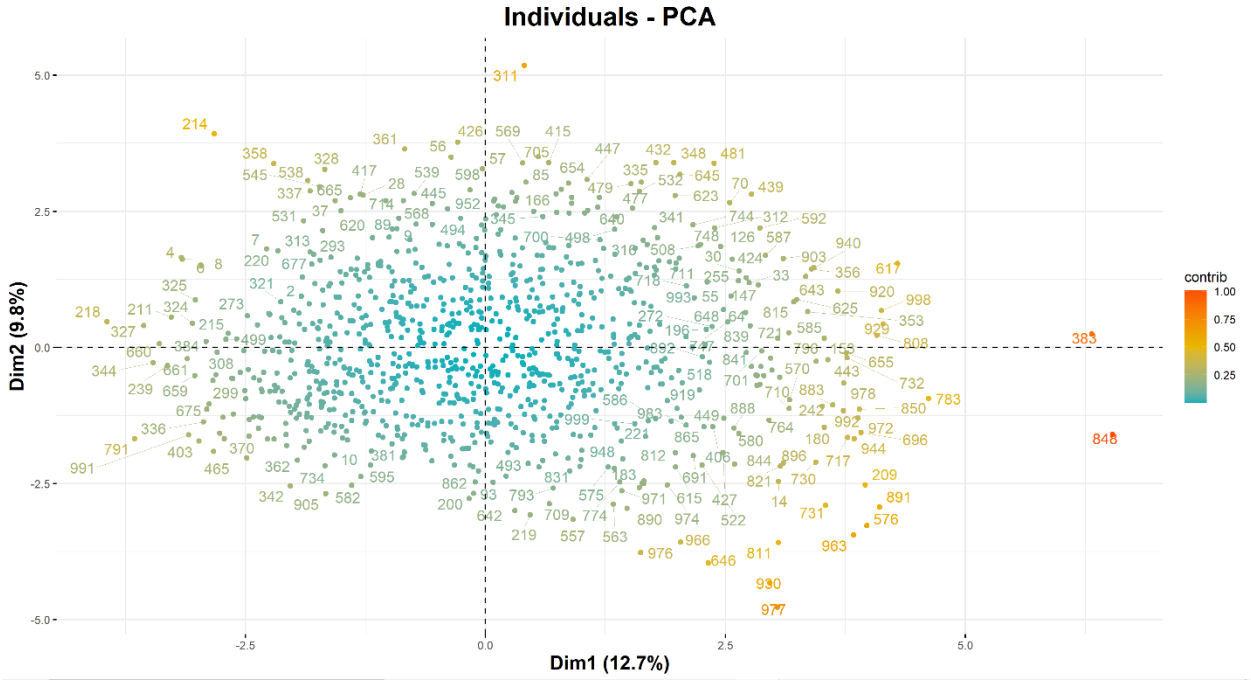


Figure 12: Individual PCA Plot

2.5 Random Forest Model

Random forest is a supervised learning algorithm used for both classification and regression tasks. It is an ensemble method that constructs multiple decision trees and combines their predictions to make a final prediction (Breiman, 2001).

We will be building the Conditional Random Forest model by using the cforest function from the party library. Training and test set is formed from the sample and is used for creating and testing the model.

Below is the code for creating this model:

```
#RANDOM FOREST MODEL
# Train the model using the training data
forest_credit <- cforest(credit_risk~., data = trainset, control =
cforest_unbiased(mtry = 10, ntree = 50))

# Predict the class labels for the test data
predictions <- predict(forest_credit, newdata = testdata, type = "response")

#Attaching probabilistic output
rf_pred <- ifelse(predictions > 0.5, 1, 0)

#Producing another confusion matrix
random_forest_cm <- table(predicted = rf_pred, actual = testset$credit_risk)
```

