CREDIT RISK: SOLVING A CLASSIFICATION BUSINESS PROBLEM USING SUPERVISED MACHINE LEARNING

ML1000: Assignment 1

Group: ML_pros

Julia Mitroi

Durai Nachiappan

Shabeeth Syed

Lingling Zhang

Credit Risk: Solving a Classification Business Problem

Using Supervised Machine Learning

1. Business understanding

1.1. Business objectives

1.1.1. Background

Much of the economy can be described by a trade-off between risk and return. There may be risks associated with all actions, with "risk" being defined as the chance that an investment, time, effort, or money will go wasted instead of used productively. The broad topic of this project is *bank investments* (with *loan* being a type of investment), and the two areas that are typically analyzed for investments are the risks associated with the investment and the potential return of that investment. The purpose of taking on risk in the first place is the chance for a greater return, thus when financial institutions give loans, they are undertaking a risk – the risk that the borrower will pay the loan back (credit risk) – in the hope of making a return.

A *credit risk* is therefore the risk of default on a debt that may result if a borrower fails to make required payments. To avoid financial loss, a bank has to evaluate the risk associated with lending money and not being fully repaid; when it receives a loan application, it has to make a decision on whether to approve the loan or not. It does so based on the applicant's profile, and it needs to use a decision support system (rule, algorithm, or model) regarding who to grant approval of the loan and who not to.

Data mining, which is the analysis step of pattern recognition and knowledge discovery in databases, is a machine learning method commonly used for this purpose. Machine learning/data mining applications based on *classification* are applied in the financial and banking areas for better decision making, including for credit risk assessment, loan approval, and fraud detection.

1.1.2. Business Objectives

The business objective that motivated this project was risk minimization and profit maximization on behalf of a bank, in the context of decision making for loan applications. Towards this business objective, this project proposes a useful solution for assessing loan

applicants, namely a machine learning classification model, that could be used by a bank to make good, business worthy decisions and mitigate financial loss.

1.1.3. Business Success Criteria

The statistical decisions made by the above-noted classification model will be translated into profit consideration for the bank. From a business point of view, our model will have a successful outcome if by using it, the bank makes accurate decisions that result in a given, measurable profit (e.g., 40%) within a given period of time, such as five or ten years.

1.2. Situation assessment

1.2.1. Inventory of Resources

Resources available to our project are as follows:

- Personnel: As this was a group project, by working on the project all members of our group assumed the roles of business, data, and data mining experts
- Data and data sources: The German credit dataset, which we downloaded from the UCI repository; access to the UCI repository
- Computing resources: Windows and Mac platforms, on which we worked on the project
- Software: R, R Studio, Shiny, Microsoft Word and Excel
- Knowledge sources: The ML1000 course material, which includes written and online documentation

1.2.2. Requirements, Assumptions, and Constraints

Key requirements of this project are:

- The project has to be completed by October 12, 2018
- The project outcomes and report have to meet or exceed the standards for ML1000, and to be deemed usable by a bank for the assessment of credit risk

Assumptions:

- The data in the German credit dataset are from a normal distribution
- The German credit dataset contains data that are amenable to applying a data mining classification algorithm
- The (hypothetical) bank for which we are developing the model has the resources to be able to use it
- As we (project members) have not directly collected the project data, we are assuming

that the data in the German credit dataset have adequate coverage of loan applicant population, applicable to any banking context

Constraints:

- The group members who worked on this project are learners of machine learning/data mining, not (yet) experts
- The time allotted to the project, although sufficient it is somewhat limited

1.2.3. Risks and Contingencies

There are no known risks associated with this class project. However, had the project been undertaken in a corporate context, there could have been the risk of not obtaining additional funding depending on the initial data mining results; and the risk of a competitor organization or consulting firms coming up with a similar/better model before we developed ours.

1.2.4. Terminology

A glossary of terminology relevant to this project is included in Appendix A.

1.2.5. Costs and Benefits

In this business context, if the bank predicts that an application is good or credit worthy and it turns out to be credit worthy, this would result in a *benefit*, or profit, for the bank. In terms of *costs*, it would be worse (costlier) to classify an applicant as good when they are bad (non-creditworthy) as it would result in financial loss for the bank, than to classify an applicant as bad when they are good; the latter would result in a loss of profit, which would be less severe for a bank than loss of money lent to a customer.

1.3. Determine data mining goals

1.3.1. Data Mining Goals

For this project, our group converted the business problem/objective noted in 1.1.2. above into a data mining (analytical) problem, with the intended analysis objective being to develop a classification algorithm for credit risk; this algorithm or model will determine with accuracy if an application is creditworthy or not, and a bank will be able to use it to make decisions on loan approvals, for risk minimization and profit maximization. In other words, the output of our analysis, or data mining process – the classification model – will enable the achievement of the

business objectives. As part of the data mining process, following the development of the classification model its performance (accuracy) will also be evaluated.

1.3.2. Data Mining Success Criteria

A correct decision would be one where the bank predicts (through our data mining model) that a customer application is good and it turns out to be credit worthy; as well as when the bank accurately predicts that an application is non-creditworthy and does not extend a loan to that applicant, avoiding incurring a financial loss.

As part of the current project, a tool used to assess cost-profit, and therefore classification model usefulness and accuracy, is a *cost matrix*. A classification cost matrix (error matrix) is useful when specific classification errors can be more severe than others; it calculates the cost of wrong prediction or right prediction, with a provision that some mis-predictions can be very costly. As noted above, it would be worse to classify an applicant as good when they are bad, than to classify an applicant as bad when they are good.

In a cost matrix, error weights are assigned to misclassifications, using class labels such as 'high risk', 'low risk' and 'safe'; weights specified must be greater than or equal to zero. The classification mining model attempts to avoid classification errors with a high error weight, and it will be considered to be successful if it did.

1.4. Project plan

1.4.1. Project Plan

This data mining goal noted above in 1.3.1. is to be achieved by this project by using a) a publicly available German credit dataset, b) R programming, and c) CRISP-DM methodology.

The broad steps of the data mining project plan are:

- Identifying the data source (described below, under Data Collection).
- Selecting the data points that need to be analyzed; or the entire dataset, if applicable.
- Extracting the relevant information from the data for the classification model.
- Identifying the best fitting classification model.
- Interpreting and reporting the results.

1.4.2 Initial Assessment of Tools and Techniques

The goal of this project is to develop a *classification model*, which itself is a tool or technique that can be defined and assessed. Classification models find a rule or set of rules to represent data into classes. As noted, financial institutions require rule(s) for making decisions, to classify customers into good or bad credit risks; and based on these decisions loan can be given to specific customers. As such, a classification model was assessed to be suitable for the business goals of this project.

