German Credit Card Analysis

By: **Kanchana Sundaralingam**

Summary

Business Problem:

Credit risk scoring and assessment is very critical in the banking sector. In order to qualify to apply for loans, a loan applicant needs to be approved by a bank. The primary aim of this project is to choose and obtain the best model to aid in the prediction of creditability in the classification of a loan application. Creditability, the class attribute, is binary and comprises two categories. They are namely, “Good” or “Bad”. There are 20 attributes used in the credit risk assessment of a loan applicant.

This report conducts an analysis on German Credit Scoring and outlines the steps involved in determining if a loan applicant is qualified to apply for loans based on his or her creditability. Data analytics tools such as R, Weka and SAS were used to predict the creditability of a loan applicant.

The data is tested and prepared by identifying outliers, in addition to finding for any missing data and deviations. Additionally, by means of correlation and ranking, the attributes which proved to be more correlated to the class attribute were found.

Classification is an important form of data analysis that extracts models describing important data classes. In light of this, for the purpose of our project, classification algorithms such as the Decision Tree and Naive Bayes were carried out on the sub datasets. In the conclusion and recommendation section, the best group was selected to find the association between the attributes and creditability.

Data Preparation

The data set that is being analyzed consists of 21 attributes including a binary class attribute (Creditability) that classifies the loan applicant into one of the two categories, namely, “Good” or “Bad”. There are 20 attributes used in judging a loan applicant in which 6 are numerical and 14 are ordinal/nominal.

- *Attribute Type*

No.	Attributes	Type
1	Account Balance	Qualitative
2	Duration of Credit (month)	Numeric
3	Payment Status of Previous Credit	Qualitative
4	Purpose	Qualitative
5	Credit Amount	Numeric
6	Value Savings/Stocks	Qualitative
7	Length of current employment	Qualitative
8	Installment per cent	Numeric
9	Sex & Marital Status	Qualitative
10	Guarantors	Qualitative
11	Duration in Current address	Qualitative
12	Most valuable available asset	Qualitative
13	Age (years)	Numeric
14	Concurrent Credits	Qualitative
15	Type of apartment	Qualitative
16	No of Credits at this Bank	Numeric
17	Occupation	Qualitative
18	No of dependents	Numeric
19	Telephone	Qualitative
20	Foreign Worker	Qualitative
21	Creditability	Qualitative

- *Missing Values*

All the values in the attributes are complete and there are no missing values in the dataset. This was further confirmed when running the dataset in Weka. From the Weka filters on the main tab on the Weka tool, Missing Values for the attributes were shown to be 0% and as such there was no need to deal with missing values.

- *Max, Min, Mean and Standard Deviation of Attributes.*

For each numerical attribute, the max, min, mean and standard deviation were calculated using SAS as shown in Figure 1 below.

The SAS System

The MEANS Procedure

Variable	Mean	Std Dev	Minimum	Maximum
Creditability	0.70	0.46	0.00	1.00
Duration_of_Credit__month_	20.90	12.06	4.00	72.00
Credit_Amount	3271.25	2822.75	250.00	18424.00
Instalment_per_cent	2.97	1.12	1.00	4.00
Sex___Marital_Status	2.68	0.71	1.00	4.00
Age__years_	35.54	11.35	19.00	75.00
No_of_Credits_at_this_Bank	1.41	0.58	1.00	4.00
No_of_dependents	1.16	0.36	1.00	2.00

Figure 1: Mean, Standard Deviation, Minimum and Maximum values for each numerical attribute in the dataset