2.5.1 Evaluation of Random Forest Model

Evaluation of model can be done using the accuracy, precision, recall, FI Score and MCC. A summary table is created and the code snippet is as follows.

```
#Evaluation of random forest model

random_forest_TP <- random_forest_cm[1,1]
random_forest_FP <- random_forest_cm[1,2]
random_forest_TN <- random_forest_cm[2,2]
random_forest_FN <- random_forest_cm[2,1]

#Calculating accuracy of our model
random_forest_accuracy <- (random_forest_TP +
random_forest_TN)/(random_forest_TP + random_forest_FP + random_forest_TN +
random_forest_FN)

#Calculating precision of our model
random_forest_precision <- (random_forest_TP)/(random_forest_TP +
random_forest_FP)

#Calculating recall (sensitivity) of our model
random_forest_recall <- (random_forest_TP)/(random_forest_TP +
random_forest_FN)

#Calculating  of our model
random_forest_sensitivity <-
(random_forest_TN)/(random_forest_TN+random_forest_FP)

#Calculating F1 score of our model
```

```
random_forest_F1_score <- 2 * ((random_forest_precision *
random_forest_recall)/(random_forest_precision + random_forest_recall))

#Calculating Matthews Correlation Coefficient of our model
random_forest_MCC <- (random_forest_TP*random_forest_FN -
random_forest_FP*random_forest_FN)/(sqrt((random_forest_TP +
random_forest_FP)*(random_forest_TN+random_forest_FN)*(random_forest_TN+rando
m_forest_FP)*(random_forest_TN+random_forest_FN)))

#Summary Table of Evaluation
metrics <- c("Accuracy","Precision","Recall","Sensitivity","F1 Score","MCC")
values <-
c(random_forest_accuracy,random_forest_precision,random_forest_recall,random_
forest_sensitivity,random_forest_F1_score,random_forest_MCC)

df_results <- data.frame(metrics, values)
df_results
View(df_results)

#Finding which features are responsible for greatest amount of variance
ForestVarImp <- varimp(forest_credit)
barplot(ForestVarImp)
```

Actual result of confusion metric is shown below.

```
> random_forest_cm <- table(predicted = rf_pred, actual = testset$credit_risk)
> random_forest_cm
         actual
predicted   0   1
        0  25  11
        1  46 108
> ForestVarImp <- varimp(forest_credit)
> barplot(ForestVarImp)
```

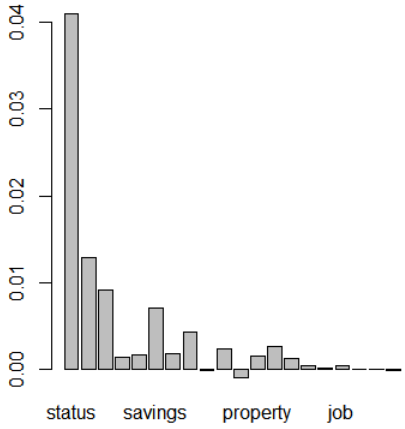


Figure 13: Barplot for important variables in Forest Tree Model

2.6 Support Vector Machine

Support vector machine (SVM) is a supervised learning algorithm used for classification and regression tasks. It is a powerful and versatile method that is based on the idea of finding a hyperplane in a high-dimensional space that maximally separates different classes (Cortes & Vapnik, 1995) (Fawcett, 2006). SVMs are effective in high-dimensional spaces and have a number of kernel functions that can be used to model non-linear relationships.

```
# Model SUPPORT VECTOR MACHINE MODEL
svm_trainset <- trainset
svm_trainset$credit_risk <- as.factor(svm_trainset$credit_risk)

#Type of kernels to use: 'linear', 'polynomial', 'radial basis', 'sigmoid')
svm_credit_risk <- svm(credit_risk~., data = svm_trainset, kernel =
"polynomial")

summary(svm_credit_risk)

#Fitting prediction model
svm_pred <- predict(svm_credit_risk, newdata = testdata, type = "response")


#Confusion matrix
svm_cm <- table(predicted = svm_pred, actual = testset$credit_risk)
svm_cm
```

Actual result of model building and confusion metric is shown below.

```
> # Model SUPPORT VECTOR MACHINE MODEL
> svm_trainset <- trainset
> svm_trainset$credit_risk <- as.factor(svm_trainset$credit_risk)
>
> #Type of kernels to use: 'linear', 'polynomial', 'radial basis', 'sigmoid')
> svm_credit_risk <- svm(credit_risk~., data = svm_trainset, kernel = "polynomial")
>
> summary(svm_credit_risk)

Call:
svm(formula = credit_risk ~ ., data = svm_trainset, kernel = "polynomial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  1
     degree:  3
     coef.0:  0

Number of Support Vectors:  510

 ( 297 213 )


Number of Classes:  2

Levels:
 0 1



>
> #Fitting prediction model
> svm_pred <- predict(svm_credit_risk, newdata = testdata, type = "response")
>
>
> #Confusion matrix
> svm_cm <- table(predicted = svm_pred, actual = testset$credit_risk)
> svm_cm
         actual
predicted   0   1
        0  15   3
        1  56 116
```

