CIND 119 – Class Final Project

German Credit Dataset Analysis and Predictive Modeling

By: Diego Tejada Cardenas & Kasra Foroughi

Class & Section: CIND 119 – DHA

June 22, 2021

Ryerson University

Summary

A German bank requires insight on determining which customers meet the necessary requirements to be approved of a bank loan. Our goal as data scientists is to analyze the German Credit dataset and through predictive modeling determine whether a customer (new or existing) is suitable to be approved of a loan. The dataset provided includes 20 defining attributes and a class attribute (Creditability) which represents a good or bad credit risk. The attributes were cross referenced with Creditability to determine the highest correlation and were selected and used to create predictive modeling via Decision Tree and Naïve Bayes classification methods. Both predicted models had a higher accuracy rating when a selected dataset was used (Decision Tree: 80% vs 85%, Naïve Bayes: 75% vs 77%). For the German bank, we recommend using Decision Tree predictive modeling assessment to determine whether a customer has met the requirements to be approved of a bank loan due to a higher accuracy, precision and recall ratings.

Tools:

In this data analysis study, we utilized the programming tool R to perform data preparation methods, predictive modeling classifications, and visualizations.

Workload Distribution

Member Name	**List of Tasks Performed**
Diego	Data Preparation, Decision Tree Classification, Predictive modeling , presentation
Kasra	Data Preparation, Naïve Bayes Classification, Predictive modeling, presentation

Data Preparation

The German Credit dataset provided (Figure 5) is compromised of 1000 observations with 20 attribute variables and 1 class attribute (Creditability). Creditability is a binary value which represents either a good credit rating (1) or bad credit rating (0). The dataset had a 700 to 300 ratio of good credit rating vs bad credit rating (Creditability of 1 and 0 respectively). The dataset provided was complete (no missing values) with a corresponding legend for categorical values. Of the 20 attribute values, 3 of them were purely numerical (Duration of Credit, Credit Amount, No of Credits at this Bank) (Figure 6.), whilst the rest were qualitative values. For the qualitative values, they were transformed into factors to better understand the data provided.

To better understand the correlation of the data, we performed a correlation plot (Figure 1) on all the variables. The correlation values were then used to determine which attributes were to be selected for the Decision Tree classification. For Naïve Bayes classification, to determine correlation a chi-square test was performed as it best represents correlation between categorical values. The selected values were based on p values

which were less than a significant level (<0.05), which were then assumed to be independent of the class attribute. The values that were excluded are explained in more detail in their respective classification sections.