Specifically, the classification model used in this project is *random forest*, which is a supervised learning algorithm that creates a forest and makes it random. The "forest" is a combination of decision trees (defined below), most of the time trained by a combination of learning models that enhances the overall result. In other words, random forest builds a number of decision trees and merges them together to get a more accurate and stable prediction. A *decision tree* (component of random forest) works as follows: Each internal node represents a "test" on an attribute/variable (e.g. whether an applicant is a foreign worker or not), each branch depicts the outcome of the test, and each leaf node is a class label (decision taken after computing all attributes). A node that has no children is a leaf. By analyzing the feature (attribute) importance, it can be decided which features to drop because they don't contribute enough or nothing to the prediction process. A general rule in machine learning is that the more features there are, the more likely the model will suffer from overfitting; therefore dropping features that don't make a significant contribution is important.

Another tool used in this project is the *CRISP-DM* methodology. CRISP-DM is a robust methodology that provides a structured approach to planning a data mining project, and is considered to be the "gold-standard" for data mining projects. (CRISP-DM stands for "CRoss-Industry Process for Data Mining".) It is described as a hierarchical process model, with tasks at four levels of abstraction, from general to specific: phase, generic, specialized, and process instance.

2. Data understanding

2.1. Initial data collection

2.1.1. Initial Data Collection Report

The dataset used for this project is a public benchmark from the UCI Machine Learning Repository, namely a credit data file donated by Professor Dr. Hans Hofmann from the University of Hamburg to the UCI repository. We collected (accessed and downloaded) the data from the UCI's FTP website, at
http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

2.2. Data description

2.2.1. Data Description Report

This dataset classifies loan applicants described by a set of attributes as good or bad credit risks. It has 1000 observations and 21 attributes (columns) – 7 continuous and 13 categorical; and the response or target variable, 'Creditability'. There are no missing values in the data. For the purposes of training and testing, 60% of the overall data were used for training and 40% for testing the accuracy of the classification of our classification algorithm.

2.3. Data exploration

2.3.1. Data Exploration Report

Before starting the analysis, we performed Exploratory Data Analysis (EDA) and data pre-processing. The UCI repository had two datasets for German credit, a raw dataset that had string value codes for the categorical variables (e.g., A40, A58), and a dataset prepared by Strathclyde University that had these codes replaced by numeric scores. For example, for the 'Account Balance' variable, with its four levels, "no running account" (A14) was replaced by a score of 1, "no balance or debit" (A14) a score of 2, "0 <= … < 200 DM" (A12) a score of 3, and "…>=200 DM" (A13) a score of 4. (DM stands for Deutsche Mark, the German currency at the time that these data were collected). For data pre-processing, the dataset with the numeric scores was further processed in Excel to obtain one variable per column; to assign column headings; and to replace the outliers by imputing the sample median.

Once we had a clean Excel (.csv) dataset, we performed exploratory data analysis using R. As

both categorical and continuous/numeric variables are part of the dataset, summary statistics were computed for both, as well as for the target variable, and the distributions charted.

Target Variable

The target (or response) variable for this dataset is 'Creditability', which is binary/categorical. 'Creditability' is the indicator variable on loan defaults, and is used as the target in the supervised machine learning we performed to solve the business problem for this project. Proportions for Creditability's two levels, Creditable (1) and Non-creditable (0), were calculated with R's *prop.table* function as being 70% and 30% respectively. The distribution for 'Creditability' is displayed using a bar plot through R's *barplot* function, as shown in Figure 1.

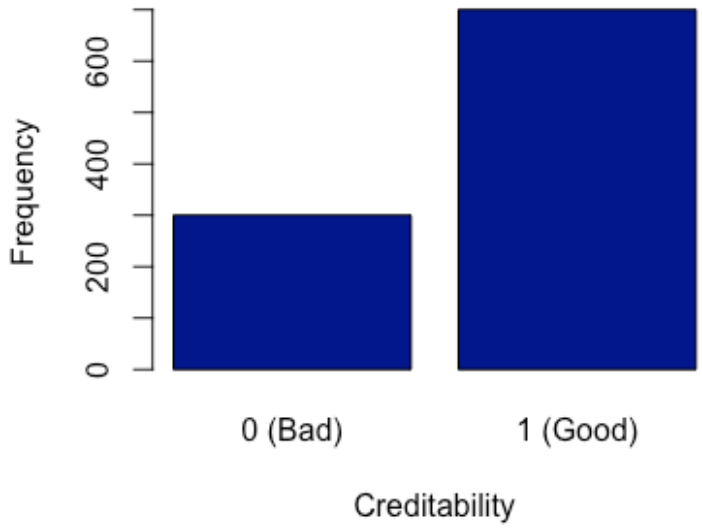

Figure 1. Bar plot depicting the distribution of 'Creditability'

Continuous Variables

Distribution of the Continuous Variables

As a first step to examine the distribution of the numeric variables, histograms and boxplots were used. The purpose of a *histogram* is to graphically summarize and display the distribution of a univariate data set; it shows whether or not the distribution is symmetrical/normal. A symmetric (bell-shaped) distribution is one in which the two halves of the histogram show as

mirror-images of one another; a skewed (non-symmetric) distribution is a distribution in which there is no such mirror-imaging.

A distribution for a variable is symmetrical about the mean and has a bell-shaped histogram if the mean equals the median (or, when the mean and median are very close it is sometimes practical to treat the distribution as symmetric). A distribution is positively skewed, with the histogram's tail on the right side being longer, if the mean is greater than the median; and is negatively skewed, with the histogram's tail on the left side being longer, if the mean is less than the median. How much the data is skewed depends on how far the mean and median differ.

A *boxplot* visually displays the data distribution through their quartiles. The lines extending from the boxes are known as the "whiskers", and are used to indicate variability outside the upper and lower quartiles. Outliers are plotted as individual dots that are in-line with the whiskers. In a boxplot, if the data are roughly the same on each side when cut down the middle by the median (which demarcates the center of the distribution in box plots), the distribution is symmetric. If there is a longer part of the box above or below the median, the data are skewed.

Duration of Credit (Month)

As the histogram in Figure 2 shows, for the 'Duration of Credit (Month)' variable, most of the applicants' credits were between 10 and 20 months. The second most frequently observed credit durations were between 20 and 30 months. The third most frequent range of credit durations was between 0 and 10 months, and the fourth most frequent range of durations was between 30 and 40. Few applicant credit durations were between 40 and 50 months, and the lowest recorded credit durations were between 50 and 60, 60 and 70, and 70 and 80 months. Histogram bin size is 10 for this variable.