3. RESULTS

Performance evaluation is an important aspect of machine learning and data analysis. It allows us to assess the effectiveness and accuracy of different models and algorithms, and helps us choose the best approach for a given task (Fawcett, 2006). Performance evaluation also helps us identify areas for improvement and optimization, and allows us to compare different models and algorithms in a consistent and objective way (Berk, et al., 2018). Therefore, it is important to carefully evaluate the performance of different models and algorithms to make informed decisions about which ones to use for a given task.

We have plotted the summary of the various evaluation parameters for Random Forest Model and Support Vector Machine. Accuracy, Recall and FI Score of Random Forest Model is better than the Support Vector Machine.

In the current context of Credit Score, we can use Random Forest Model for prediction.

Table 3: Evaluation Parameters for Random Forest Model

	metrics	values
1	Accuracy	0.70000000
2	Precision	0.69444444
3	Recall	0.35211268
4	Sensitivity	0.90756303
5	F1 Score	0.46728972
6	MCC	0.06389111

Table 4: Evaluation Parameters for Support Vector Machine

	metrics	values
1	Accuracy	0.68947368
2	Precision	0.83333333
3	Recall	0.21126761
4	Sensitivity	0.97478992
5	F1 Score	0.33707865
6	MCC	0.08441723

4. CONCLUSION

In this investigation, we have performed an in-depth analysis of the German Credit Score Data. Problem definition of prediction of Credit_Risk was established. The data was thoroughly analysed using multiple techniques.

As the response variable is categorical, we have used Random Forest & Support Vector Machine. We have done an extensive calculation for 6 evaluation parameters.

Both the model – Random Forest and SVM has generalised it, wherein Random forest has better performance in Accuracy, Recall and FI Score.

The current research work was limited by the data and time. The work can further be expanded to have more data, multiple models and a comparative analysis of these models.

References

Agresti, A. F. C., 2013. *Statistics: The Art and Science of Learning from Data.* s.l.:Pearson Education.

Agresti, A. & Finlay, B., 2009. *Statistical Methods for the Social Sciences.* s.l.:Pearson Education.

Bansal, S., 2022. *What Is Data Science Process and Its Significance?.* [Online] Available at: https://www.analytixlabs.co.in/blog/data-science-process [Accessed 2023].

Berk, R. A., Brown, L. & Zhao, Y., 2018. *Performance evaluation and optimization for machine learning models.* s.l.:Cambridge University Press.

Bowerman, B. L., O'Connell, R. T. & Murrell, P., 2018. *Applied Statistics for the Six Sigma Green Belt.* 2nd ed. s.l.:ASQ Quality Press.

Breiman, L., 2001. Random Forests. *Machine Learning,* 45(1), pp. 5-32.

Cortes, C. & Vapnik, V., 1995. Support-vector networks. *Machine learning,* 20(3), pp. 273-297.

Dhar, V., 2017. *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking.* s.l.:O'Reilly Media.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern recognition letters,* 27(8), pp. 861-874.

Gelman, A. & Hill, J., 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* s.l.:Cambridge University Press.

Ghosh, D. & Vogt, A., 2012. *Outliers: An evaluation of methodologies.* s.l., Joint statistical meetings.

Gonick, L. & Smith, W., 1993. *The Cartoon Guide to Statistics.* s.l.:HarperCollins Publishers.

Hand, D. J., 2013. *The Palgrave Handbook of Economics and Complexity.* s.l.:Palgrave Macmillan.

Manyika, J. et al., 2011. *Big data: The next frontier for innovation, competition, and productivity,* s.l.: McKinsey & Company.

McGill, R., Tukey, J. W. & Larsen, W. A., 1978. Variations of Box Plots. *The American Statistician,* 32(1), pp. 12-16.

Pallant, J., 2011. *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS.* s.l.:Allen & Unwin.

Siddharthan, A., 2018. *Data Science for Healthcare: An Introduction to Key Methods, Challenges, and Opportunities.*. s.l.:The MIT Press.

Witten, I. H., Frank, E. & Hall, M. A., 2011. *Data Mining: Practical Machine Learning Tools and Techniques.* s.l.:Morgan Kaufmann.