Predictive Modeling

Data Splicing

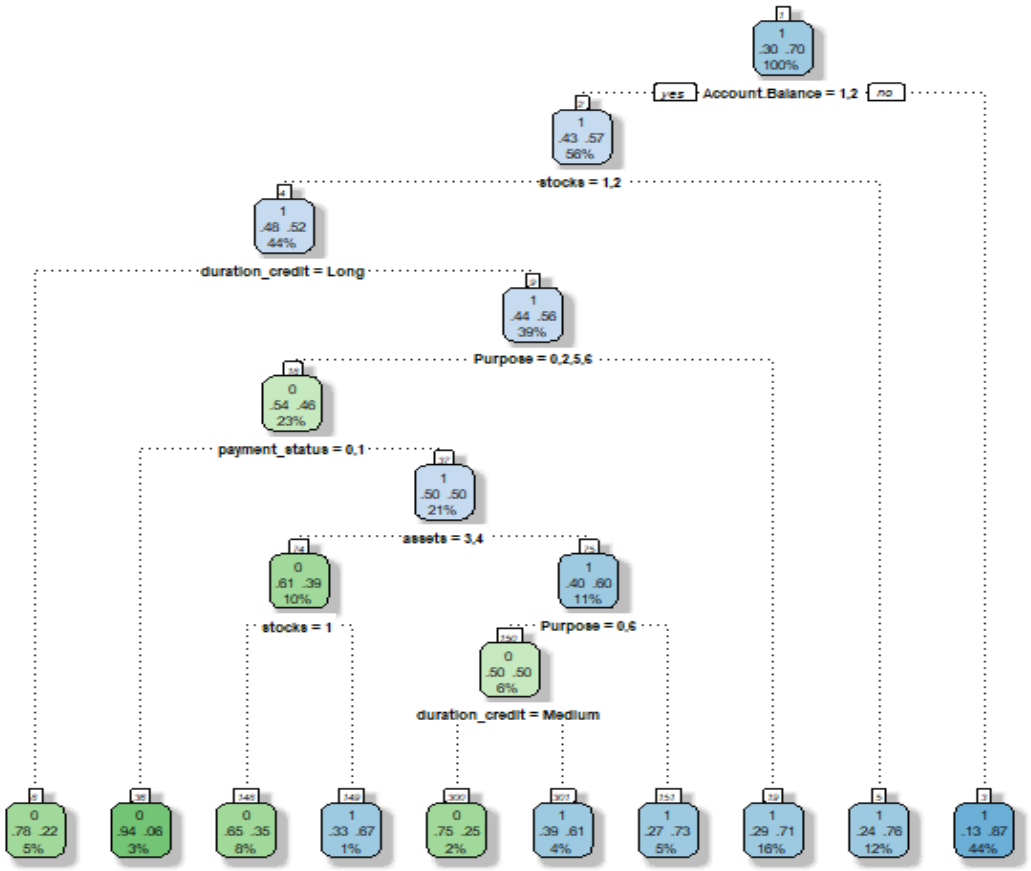
For our data split strategy, we used the test-train set split in both classification methods. The training set compromised 70% of the dataset being used and the test set compromised the remaining 30% (the split percentages were used based on common data split strategies). The training set was then used to predict the machine learning model and the test set was used to evaluate the training model and predict how accurate the model was in relation.

Decision Tree Classification

The decision tree in this model was run with the original values and a training set, which include selected variables based on correlation test (Figure 1), regression test, and graph exploration (plots, histograms, and kernel density plots). The original values without balancing result in an accuracy average of 95%, suggesting that the data is overfitting the model. Moreover, it required other approaches such as factoring the variables. The result of balancing and factoring the original values is a reduction accuracy to 80% average that allow us to assess a minimum baseline for the selected variables. The selected variables were the account balance, duration of credit, payment status, valuable assets, stocks, installments, purpose, and age. The accuracy result of the selected variables is 85% average, moreover, the other metrics such as recall, and precision indicated higher performance of the selected variables in relation to entire dataset. The complete result of the metrics can be observed in the graph below.

	Balance Dataset	Selected Dataset
Accuracy	**0.80**	**0.85**
TP Rate	**0.63**	**0.67**
FP Rate	**0.13**	**0.11**
Precision	**0.82**	**0.85**
Recall	**0.91**	**0.94**

The decision tree suggest that the class attribute (Creditability) is affected positively by the higher age, money in their account, credit record, assets, and stocks. On the other hand, it was negativity affected by installments percent and purpose. The variable purpose was only negative if the purpose was to buy assets that depreciate such as cars, electronic and similar but it changes with assets such as houses.