- *Outlier values (records) for each the attributes*

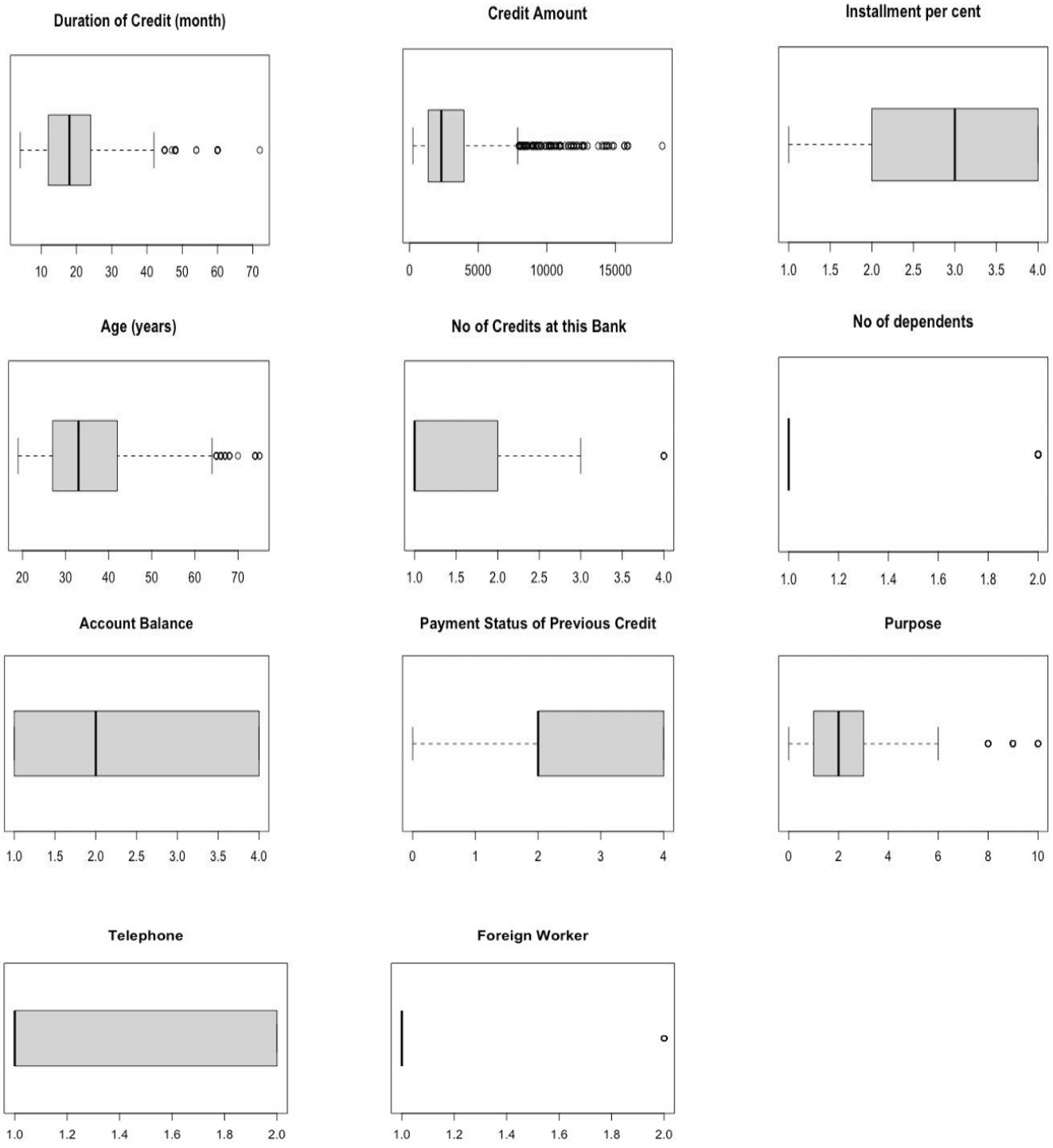

Figure 2: Boxplots of all the numerical attributes in the dataset

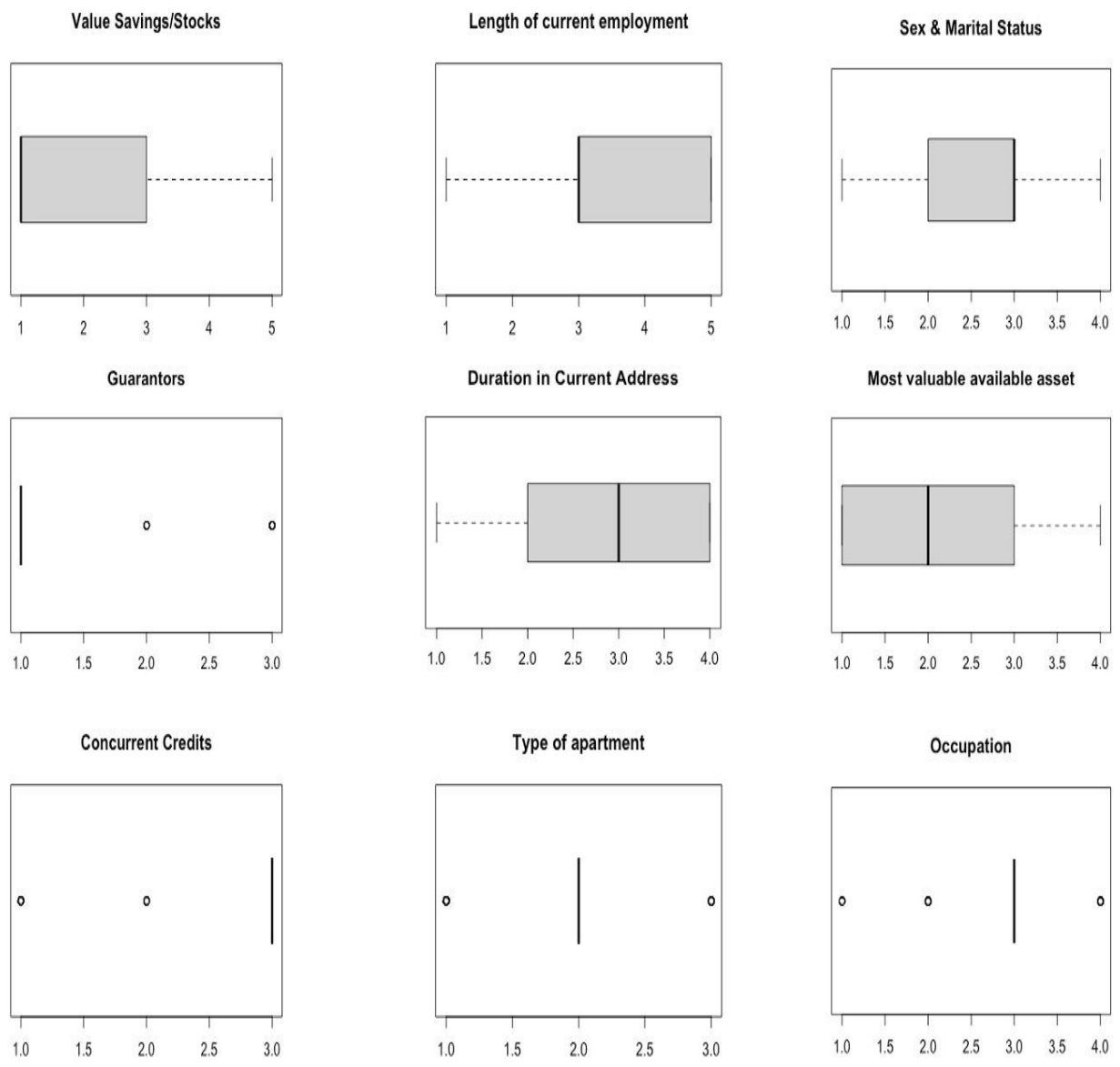

Figure 2: Boxplots of all the numerical attributes in the dataset

As seen from the boxplot diagrams in Figure 2, and in particular, the boxplot diagrams of the numerical attributes, there were low percentages of outliers and as such, the values were left in place.

- *Distribution of numeric attributes*

Using R, the histogram of all numeric attributes were modelled and their distribution were analyzed.