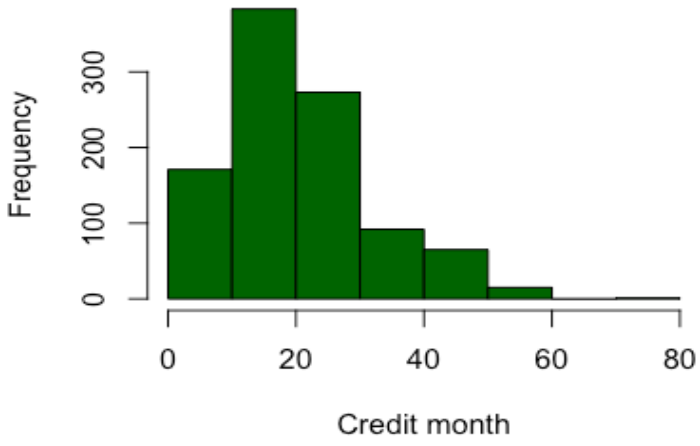

Figure 2. Histogram depicting the distribution of 'Duration of Credit (Months)'

The boxplot for 'Duration of Credit (Month)' (Figure 3) depicts this variable's distribution through the quartiles. The red dot and thin line on the boxplot indicate the sample mean.

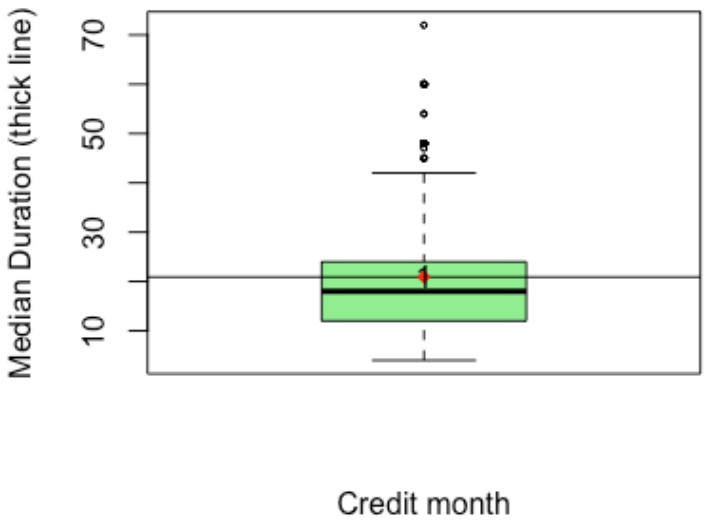

Figure 3. Box plot depicting the distribution of 'Duration of Credit (Months)'

Both the histogram and boxplot for Duration of Credit show that the data are skewed to the right. If a histogram is skewed, the median (second quartile) is a better estimate of the "center" of the distribution than the sample mean. Additionally, if the histogram/boxplot indicates a skewed data set, a recommended next step is to quantitatively summarize the data by computing and reporting the mean, the median (Q2), and sample mode. These numeric measures, for each of the three continuous variables, are included in a summary table below.

Amount of credit (DM)

The histogram in Figure 4 depicts that for the 'Amount of credit (DM)' variable, most of the applicants' credit amount was between 1000 and 2000 DM. (Histogram bin size is 1000 for this variable.) The second most frequently observed credit amount was between 2000 and 3000 DM. The third most frequent range of credit amount was between 3000 and 4000 DM, and the fourth most frequent range of prices was between 0 and 1000.

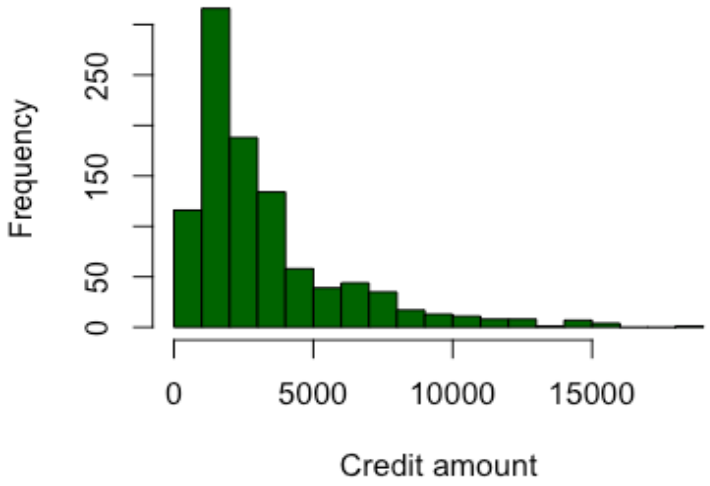

Figure 4. Histogram depicting the distribution of 'Amount of Credit (DM)'

The boxplot for 'Amount of credit (DM)' (Figure 5) depicts this variable's distribution through the quartiles. The red dot and thin line on the boxplot indicate the sample mean.

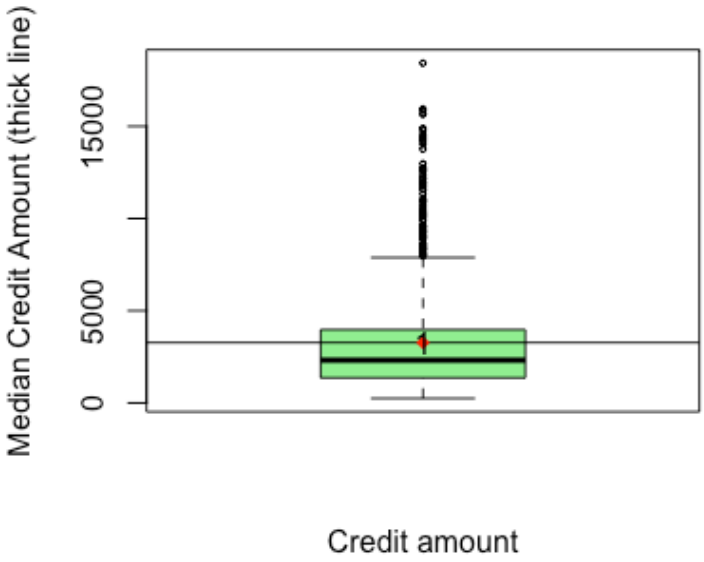

Figure 5. Box plot depicting the distribution of 'Amount of Credit (DM)'

Age (of Applicant)

The histogram in Figure 6 depicts that for the 'Age (of Applicant)' variable, most of the applicant ages were between 20 and 30 years; the age was in this range for 393 of the applicants in this sample. The second most frequently observed ages were between 30 and 40; the age was in this range for 319 applicants. The third most frequent age range was between 40 and 50, and the fourth most frequent age range was between 50 and 60. Few applicant ages were between 60 and 70, and the lowest recorded ages were between 10 and 20, and 70 and 80. Histogram bin size is 10 for this variable.