Selected variables

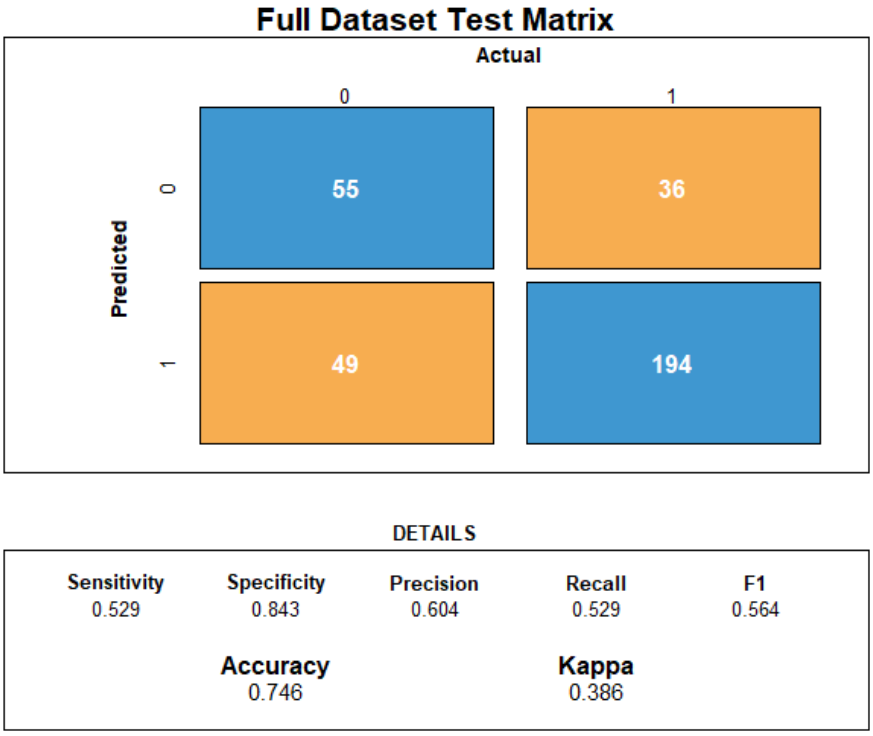
Naïve Bayes Classification

Naïve Bayes classifier is a statistical predictive model based on the Baye's theorem which assumes that the variables being used to predict the model are independent of one another. The class attribute (Creditability) was cross referenced with the 20 attribute values and a chi-square test was performed (Figure 2). The reason for this is that a majority of the attributes are classified as qualitative (categorical), thus a normal correlation/linear regression can't be used. 2 predictive modeling algorithms were used in this section, one with the entire data set and one with selected values.

The selected values used were; **Creditability, Payment Status of Previous Credit, Account Balance, Value Savings/Stocks, Length of Current Employment, Most valuable Available Asset, Sex & Marital Status, and Guarantors** (Figure 3)**.** These values all had a significant p value (< 0.05) and were deemed independent which meets the necessary assumption of Naïve Bayes. 3 numerical values (Credit Amount, Age, and Duration of Credit) were excluded from this selection as they weren't categorical and would in turn skew results. The attributes of purpose, concurrent credits, type of apartment, and foreign worker did have significant p-values, but were not included as they hindered the accuracy of the modeling process.

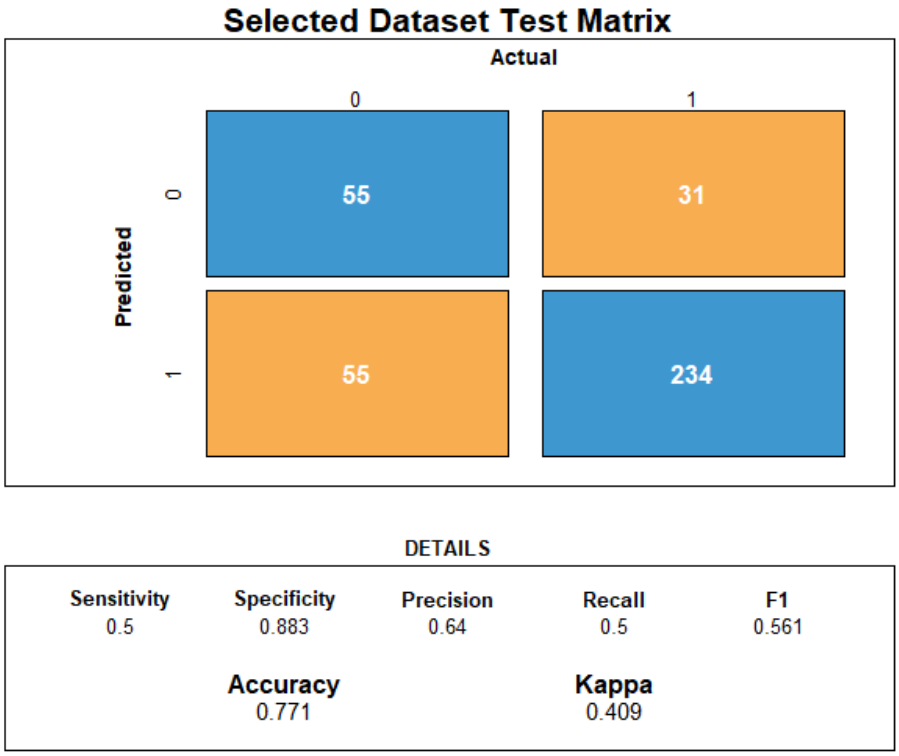
a) Entire Dataset Naïve Bayes Classification