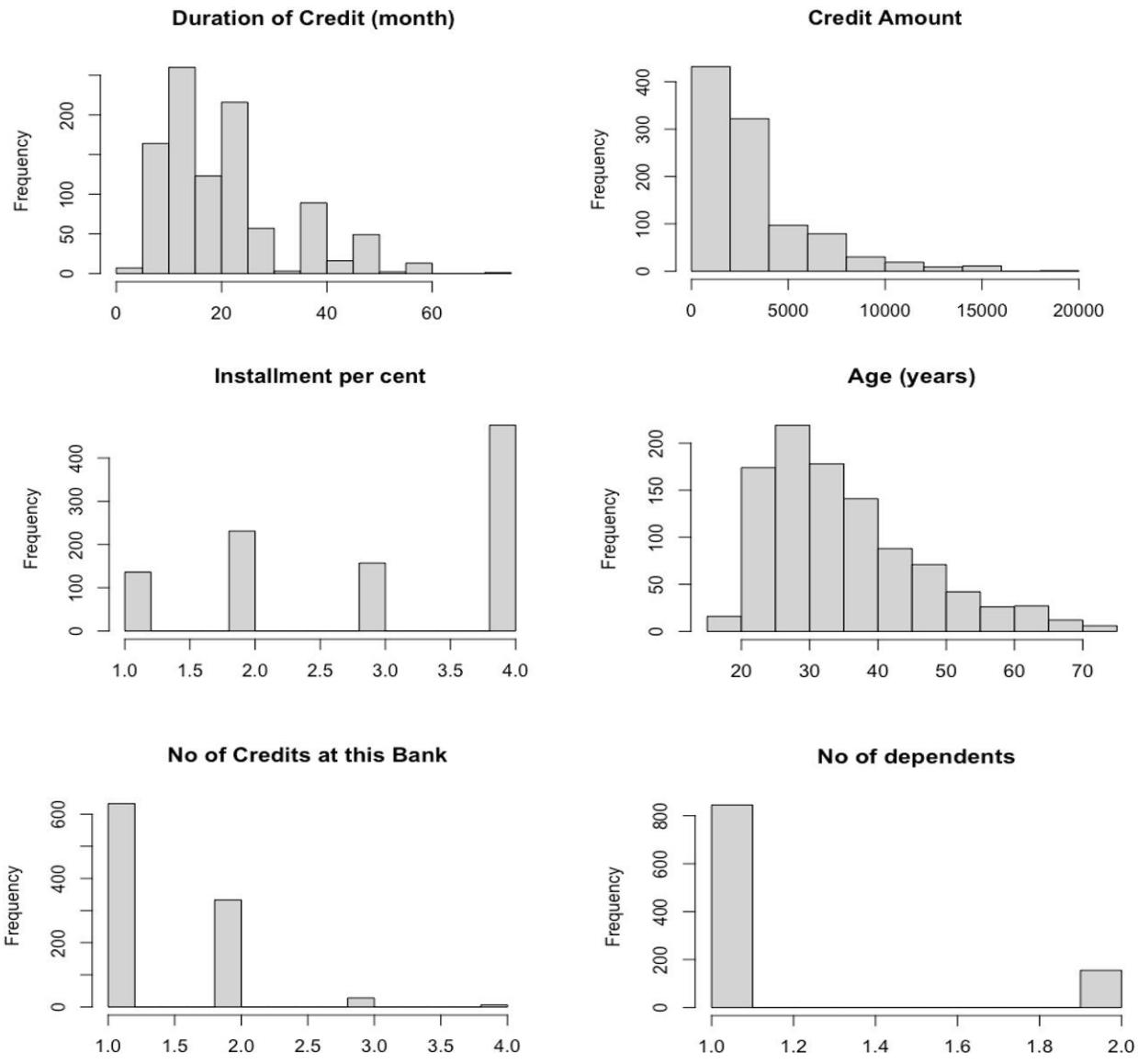

Figure 3: Histogram distributions of the numerical attributes in the dataset

As seen from Figure 3, the histogram distributions of “Installment per cent”, “No of Credits at this Bank” and “No of dependents” are non-normal. Also, the distributions of “Duration of Credit (month)”, “Credit Amount” and “Age (years)” are right skewed. It can be seen that most attributes can influence the class attribute (Creditability). However, the histograms cannot be solely used to determine