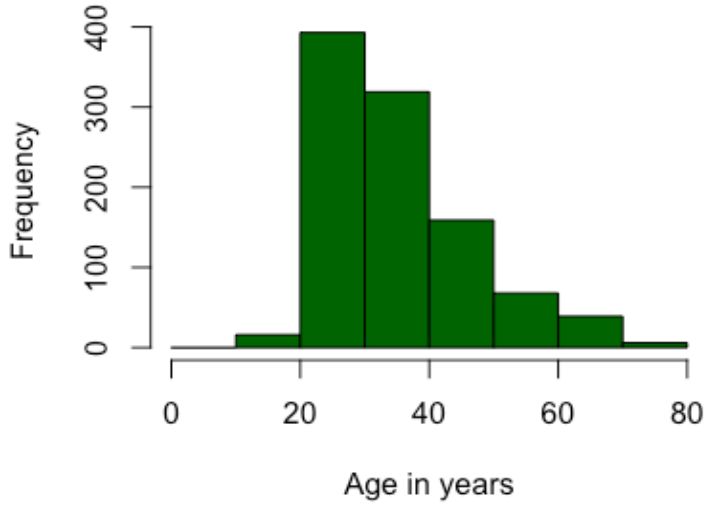

Figure 6. Histogram depicting the distribution of 'Age (of Applicant)'

The boxplot for 'Age (of Applicant)' (Figure 7) depicts this variable's distribution through the quartiles. The red dot and thin line on the boxplot indicate the sample mean.

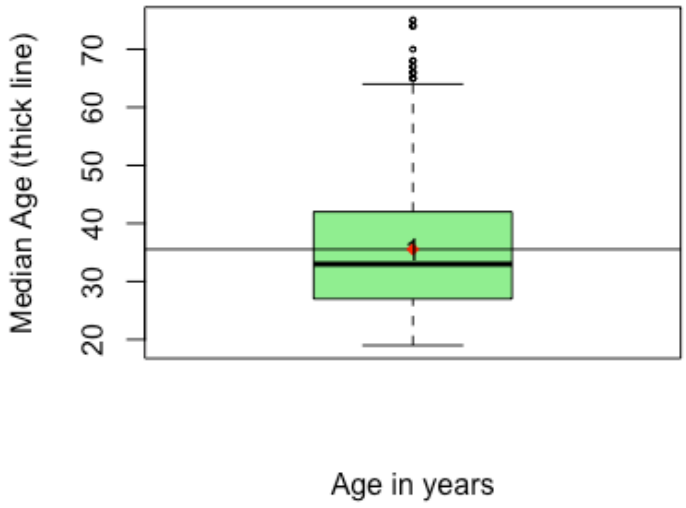

Figure 7. Box plot depicting the distribution of 'Age (of Applicant)'

The histograms and boxplots for all three continuous variables show positive skewness, that is, the distribution is not symmetrical (where the data points would be roughly equally balanced around the mean). For positively skewed distributions, as in the German credit data, the mean is greater than the median; the tail of the distribution on the right hand (positive) side is longer than on the left-hand side (as shown in the histograms); and the median is closer to the first quartile than the third quartile (seen in the boxplots).

Interpreting the positive data skew further, assuming that our credit data is from a normal distribution, the positive skewness for:

- 'Age (of Applicant)' means that the average age of the applicants in this dataset is higher than the median age of the applicants. In other words, more than half of the applicants have a lower age than the average age. There are more younger applicants than older applicants.
- 'Amount of Credit (DM)' means that the average credit amount of the applicants in this dataset is higher than the median credit amount of the applicants. Or, more than half of the applicants have a lower credit amount than the average credit amount. There are more applicants with lower credit amount than applicants with higher credit amount.
- 'Duration of Credit (Month)' means that the average duration (in months) of the applicants' credit is higher than the median duration of credit. In other words, more than half of the applicants have had credit for a shorter amount of time than the average duration of credit. There are more applicants with shorter credit duration than applicants with a higher number of months of credit period.

A *summary* of the continuous variables, based on running the R program for data exploration, is depicted in the table below.

Predictor	**Min**	**Q1**	**Median**	**Q3**	**Max**	**Mean**	**SD**
Duration of Credit	4	12	18	24	72	20.9	12.06
Credit Amount	250	1366	2320	3972	18420	3271	2822.75
Age (of Applicant)	19	27	33	42	75	35.54	11.35

Categorical variables

Proportions of applicants belonging to each level of a categorical variable were graphed in a grid format through R's ggplot funtion (Figure 8).

Figure 8. Matrix of bar plots depicting the distribution of the dataset's Categorical variables

The proportions were also computed using R's *dplyr* library. For example, for 'Account Balance', the proportions were:

Predictor	**Levels and Proportions**			
Account Balance	No Account	None	Below 200 DM	200 DM or Above
%	27.4%	26.9%	6.3%	39.4%

As most of the categorical variables in this dataset have several levels, the full cross-classification of all variables would result in no observations in many cells. As such, only a subset of predictor variables was considered, that may have an impact on the response variable, Creditability, based on Chi-square and t-tests p-values; explained below.

Two-dimensional (K1 x K2) contingency tables were calculated using R's *CrossTable* function. A *contingency table* is a special type of frequency distribution table, that summarizes the relationship between categorical variables, where two variables are shown simultaneously.

For example, for 'Account Balance', the cross-tabulation with 'Creditability' is depicted as below.

```
> CrossTable(Creditability, Account.Balance, digits=1, prop.r=F, prop.t=F, prop.chisq=F, chisq=T)


   Cell Contents
|-------------------------|
|                       N |
|           N / Col Total |
|-------------------------|


Total Observations in Table:  1000


             | Account.Balance
Creditability |         1 |         2 |         3 |         4 | Row Total |
--------------|-----------|-----------|-----------|-----------|-----------|
            0 |       135 |       105 |        14 |        46 |       300 |
              |       0.5 |       0.4 |       0.2 |       0.1 |           |
--------------|-----------|-----------|-----------|-----------|-----------|
            1 |       139 |       164 |        49 |       348 |       700 |
              |       0.5 |       0.6 |       0.8 |       0.9 |           |
--------------|-----------|-----------|-----------|-----------|-----------|
 Column Total |       274 |       269 |        63 |       394 |      1000 |
              |       0.3 |       0.3 |       0.1 |       0.4 |           |
--------------|-----------|-----------|-----------|-----------|-----------|
```

The levels of 'Account Balance are:
"no running account" = 1
"no balance or debit" = 2
"0 <= … < 200 DM" = 3
"…>=200 DM" = 4

'Creditability' has two levels, 1 and 0, or creditable and non-creditable.

