MACQUARIE
University
SYDNEY·AUSTRALIA

BUSA3020 Advanced Analytics Techniques

Assignment #4: Group Project
German Credit Risk Evaluation

GROUP MEMBERS	STUDENT ID
Alisha Sestak	45159823
Chris Jerylle Vargas Oidem	45476624
Nguyen Minh Hanh Do	45524866
Nicole El-Aswad	45205302

Due Date: Wednesday, 10 June 2020, 11.55pm

Word Count: 1,499 Words

Table of Contents

Executive Summary

The report presents an analysis and evaluation of the loan approval system to the management. This will help them decide whether to go ahead with a loan approval or not. Univariate, bivariate, and multivariate analysis were conducted through Microsoft Excel, Orange Data Miner, and RStudio.

It was found that there were no missing values but there was skewed data. Log transformation was then applied for variables “Age”, “Credit Amount” and “Duration” to make the data more normally distributed.

The use of algorithms illustrated that Random Forest was the best predictor to use and provided the most significant result based on metric scores and the Receiver Operating Characteristics (ROC) curve with pre-processed data. It is also recommended to evaluate the data using a single predictive model with 77.3% accuracy.

Furthermore, findings outlined that there were major clusters within the market.

The following risks should be considered and undertaken by the bank to efficiently maximise profit/ reach optimal risk tradeoff:

- Good credibility is associated with shorter credit durations in comparison to bad credibility.
- The higher the credit borrowed, the more likely it is that the applicant is bad.
- Older applicants have better credibility than younger applicants.

1.0 Background Information

Upon receiving loan applications banks need to decide whether or not to approve the applicant based on various demographic and socio-economic criteria including in the application. Making this approval decision comes with two inherent risks: rejecting customers that would have been likely to repay the loan, and approving customers likely to default. This second risk is five times more costly to a bank.

2.0 Situation Review

A dataset with 1,000 loan applicants has been provided. The dataset contains 3 numerical and 17 categorical variables and classifies applications as having either good or bad credibility. This data has been used throughout this study to try and generate an analytical model and identify distinct market segments to improve the accuracy of applicants' credibility. Furthermore, the condition “it is worse to class a customer as good when they are bad than it is to class a customer as bad when they are good” will also be explored to ascertain the tradeoffs required to minimise risk of loss.

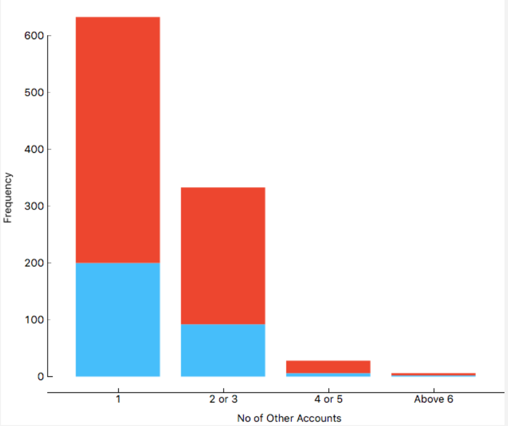
2.1 Data Handling

Prior to applying predictive algorithms, exploratory analysis was conducted and the structure of each variable examined to ensure that the data was suitable and the results of any investigation would be reliable.

Variable Selection

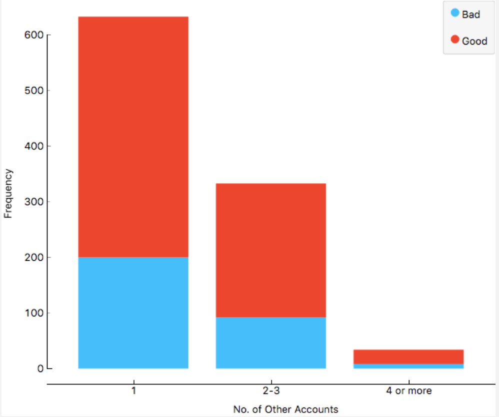
Throughout our preliminary analysis it was apparent that certain variables/variable levels were redundant. An example of this and the corrective treatment applied is shown below:

Original Variable



The above diagram depicts the original distribution of ‘No. of Other Accounts at this Bank’. The level “Above 6” accounts for less than 1% of applications and therefore will not have a significant impact on model fit.

Treatment



Logical thinking, and the similar probability distribution for credibility between the levels “4 or 5” and “Above 6” indicate that it is appropriate to combine these levels. This will also allow greater simplification of the data without impacting model fit.