this. Additionally, by looking at the two right skewed histograms, Duration of Credit (month)”, “Credit Amount” and “Age (years)”, it can be inferred that the frequency is exponentially decreasing.

- *Correlation of attributes in relation to the class attribute*

The ggcorrplot Package in R was used to visualize the correlation matrix to see which attributes seem to be correlated.

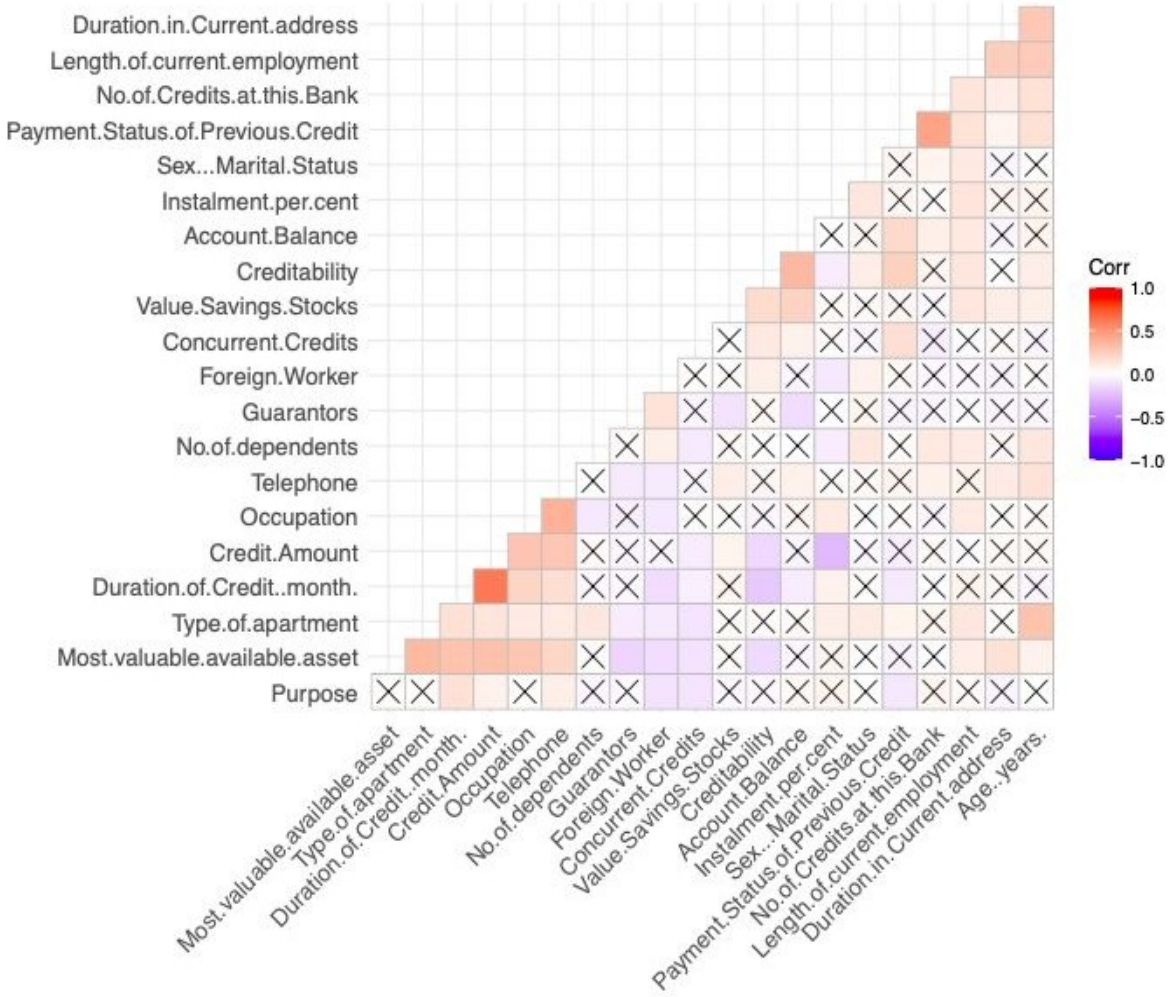

Figure 4: Visualization of the correlation matrix of all the attributes in the dataset