The proportions displayed in the cross tabulation of 'Account Balance' and 'Creditability' show for example, in the column proportions, that 30% of 1000 applicants have no account, and another 30% have no balance; while about 40% have some balance, either less than or greater than 200 DM. From those who have no account, 135 (50%) are not credit worthy and 139 (50%) are found to be Creditable. Also, in the group with no account balance, 40% were found to be non-Creditable.

This example described features of the contingency table for 'Account Balance' with Creditability. We created two-variable relationship tables for all of the dataset's variables (as shown in the enclosed R program for the exploratory data analysis) – using cross tabulation for the categorical variables, and Welch Two Sample t-tests for the continuous variables – and when running the *CrossTable* and *t.test* functions in R, they also output the *chi-square p-values* and *t-test p-values* for each contingency table and continuous-target variable t-test, depicting the significance of a predictor variable on 'Creditability'.

For example, for Account Balance:

```
Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  123.7209     d.f. =  3     p =  1.218902e-26
```

Only significant predictors were included in the development of our classification model. The following table summarizes the chi-square p-values for each contingency table we created, with the significant predictors depicted in red font.

Predictor variable	**Chi-square P-value**
Account Balance	< 0.001
Payment Status	< 0.001
Purpose	< 0.001
Savings/Stock Value	< 0.001

Length of Employment	< 0.001
Installment %	0.14
Sex and Marital Status	0.01
Duration in Current Address	0.86
Type of Apartment	< 0.001
Most Valuable Asset	< 0.001
Number of Credits at Bank	0.15
Guarantor	0.98
Occupation	0.42
Concurrent Credits	< 0.001
Number of Dependents	0.92
Telephone	0.28

The following table summarizes the t-tests results, including the p-values for each relationship between a continuous variable and the target variable, with the significant predictors depicted in red font.

Predictors	Mean (Creditworthy Group)	Mean (Non-Creditworthy Group)	P-value (T-test)
Duration of Credit	19.0	24.9	< 0.001
Amount of Credit	3928.1	2985.4	< 0.001
Age	33.9	36.2	0.003

2.4. Verify data quality

2.4.1. Data Quality Report

There are no missing values in the German credit dataset, verified using R's *is.na* function. An assumption is being made that the data are coming from a normal distribution, have good coverage of loan applicant population, and is applicable to most banking contexts.

Based on the boxplots, it was observed that there were outliers in the data for the continuous variables. The outliers were imputed with each variable's median, to normalize the data values for improved reflection of a normal distribution.

3. Data preparation

3.1. Dataset Description

The original dataset had both characters and numerical. The characters were assigned ranks and the table below summarizes the interpretation of the ranking system.

Attribute	Variable Name	Ranking System
1 (qualitative)	Account Balance	1 : ... < 0 DM 2 : 0 <= ... < 200 DM 3 : ... >= 200 DM / salary assignments for at least 1 year 4 : no checking account
2(numerical)	Duration of Credit	
3(qualitative)	Payment Status of Previous Credit	0 : no credits taken/ all credits paid back duly 1 : all credits at this bank paid back duly 2 : existing credits paid back duly till now 3 : delay in paying off in the past 4 : critical account/ other credits existing (not at this bank)
4(qualitative)	Purpose	0 : car (new) 1 : car (used) 2 : furniture/equipment 3 : radio/television 4 : domestic appliances 5 : repairs 6 : education 7 : (vacation - does not exist?) 8 : retraining 9 : business 10 : others
5(numerical)	Credit amount	
6(qualitative)	Value Savings/Stocks	1 : ... < 100 DM 2 : 100 <= ... < 500 DM 3 : 500 <= ... < 1000 DM 4 : .. >= 1000 DM

		5 : unknown/ no savings account
7(qualitative)	Length of current employment	1 : unemployed 2 : ... < 1 year 3 : 1 <= ... < 4 years 4 : 4 <= ... < 7 years 5 : .. >= 7 years
8 (numerical)	Instalment per cent	
9 (qualitative)	Sex & Marital Status	1 : male : divorced/separated 2 : female : divorced/separated/married 3 : male : single 4 : male : married/widowed 5 : female : single
10(qualitative)	guarantors	1 : none 2 : co-applicant 3 : guarantor
11(numerical)	Duration in Current address	
12(qualitative)	Most valuable available asset	1 : real estate 2 : building society savings agreement/ life insurance 3 :car or other 4 : unknown / no property
13(numerical)	Age..years.	
14(qualitative)	Concurrent.Credits	1 : bank 2 : stores 3 : none
15(qualitative)	Type.of.apartment	1 : rent 2 : own 3 : for free
16(numerical)	No.of.Credits.at.this.Bank	
17(qualitative)	Occupation	1 : unemployed/ unskilled - non-resident 2 : unskilled - resident 3 : skilled employee / official 4 : management/ self-employed/highly qualified employee/ officer
18(numerical)	No.of.dependents	
19(qualitative)	Telephone	1 : none 2 : yes, registered under the customers name
20(qualitative)	Foreign.Worker	1 : yes

		2 : no

3.2. Select data

3.2.1. Rationale for Inclusion/ Exclusion

In the data exploration section, we identified the a subset of potential predictor variables for building the classification model, tabulated below, based on their impact on 'Creditability', as determined by the significance of t-tests and chi-square tests performed during the exploratory data analysis step.

Variable
Account.Balance
Payment.Status.of.Previous.Credit
Value.Savings.Stocks
Length.of.current.employment
Purpose
Value.Savings.Stocks
Length.of.current.employment
Sex...Marital.Status
Most.valuable.available.asset
Type.of.apartment
No.of.Credits.at.this.Bank
Concurrent.Credits
Duration.of.Credit..month.
Credit.Amount

3.3. Clean data

3.3.1. Data Cleaning Report

While selecting the data, it was noticed that there were some categorical data and three numerical variables. Hence, we had to treat them separately. The categorical data was converted into factors to differentiate them from numerical. In either case, the dataset was checked for missing variables. Since there were no missing variables in the dataset, we moved to checking for outliers in the numerical section of the dataset. The following boxplots were created to visually check for outliers.