The Naïve Bayes model was implemented on the entire dataset, it resulted in an **accuracy** of 75% (test matrix) with a 95% CI range of 70 % to 79%. This prediction indicates that when using the entire data set of attributes, the model correctly predicts true positive (TP) and true negative (TN) values 74.6 % of the time. The **true positive rate (TPR) of the dataset is 53%** and a **true negative rate (TNR) is 84%.**



b) Selected Dataset Naïve Bayes Classification

The Naïve Bayes classification implemented on the selected values resulted in an accuracy of 77% (test matrix) with a 95% CI : 72% to 81%. Using the selected values to perform a Naïve Bayes classifier nets an increase in accuracy of the prediction model by roughly ~2%. The **TPR is 50% and a TNR of 88%**



c) Naïve Bayes Conclusion

After performing the Naïve Bayes prediction model, the selected dataset is deemed to have a better performance rating with a higher accuracy (2% increase) and TNR (4% increase). This indicated that the selected values have a higher chance of predicting whether a customer is suitable to be approved of a loan rather than using the entire data set.

Conclusion

Through data preparation, we were able to deduce the best values for Decision Tree classification utilizing correlation and chi-square test for Naïve Bayes classification. Overall, there was no attribute that had a significant direct correlation with Creditability, with the highest being Account Balance. However, in the data, a portion of the customers who had a good credit (Creditability of 1) had no checking account (Account Balance value of 4) (Figure A). We would recommend a more thorough dataset that can differentiate why these individuals are deemed to have a higher Creditability rating.

Figure A.