As seen from Figure 4, a blank with a cross is shown where there is no significant correlation. The darker the orange color (positive correlation) of the square or darker the purple color (negative correlation) of the square, the more significant the correlation between the attributes. In alignment with this concept, it can be deduced that “Duration of Credit Month” and “Credit Amount” have the greatest positive correlation, while “Credit Amount” and “Installment per cent” have the greatest negative correlation. Significant correlation can also be observed between “Payment Status of Previous Credit” and “No of Credits at this Bank”, as well as between “Type of apartment” and “Age”.

Most importantly, the attribute most linked to the class attribute (Creditability), is shown to be “Account Balance”. Another attribute with significant correlation to the class attribute (Creditability) is “Payment Status of Previous Credit”.

To further confirm the correlation between attributes, Weka was used. The Chi-square/Correlation Method test was utilized to rank order the attributes according to correlation, with the most correlated being at the top of the rank.

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
111.714 +- 6.025      1   +- 0        2 Account Balance
55.838 +- 3.599       2   +- 0        4 Payment Status of Previous Credit
37.093 +- 6.289       3.7 +- 1        3 Duration of Credit (month)
32.855 +- 3.986       4.1 +- 0.7      7 Value Savings/Stocks
30.862 +- 1.926       4.6 +- 0.8      5 Purpose
24.281 +- 9.161       7.2 +- 4.09     6 Credit Amount
21.644 +- 3.2         7.2 +- 0.87    13 Most valuable available asset
17.194 +- 4.503       8.5 +- 1.5      8 Length of current employment
17.105 +- 2.837       8.8 +- 0.87    16 Type of apartment
11.769 +- 1.962      10.8 +- 0.87    15 Concurrent Credits
 9.131 +- 2.204      11.6 +- 1.11    10 Sex & Marital Status
11.343 +- 7.687      11.6 +- 4.9     14 Age (years)
 6.133 +- 1.476      12.9 +- 0.54    11 Guarantors
 6.12  +- 1.196      13   +- 0.89    21 Foreign Worker
 1.956 +- 0.541      15   +- 0.63    18 Occupation
 1.237 +- 0.45       15.8 +- 0.87    20 Telephone
 1.015 +- 0.636      16   +- 1       12 Duration in Current address
 0     +- 0          17.6 +- 0.49    19 No of dependents
 0     +- 0          18.6 +- 0.49     9 Instalment per cent
 0     +- 0          20   +- 0       17 No of Credits at this Bank
```

Figure 5: Rank order of the attributes in the dataset in relation to correlation to the class attribute

Using the “ChiSquaredAttribute” Evaluator in Weka, the attributes have been rank ordered and a merit weight was given for each as shown in Figure 5. The attribute most linked to the class attribute is confirmed to be “Account Balance”, followed by “Payment Status of Previous Credit”, and then “Duration of Credit (month), followed by “Value Savings/Stocks” and “Purpose”. These attributes can be said to be the top five with regards to correlation ranking.

These attributes were also initially predicted to have a strong correlation to the class attribute (Creditability) in Figure 4. As such, this shows that the top five attributes shown in Figure 5, are very useful in the prediction of a loan applicant’s creditability that would in turn determine if their loan application will be approved by a bank.

Predictive Modeling (Classification)

Model	TP	TN	FP	FN	Accuracy	Precision	Recall	Specificity	Mis-Class	Total FP+FN
Decision Tree (baseline)	660	115	185	40	77.5%	78.1%	94.3%	38.3%	22.5%	225
Decision Tree (validation)	646	125	175	54	77.1%	78.7%	92.3%	41.7%	22.9%	229
Naïve Bayes (baseline)	550	202	150	98	75.2%	78.6%	84.9%	67.3%	24.8%	248
Naïve Bayes (validation)	536	214	142	108	75.0%	79.1%	83.2%	71.3%	25.0%	250
Naïve Bayes training split 85	496	115	104	135	71.9%	82.7%	78.6%	38.3%	28.1%	239
Naïve Bayes test split 15	87	30	13	20	78.0%	87.0%	81.3%	10.0%	22.0%	33
Total	**583**	**145**	**117**	**155**	**72.8%**	**83.3%**	**79.0%**	**48.3%**	**27.2%**	**272**
Naïve Bayes training split 75	434	120	95	101	73.9%	82.0%	81.1%	40.0%	26.1%	196
Naïve Bayes test split 25	143	42	28	37	74.0%	83.6%	79.4%	14.0%	26.0%	65
Total	**577**	**162**	**123**	**138**	**73.9%**	**82.4%**	**80.7%**	**54.0%**	**26.1%**	**261**