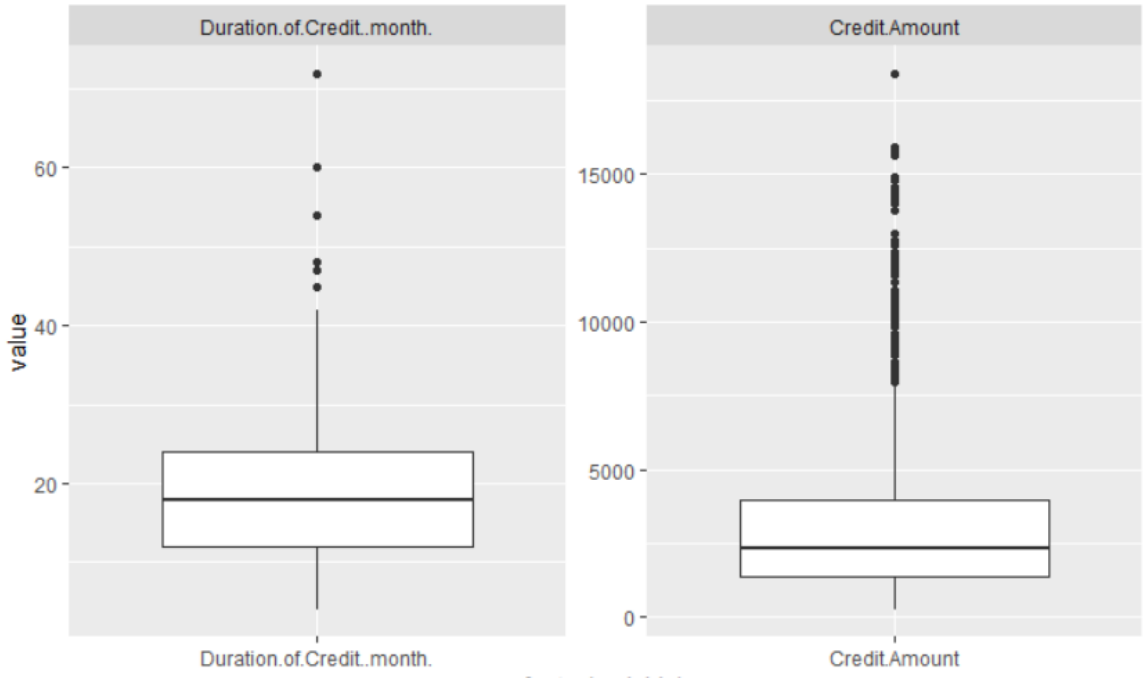

Figure 9. Box plots depicting the distribution of 'Duration of Credit (Month)' and 'Amount of Credit'

Based on the boxplots (Figures 3, 5, 7) we can see that all three continuous variables in the German credit dataset have outliers (extreme values), as there are data values that are far away from the quartiles, extending beyond the boxplots' whiskers. However, only Credit Amount and Duration of Credit Month has the most outliers (Figure 9). So, we imputed the column outliers with the column median for these continuous variables. Specifically, an interquartile range function was created and the median was imputed for the outliers, as shown in the enclosed R program.

3.4. Construct data

The continuous variables were then normalized in order to have the weights and scaling equal between them, after which we performed Principal Component Analysis to see if one of them could be discarded (Figure 10).

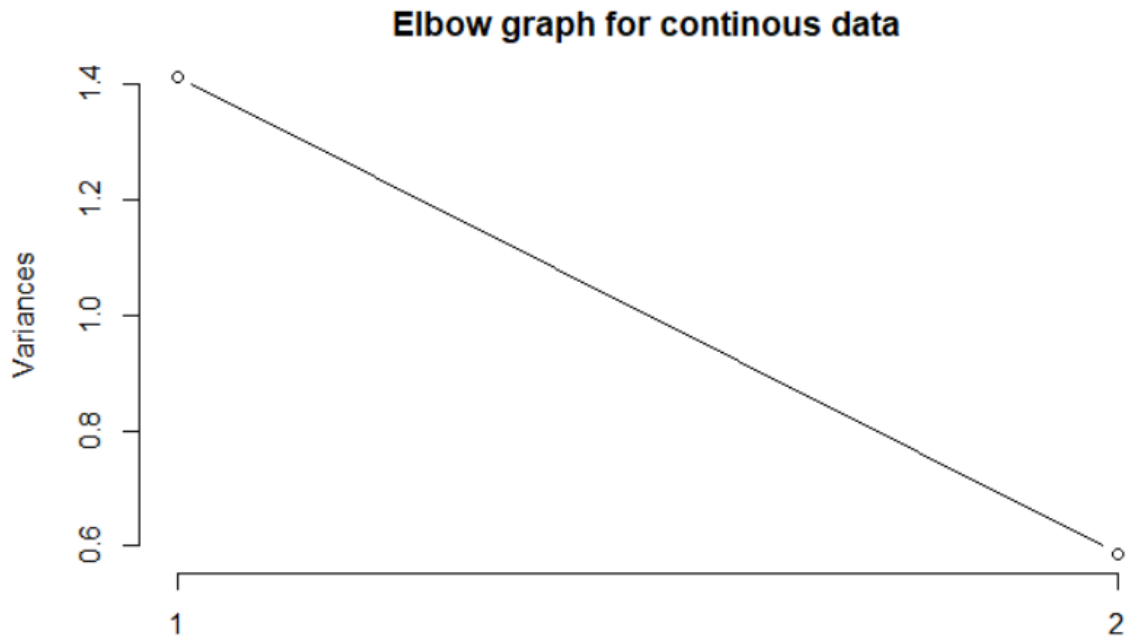

Figure 10. Elbow graph for continuous data

We then did a PCA on the categorical data by using a one hot encoding on the variables (Figure 11).