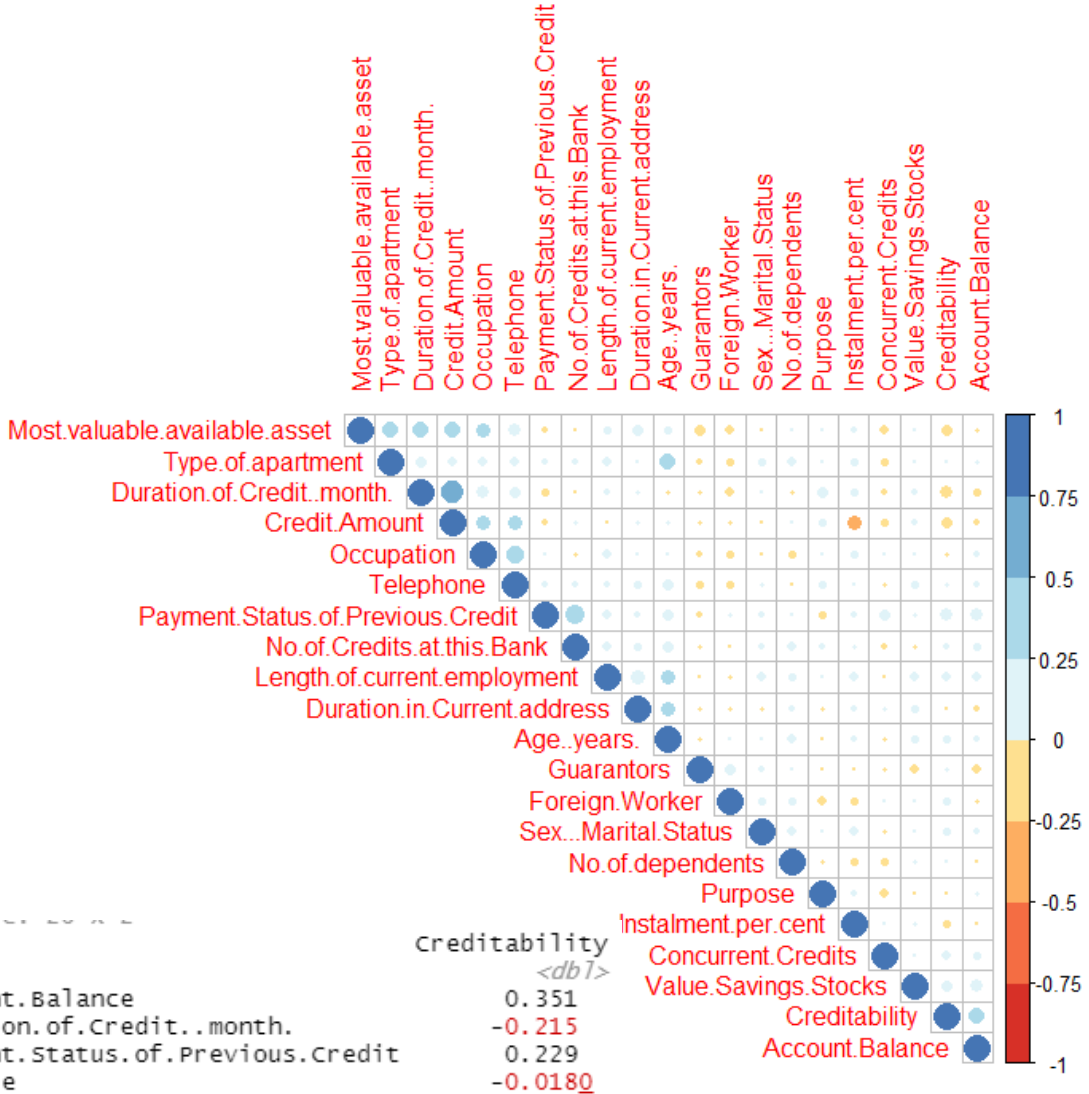
	1	2	3	4
0	135	105	14	46
1	139	164	49	348

Through our data analysis study of the German Credit dataset, we have concluded that both the Decision Tree and Naïve Bayes classification have a high accuracy rating when utilizing selected data (85% and 77% respectively). Decision Tree method is a better classification method in this scenario as it nets a higher Recall and Precision value in comparison to Naïve Bayes classifier. The selected variable (Decision Tree) show us that Creditability was highly affected by higher age, money in their account, credit record, assets, and stocks. This makes sense as all those factors showcase the person has experience with handling money and are good candidates to be approved of loans. For Naïve Bayes prediction model, the highest evaluation metric was the specificity which is useful to determine the customers who don't have a good credit rating and will not be getting a bank loan.

We have noticed that the majority of possible customers are young adults (20-40), which have the propensity to increase their consumption capacity and investment. However, the lack of robust credit history does not allow them to obtain higher levels of loans. The size of their loans and time is too short to determine the risk, for that reason, we suggest the inclusion of new metrics, focus on these uncapped customers, such as investment/consumption or risk associated with major possible expenditures like driving record, work hazards, etc. According to the decision tree, the most important nodes include the age, money in their account and credit record. The inclusion of the metrics would partially solve the uncapped customer issue. To fully obtain the benefits, we suggest the creation of a micro-loan system (for investments) in which the micro-loans are tied to an account with higher data access (transaction). Allowing the bank to increase their risk assessment of this target population and the number of loans given, but at the same time, obtaining the probability of getting new customers with a higher credit record that in the long term will be more beneficial.