Naïve Bayes training split 65	382	102	68	98	74.5%	84.9%	79.6%	34.0%	25.5%	166

Naïve Bayes test split 35	201	43	49	57	69.7%	80.4%	77.9%	14.3%	30.3%	106
Total	**583**	**145**	**117**	**155**	**72.8%**	**83.3%**	**79.0%**	**48.3%**	**27.2%**	**272**
Naïve Bayes training split 55	312	85	75	78	72.2%	80.6%	80.0%	28.3%	27.8%	153
Naïve Bayes test split 45	271	60	42	77	73.6%	86.6%	77.9%	20.0%	26.4%	119
Total	**583**	**145**	**117**	**155**	**72.8%**	**83.3%**	**79.0%**	**48.3%**	**27.2%**	**272**

Figure 6: Classification Model Parameters

Baseline Set – All variables used in the model

Validation Set - Only 7 variables used based on Part One correlation analysis – account balance, payment status of previous credit, duration of credit, value savings and stocks, length of current employment, concurrent credits and purpose.

Training and Test Sets – 4 splits (85/15, 75/25, 65/35, 55/45)

Post-Predictive Analysis

The 4 contributing variables for the baseline model is ranked in the order of importance:

1. Account balance 2. Age 3. Purpose 4. Duration of Credit

The 5 contributing variables for the validation model is ranked in the order of importance:

1. Account balance 2. Duration of Credit 3. Value Savings/Stocks 4. Status of Previous Credit 5. Purpose

Concurrent credits and Length of employment was not listed on the output as they ranked lower.

Decision Tree Model

There is very little difference in the accuracy of both the baseline and validation sets. Both models were able to correctly predict about 77% of the true positive and negative outcomes. This is also true for precision with baseline at 78.1% vs validation at 78.7%. The number of observations that both models predicted correctly is very similar as well. Both models also predicted a fairly high rate of false positives compared to the entire population (baseline 185/1000=18.5% vs validation 175/1000=17.5%). The cost of a false positive (someone who has bad credit but was predicted as having good credit) is higher than the cost of a false negative (someone who has good credit but was predicted as having bad credit). This is because the bank would issue loans to these individuals and may not get paid back on time or in full since these individuals have bad creditability. False negatives are lower in the baseline set at 40 vs 54 in the validation set. Although the bank would want this number to be low, the only loss they would be incurring is the lost interest by not issuing the loans to these individuals with good credit but predicted as having bad credit. The cost of losing interest on a loan is far less than the cost of losing the principle plus interest. Recall is high for both models at over 92%, which is good, meaning those that are predicted as creditable are indeed creditable. Specificity is quite low at 38.3% and 41.7%. This means the model is not very accurate in predicting true negatives. The direct result is also having high rates of false positives.

It's no surprise that "account balance" is the single most important attribute in determining if someone is creditability or not. What is surprising is a majority of people who have good credit actually do not have an account at this bank. This shows the importance of other tree nodes such as purpose of the loan (both being an important attribute in both models). Just based on account balance alone would not be sufficient since the majority of individuals have a purpose of buying a new or used car, furniture/equipment, and radio/tv which could cost well over $200 DM.

Although account balance is the top attribute being split from the Decision Tree models as well as the Naïve Bayes models. Observations 3 and 4 were grouped together for the split but they actually represent very different things. Observation 3 represents an account balance having the most amount of money at >$200 DM whereas observation 4 represents someone who doesn't have an account with this

bank. That does not mean they don't have an account elsewhere, but there's not enough information to determine their creditability.

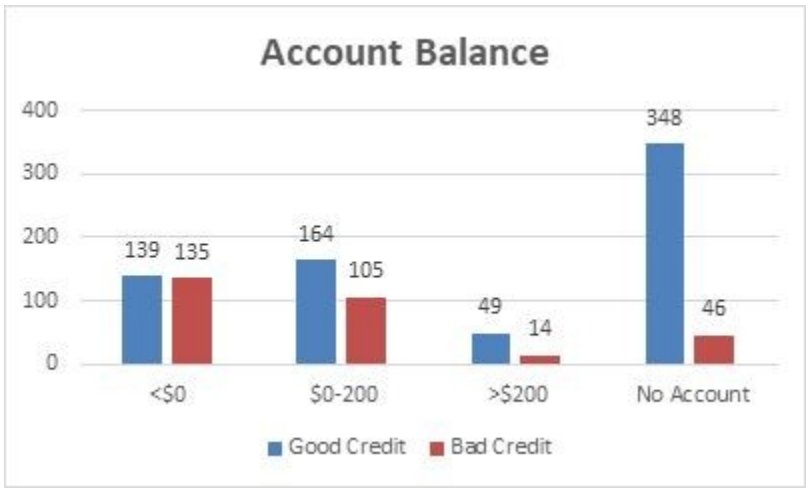

Figure 7: Bar Chart Depicting Account Balance

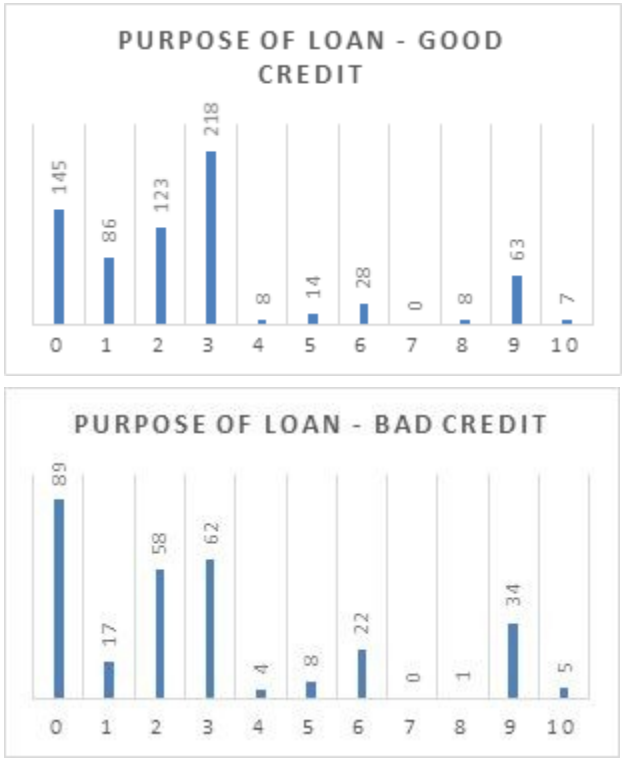