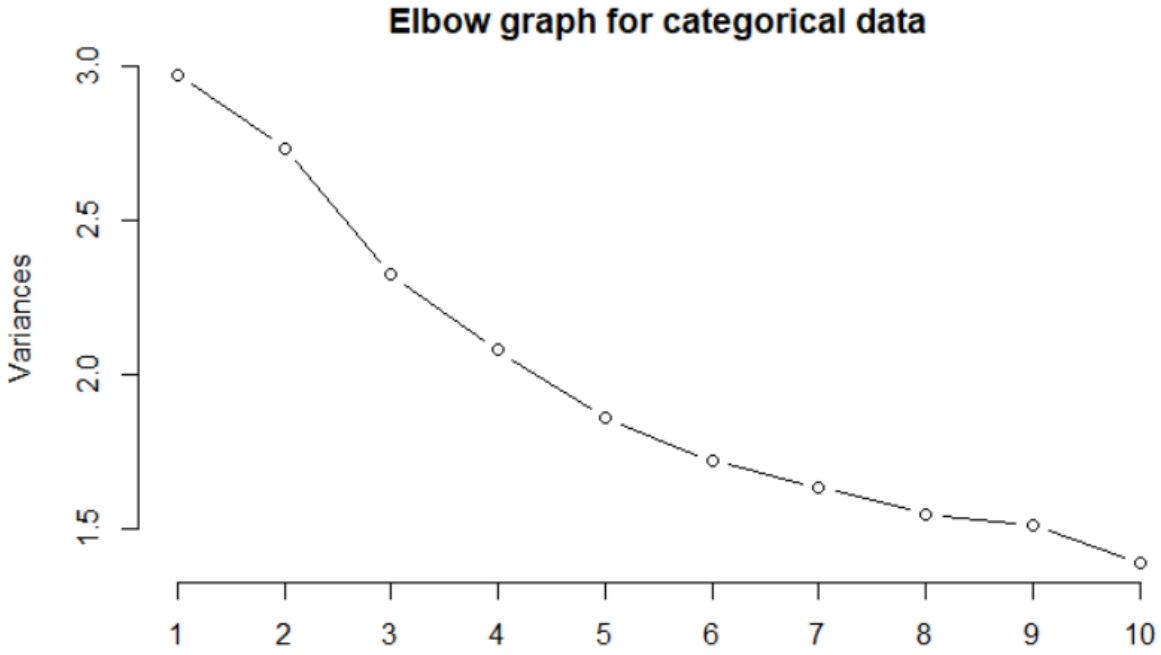

Figure 11. Elbow graph for categorical data

Based on the above two graphs, it was determined a total of 9 variables would be suffice for inclusion in the model. The list of the variables is included in the modeling section (Section 4).

3.5. Integrate data

3.5.1. Merged Data

After preparing the data, we merged the continuous and the categorical data into one dataset.

3.6. Format data

3.6.1. Reformatted Data

The data did not need to be formatted because it was transformed earlier in the stage.

4. Modelling

4.1. Selecting the Modelling technique

4.1.1. Modelling technique

Since this project is about a Supervised classification model where the Creditability (i.e. whether the bank should grant loan to a consumer or not) of a consumer record is the Target variable, and also because we are trying to predict the Creditability of the consumer, we decided to use the Random Forest algorithm. As noted in Section 1.4.2, 'Initial Assessment of Tools and Techniques, a Random Forest classification model was assessed to be suitable for the business goals of this project, as it is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time.

4.1.2. Modelling assumptions

Random Forest uses Bootstrap aggregation method, and assumes that sampling is representative. The German credit data we selected for this project has 70% records with Creditability = 1 and 30% records with Creditability =0. But since Random Forest is an ensemble of multiple Decision trees and does not depend on the variables to be independent, it performs better than Decision tree and Naïve Bayes algorithm. We also included Random sampling techniques to make sure the Test and Training sets are representative, which in turn validates the assumption that sampling is representative.

4.2. Generate test design

4.2.1. Test design

The original dataset was divided into Training and Testing sets in 7:3 ratio. Also, Random sampling techniques (using Caret package) were added to ensure the both Training and Test sets are representative of the original data's Creditability ratio i.e. both Training and Test sets contains the same 30% of records with Creditability = 0. Specifically, using R's createDataPartition() function, the random sampling was done within the levels of Creditability to balance the class distributions within the splits. Model was evaluated using a Confusion matrix.

4.3. Build model

4.3.1. Parameter settings

The model is created using the "randomForest" function from "randomForest" package in R. The following 9 variables from the dataset were passed to the function as Predictor variables for creating the trees:

- Account.Balance
- Payment.Status.of.Previous.Credit
- Purpose
- Value.Savings.Stocks
- Sex...Marital.Status
- Most.valuable.available.asset
- No.of.Credits.at.this.Bank
- Duration.of.Credit..month.
- Credit.Amount

These variables were identified as significant based on our data pre-processing.

4.3.2. Models

The Random Forest model was created using the training dataset, and using the 9 variables described above as predictors. The plot of the model is shown in Figure 12 below.

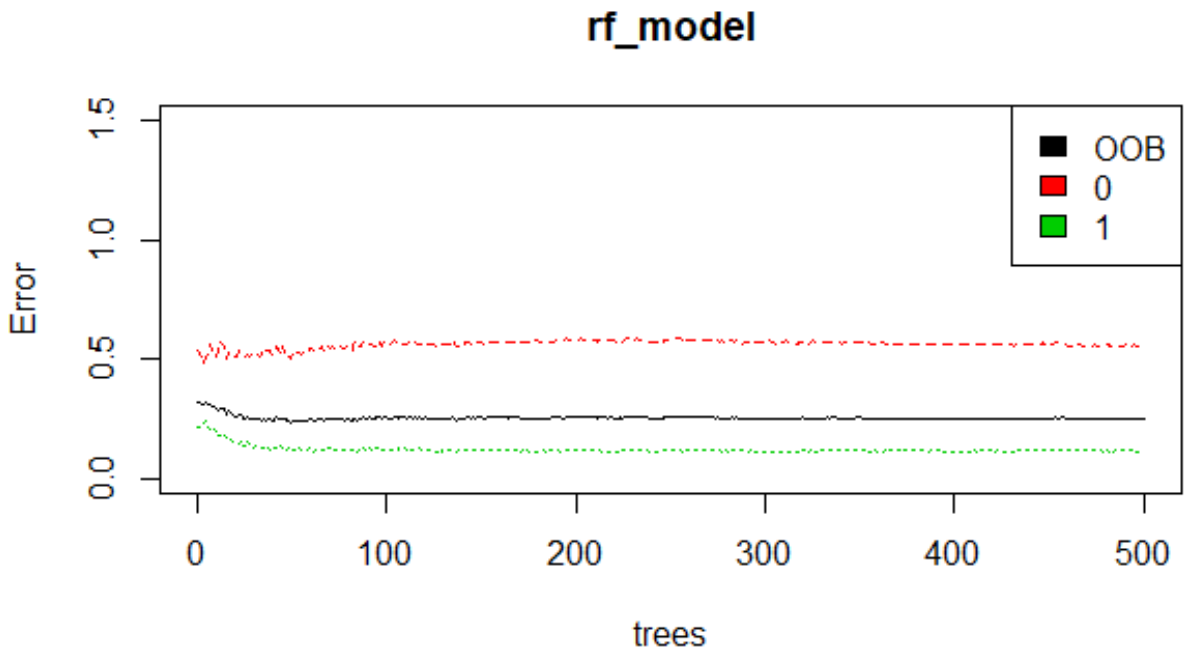

Figure 12. Plot of the random forest model

4.3.3. Model description

From the plot of the model, we can see that Random Forest has used up to 500 trees for making the prediction and we can also see that the error rate for records with Creditability = 0 is higher than the error rate for records with Creditability = 1. This is mainly because of the composition of the original dataset, we only had 30% records with Creditability = 0 in the original dataset. The model is able to predict more accurately when the number of records is higher.

4.4. Assess model

4.4.1. Model assessment

Let us assume that a correct decision of the bank would result in 35% profit at the end of 5 years. A correct decision here means that the bank predicts an application to be good or credit-worthy and it actually turns out to be credit worthy. When the opposite is true, i.e. bank predicts the application to be good but it turns out to be bad credit, then the loss is 100%. If the bank predicts an application to be non-creditworthy, then loan facility is not extended to that applicant and bank does not incur any loss (opportunity loss is not considered here).