Appendix

Figure 1. Correlation between all variable types (pre factor)



	term	Creditability
	<chr>	<dbl>
1	Account.Balance	0.351
2	Duration.of.Credit..month.	-0.215
3	Payment.Status.of.Previous.Credit	0.229
4	Purpose	-0.0180
5	Credit.Amount	-0.155
6	Value.Savings.Stocks	0.179
7	Length.of.current.employment	0.116
8	Instalment.per.cent	-0.0724
9	Sex...Marital.Status	0.0882
10	Guarantors	0.0251
11	Duration.in.Current.address	-0.00297
12	Most.valuable.available.asset	-0.143
13	Age..years.	0.0913
14	Concurrent.Credits	0.110
15	Type.of.apartment	0.0181
16	No.of.Credits.at.this.Bank	0.0457
17	Occupation	-0.0327
18	No.of.dependents	0.00301
19	Telephone	0.0365
20	Foreign.Worker	0.0821

Figure 2. Chi-square test – Comparing correlation between categorical values

Row	Column	Chi.Square	df	p.value
Creditability	Account.Balance	123.721	3	0.000
Creditability	Duration.of.Credit..month.	78.887	32	0.000
Creditability	Payment.Status.of.Previous.Credit	61.691	4	0.000
Creditability	Purpose	33.356	9	0.000
Creditability	Credit.Amount	931.746	922	0.405
Creditability	Value.Savings.Stocks	36.099	4	0.000
Creditability	Length.of.current.employment	18.368	4	0.001
Creditability	Instalment.per.cent	5.477	3	0.140
Creditability	Sex...Marital.Status	9.605	3	0.022
Creditability	Guarantors	6.645	2	0.036
Creditability	Duration.in.Current.address	0.749	3	0.862
Creditability	Most.valuable.available.asset	23.720	3	0.000
Creditability	Age..years.	57.627	52	0.275
Creditability	Concurrent.Credits	12.839	2	0.002
Creditability	Type.of.apartment	18.674	2	0.000
Creditability	No.of.Credits.at.this.Bank	2.671	3	0.445
Creditability	Occupation	1.885	3	0.597
Creditability	No.of.dependents	0.000	1	1.000
Creditability	Telephone	1.173	1	0.279
Creditability	Foreign.Worker	5.822	1	0.016