Figure 8: Bar Chart - Purpose of Loan

One interesting but not surprising area is the attribute "age" is second in importance in the baseline model. One would expect age to be a significant factor in determining whether an individual has good or bad credibility, given someone who's younger and less established would have less account balance, less savings, less assets and more outstanding credits such as student or credit card loans. But one can also

argue that someone who's older may have been married and have dependents, and also loans such as a mortgage making them less credible.

The test set predicted very similar results without age as an attribute. This does not lessen the importance of age as an attribute, but indicates that there are other attributes such as "value of savings/stocks" and "status of previous credit" that may be correlated to age.

An ideal situation would be the model predicting the least amount of false positives and the least amount of false negatives to maximize credit income through loans. But there seems to be a tradeoff as the training set had more false positives but less false negatives whereas the test set had more false negatives but less false positives. Overall the test set would be favoured slightly over the training set because the cost of false positives is higher than the cost of false negatives.

The Naïve Bayes model assumes that all the variables are independent of each other, and it resulted in slightly less accuracy than the Decision Tree model. Precision is about the same. Recall is much lower in the 83% range.

This indicates that one or more variables are correlated or dependent with one another. With 20 attributes it is likely that some variables are correlated with each other. As described earlier, someone in their 20s and early 30s will most likely not be married, and therefore have no dependants, and will likely have less account balance, have a longer payment status of a previous credit, will be employed at a shorter period of time, less skilled, have less assets, savings or investments. Similarly, someone in their late 30s and beyond will most likely reflect the opposite situation. Although this may not be true for every individual in the dataset or the real world, this assumption can hold true for the majority of the sample size or population. Below is a graph showing the breakdown in credibility based on those individuals that are 30 years of age or lower and those that are greater than 30 years old. Although a similar amount of data resulted from both scenarios 148 vs 153 individuals with bad credit, but the percentage is much higher at 36.2% for those that are less than or equal to 30 years old vs 25.9% for those greater than 30 years old.

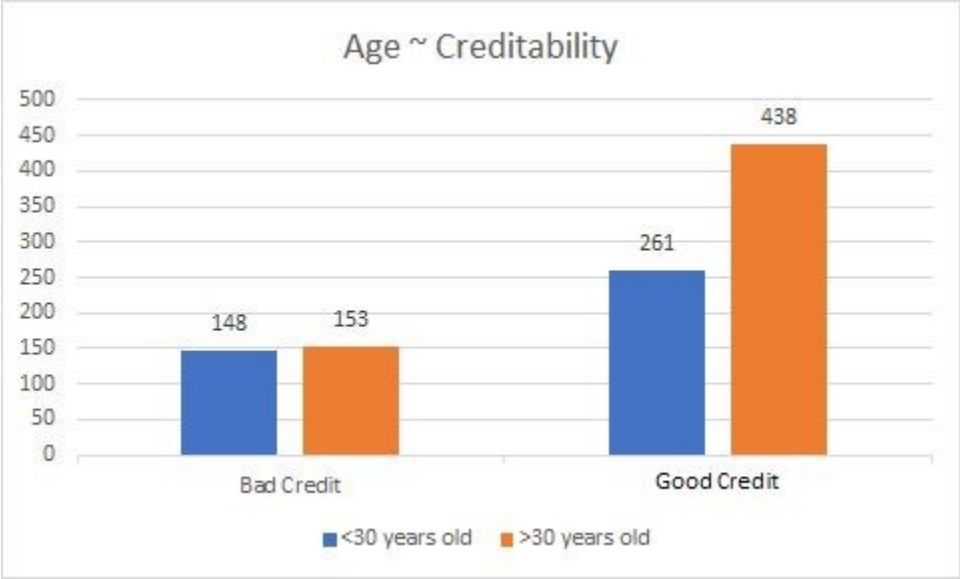

Figure 9: Bar Chart (Age ~ Creditability*)*

The Decision Tree model predicted 646/700=92.3% individuals with good credit as having good credit whereas Naïve Bayes predicted 536/700=76.6%. That is over 100 people (over 15%) that the bank would have lost in interest revenue. This amount would be exponential if it's a loan with an amortization period of 20 or 30 years.

But on the other hand, the Decision Tree model was only able to predict 125/300=41.7% individuals that have bad credit as having bad credit whereas Naïve Bayes predicted 214/300=71.3%. The accuracy of the decision tree model on true negatives is far from desirable. Chances are if they're not being predicted as true negatives, they may be predicted as false positives, thus costing the bank in both the short term and long term.

The Decision Tree model predicted 175 individuals with bad credit as having good credit whereas the Naïve Bayes model predicted 142. This is a cost to the bank as these individuals are mistakenly being issued loans to but they may not be able to pay it back on time or in full. The predictions are fairly similar for both models.

The Decision Tree model predicted 54 individuals with good credit as having bad credit whereas the Naïve Bayes model predicted 108 (double). This is also a cost to the bank in terms of lost interest.