Out of 1000 applicants, 70% are creditworthy. A loan manager without any model would incur [0.7*0.35 + 0.3 (-1)] = - 0.055 or 0.055 unit loss. If the average loan amount is 3200 DM (approximately), then the total loss will be 1760000 DM, and per applicant loss is 176 DM.

The cost matrix for our model is as follows:

		Predicted	
		Credit worthy	Non-Credit worthy
Actual	Credit worthy (1)	179 (Weightage= +.35)	30
	Non-Credit worthy(0)	53 (Weightage=-1)	38

Unit profit = ((179/300) * 0.35) – ((53/300)*1) = 0.038

Per Applicant profit = Unit profit * Average loan amount = 121.6 DM

5. Evaluation

5.1. Results Evaluation and Conclusion

5.1.1. Assessment of data mining results with respect to business success criteria

As per the business success criteria defined earlier, a successful outcome is, when the bank makes accurate decisions that result in a measurable profit within a period of time. Based on the Cost matrix described in the above section we can see that we are able to create a profitable model with a per applicant profit of 121.6 DM.

We also created a Confusion matrix using the Caret package and from the Confusion matrix we can see that the model is 72.33% accurate. Given below are some of the statistics generated from Caret package's Confusion matrix:

```
               Accuracy : 0.7233
                 95% CI : (0.669, 0.7732)
    No Information Rate : 0.6967
    P-Value [Acc > NIR] : 0.17334

                  Kappa : 0.2951
 Mcnemar's Test P-Value : 0.01574
```

```
            Sensitivity : 0.4176
            Specificity : 0.8565
         Pos Pred Value : 0.5588
         Neg Pred Value : 0.7716
             Prevalence : 0.3033
         Detection Rate : 0.1267
   Detection Prevalence : 0.2267
      Balanced Accuracy : 0.6370

       'Positive' Class : 0
```

5.2. Next steps and Process review

5.2.1. Review of Processes

The goal of this project was to create a profitable model that could be used by the bank to make better decisions and this would in turn ensure risk minimization and profit maximization. We had selected the German Credit data from the UCI repository as our data source, and used Random Forest algorithm from the "randomForest" R package to create our model using CRISP-DM methodology.

Though the project has met the Business success criteria we can still see that there is a lot of scope for improvement. The Error rates for non-credit worthy records (Creditability = 0) is much higher when compared to the Error rates for Credit worthy records (Creditability = 1). The Error rates for non-credit worthy records can be improved in future models if we can train the model with a higher number of non-credit worthy records.

5.2.2. Next steps

The model has per applicant profit of 121.6 DM and an accuracy of 72.33% and hence it meets the Business success criteria and Data mining success criteria set at the outset of our project, described in the first two sections of this report. Our model could be deployed in any financial institution that requires a classification tool for credit risk or loan extension decision making; and based on feedback from using the model with new banking data, the model could be improved, through additional machine learning and data mining steps, for an even higher classification performance.

Appendix A

Glossary of Terms

Algorithm: A machine learning algorithm is any algorithm that can produce a model by analyzing a dataset.

Bar Plot: A chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

Boxplot: A standardized way of displaying the distribution of data based on the five-number summary: minimum, first quartile, median, third quartile, and maximum.

Categorical Variable: A variable that can take on one of a limited and usually fixed number of possible values.

Chi-square Statistic: A nonparametric statistical measurement of how expectations compare to results.

Classification Algorithm: A machine learning algorithm that uses labeled examples to create a model that can be used for classification.

Confusion Matrix: A table that summarizes how successful the classification model was at predicting examples belonging to various classes; one axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

Contingency Table: A table showing the distribution of one variable in rows and another in columns, used to study the association between the two variables.

Continuous Variable: A variable that has an infinite number of possible values; any value is possible for the variable.

Cost Matrix: An evaluation tool for classification models used when specific classification errors can be more severe than others; it calculates the cost of wrong prediction or right prediction, with a provision that some mis-predictions can be very costly.

Credit Risk: The risk of default on a debt that may result if a borrower fails to make required payments.

CRISP-DM: A methodology that provides a structured approach to planning a data mining project.

Cross-classification: A technique that measures the changes in one variable in a dataset when other variables are accounted for.

Data Mining: The process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Data Preprocessing: A data mining technique that involves transforming raw data into an understandable format.

Data Quality: An assessment of data's fitness to serve its purpose in a given context.

Decision Tree: A decision support tool that uses a tree-like graph or model of decisions and their possible consequences.

Distribution: The distribution of a statistical dataset (or a population) is a listing or function showing all the possible values (or intervals) of the data and how often they occur.

Exploratory Data Analysis: An approach to analyzing datasets by summarizing their main characteristics; it is used for seeing what the data can depict beyond the formal modeling or hypothesis testing task.

Feature: An attribute of a data point, usually a part of a feature vector; it can be numerical or categorical.

Histogram: A diagram consisting of rectangles whose area is proportional to the frequency of a variable and whose width is equal to the class interval.

Imputation: The process of replacing missing or outlier data with substituted values.

Machine Learning: A subfield of computer science, mathematics, and statistics that focuses on the design of systems that can learn from and make decisions and predictions based on data.

Mean: The mean or average that is used to derive the central tendency of a dataset; determined by adding all the data points in a population and then dividing the total by the number of points.

Median: The value separating the higher half from the lower half of a data sample.

Missing Values: Occur when no data value is stored for the variable in an observation.

Normal Distribution: A function that represents the distribution of many random variables as a symmetrical bell-shaped graph.

Outlier: An observation that lies outside the overall pattern of a distribution.

Overfitting: Occurs when the machine learning algorithm learns a model that fits the training data too well by incorporating details and noise specific to the training data, which results in poor prediction of the labels of examples from the validation set.

P-value: The level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event.

Predictor Variable: A variable whose values will be used to predict the value of the target variable.

Quartile: Each of four equal groups into which a population can be divided according to the distribution of values of a particular variable.

Random Forest: A supervised learning algorithm that creates a forest and makes it random; the "forest" is a combination of decision trees, most of the time trained by a combination of learning models that enhances the overall result.

Range: The difference between the lowest and highest values in a dataset.

Summary Statistics: Are used to summarize a set of observations, in order to communicate the largest amount of information as simply as possible.

Target Variable: The variable that is or should be the output of the machine learning model.

T-test: Test used to determine whether the mean of a population significantly differs from a specific value (called the hypothesized mean) or from the mean of another population.

Welch Two Sample T-test: A two-sample location test used to test the hypothesis that two populations have equal means.