Figure 3. Naïve Bayes Selected Variables Data Points

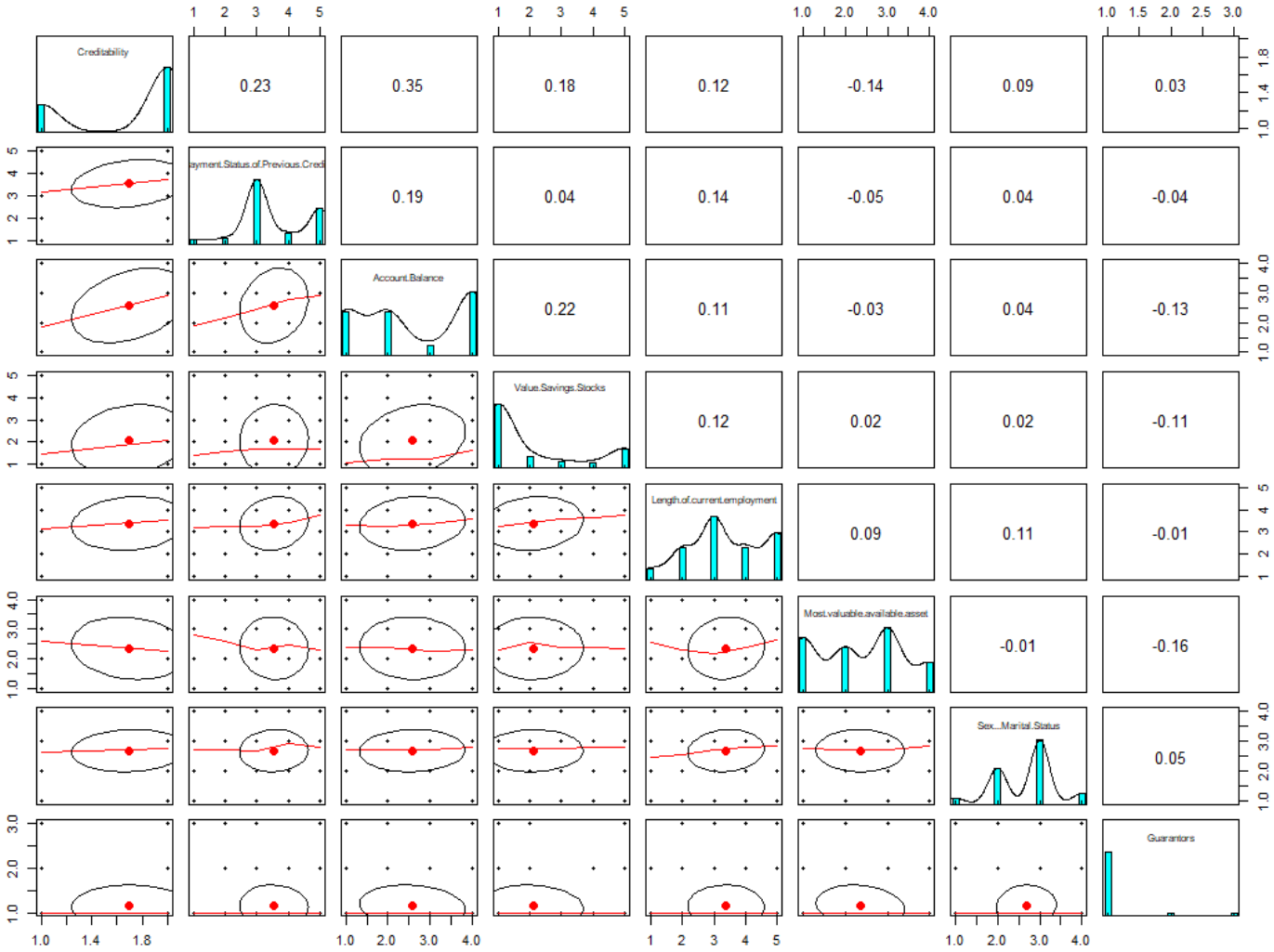
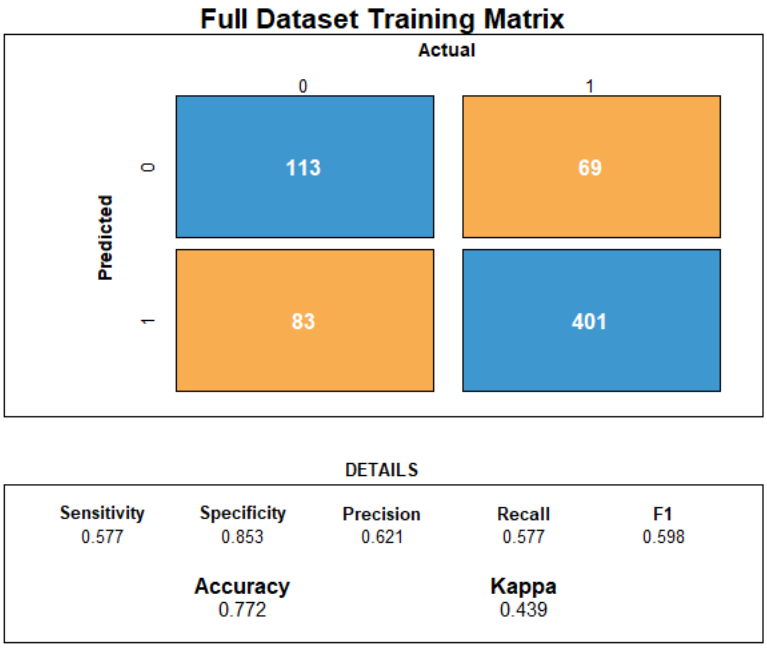