If we were to add up both false positives and false negatives the Decision Tree model would be 229 and the Naïve Bayes model would be 250. A lower number would be ideal as the Naïve Bayes model predicted a quarter of the people incorrectly.

Conclusions and Recommendations

Correctly predicting the creditability of customers is an integral part of a bank's business and success. If a large majority of a bank's customer base is unable to pay their loans back, they'd be out of business. The models being used are by no means the be all and end all to providing loans to customers. This is only the first step. It is recommended that the bank collect more data, such as individual income or household income to further assist them in making these decisions. This would especially be beneficial for larger sums and for longer term loans. The bank should also use different classification models and compare the results. This should be updated periodically as some information (attributes) is real time and is subject to change constantly, thus affecting the prediction results.

From the data preparation stage, it was known that 5 variables were the most correlated with the class variable "creditability". These five variables are Account Balance", "Payment Status of Previous Credit", "Duration of Credit (month), "Value Savings/Stocks" and "Purpose,. It is not surprising to see many of these above mentioned attributes as having the most impact on someone's creditability. These take into consideration the current status of someone's financial health as well as looking into their past history which are both important factors in determining if someone is creditable or not.

In the Decision Tree model, the baseline model splits all the variables and rated 4 variables as the most important predictors of creditability (account balance, age, purpose and duration of credit). The validation model used 7 variables (mentioned above). The results were very similar with neither model overshadowing the other in terms of accuracy and precision. This indicates that the model knows what attributes are not important when determining their prediction. These include installment percent, sex and marital status, guarantors, duration in current address, type of apartment, occupation, number of dependents, telephone and foreign worker. This is to be expected as they offer little to no correlation to the class label.

The Naïve Bayes models used the same criteria to split the data. The results were quite similar to the Decision Tree model. Accuracy is slightly compromised with only a 2% difference. Precision is slightly higher for both the validation, training and test sets. Recall was much lower and specificity is much higher. Interestingly, these two metrics seem to show a bit of a tradeoff.

Splitting the data is a viable option however it does not yield better results especially when the split keeps getting closer to one another as shown in the 65/35 and 55/45 splits.

In conclusion, neither model is perfect at predicting the creditability of customers but does a good job at filtering through some of the important attributes associated with determining if someone is credible or not. The dataset contains a 70/30 split with 70% of the people having good credit and 30% having bad credit. The range of accuracy for both models falls between 74-78% which is a fairly close range for acceptance. However this does not paint the entire picture, as the true positives are much higher in the Decision Tree model. Although recall is much higher in the Decision Tree model, in this particular case, it

is not desired as it is not a life and death prediction such as predicting cancer in patients. What is needed is a higher precision thus resulting in less false positives. This is achieved through the Naïve Bayes models. A bank is a risk averse organization, so they would not risk borrowing money to people who potentially could not pay them back. The money can be used in other investments which would yield higher ROI. It is concluded that the Naïve Bayes validation model would yield the ideal results as every statistic is more balanced. It also has the highest specificity which means those with bad credit are being predicted as so, thus resulting in less false positives, which is a high cost to the bank.