Figure 4. Training set Confusion Matrix for Naïve Bayes

4a)



4b)

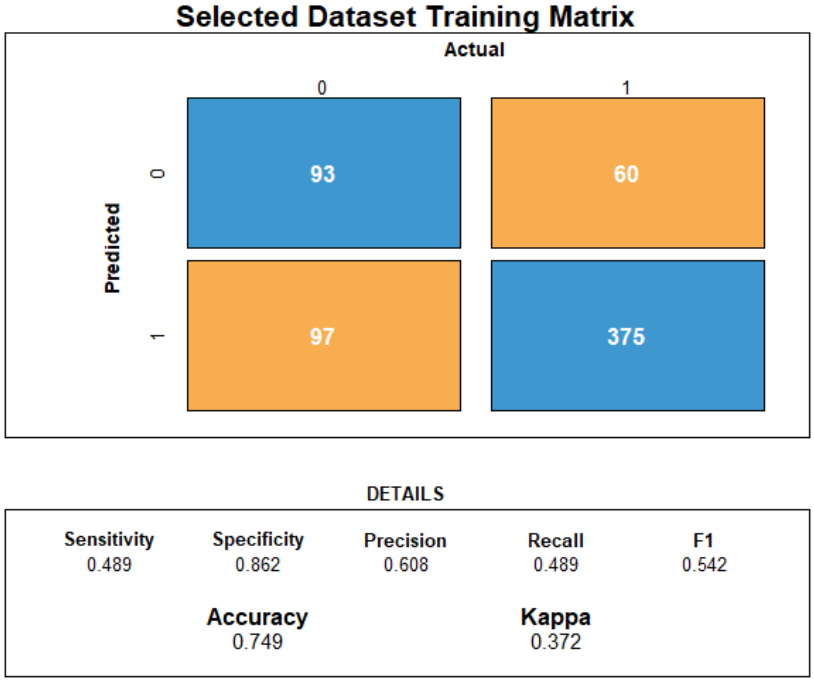


Figure 5 – Dataset values and attribute types

Variables	Type	Mean	St. Dev	Min	Percentile (25)	Percentile (75)	Max
Creditability (Class Attribute)	Discrete	-	-	0	-	-	1
Duration of Credit month.	Discrete	20.903	12.059	4	12	24	72
Payment Status of Previous Credit	Nominal	2.545	1.083	0	2	4	4
Purpose	Nominal	2.828	2.744	0	1	3	10
Credit Amount	Continuous	3,271	2,822	250	1,365.5	3,972	18,424
Value Savings/Stocks	Ordinal	2.105	1.580	1	1	3	5
Length of current employment	Nominal	3.384	1.208	1	3	5	5
Instalment per cent	Nominal	2.973	1.119	1	2	4	4
Sex	Nominal	2.682	0.708	1	2	3	4
Guarantors	Nominal	1.145	0.478	1	1	1	3
Duration in Current address	Discrete	2.845	1.104	1	2	4	4
Valuable asset	Continuous	2.358	1.050	1	1	3	4
Age.	Continuous	35.542	11.353	19	27	42	75
Concurrent Credits	Nominal	2.675	0.706	1	3	3	3
Type of apartment	Ordinal	1.928	0.530	1	2	2	3
No of Credits at this Bank	Discrete	1.407	0.578	1	1	2	4
Occupation	Nominal	2.904	0.654	1	3	3	4
No of dependents	Discrete	1.155	0.362	1	1	1	2
Telephone	Nominal	1.404	0.491	1	1	2	2
Foreign Worker	Nominal	1.037	0.189	1	1	1	2
Account Balance	Nominal	2.577	1.258	1	1	4	4

Figure 6 – Numerical values Box plots