Additional variables that exhibited similar characteristics were considered using the same logic and redundant levels removed from the dataset. An overview of these changes is shown below:

Variable	Original Levels	Treatment	Adjusted Levels
Loan Purpose	Furniture/Equipment Business Domestic Appliances Radio/TV Car (new) Car (used) Repairs Education Retraining Other	Equipment & Business combined Appliances & Radio/TV combined Cars (new & used) & Repairs combined Education, Training & Other combined	Business/Equipment Appliances/TV Cars Education/Training/Other
Other Debts	Bank Stores None	Bank & Stores combined	Bank_Stores None
Occupation	Unemployed/Unskilled (non-res) Management/Highly Qualified Unskilled (res.) Skilled/Official	Unemployed/Unskilled (non-res.) & Unskilled (res.) combined	Unemployed/Unskilled Highly Qualified Skilled
Sex / Marital Status	Male (divorced/separated) Female (divorced/married) Male (single) Female(single) Male (married/widowed)	Marital status disregarded as very little additional information added to the data set. All male and female levels grouped appropriately.	Male Female
Credit History	No Credits/All Repaid Duly All Repaid Duly (until now) All Repaid Duly Critical Delay in Paying	All three variables associated with credits being repaid duly combined.	All Credits Repaid Duly Delay in Paying Critical Account

<u>Data Transformation</u>

Once redundancies were removed from the data, the transformation was conducted on both the numerical and categorical variables to remove skewness and outliers, which ensure that the values within the dataset were both logical and appropriate for predictive analysis.

- Numerical Variables

Each of the three numerical variables within the original data set i.e. “Duration”, “Credit Amount” and “Age”, contained outliers and possessed heavily right-skewed distributions:

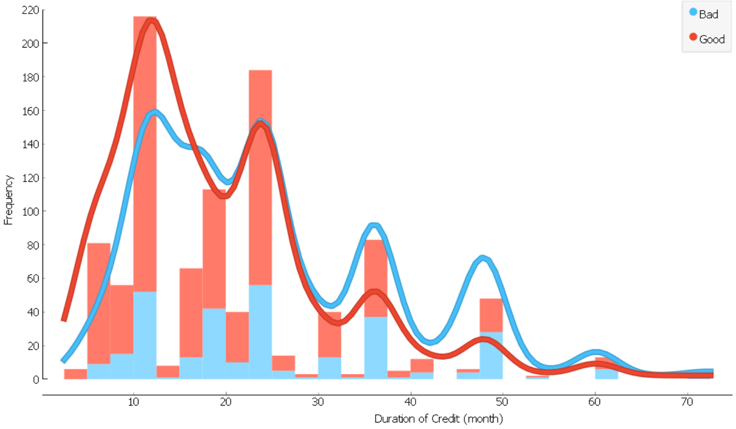


Fig 1. Positively skewed data distribution for Duration of Credit (months)

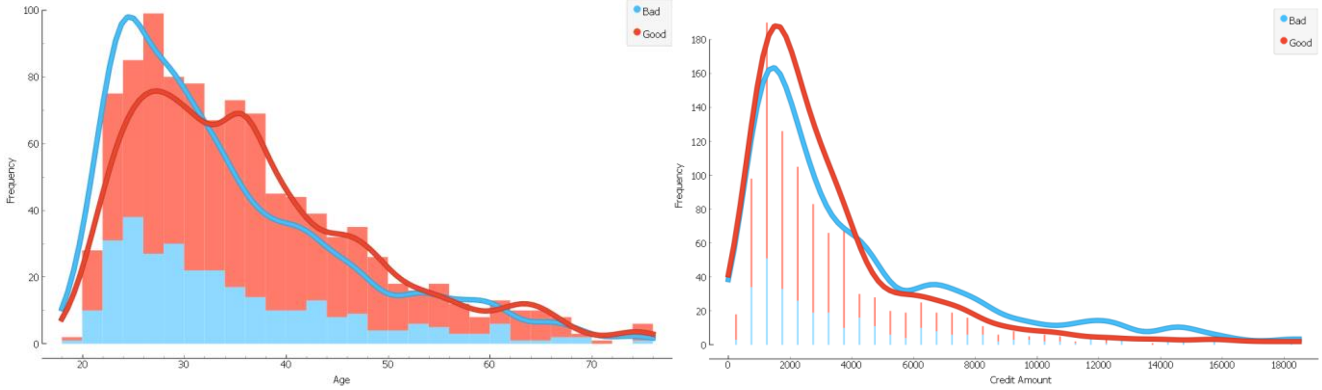


Fig 2 & 3. Positively skewed data distribution for Age (left) and Credit Amount (right)

As a high level of skewness can generate misleading results in statistical tests, the positive skewness exhibited in the above figures is undesirable. To improve results, we transformed the numerical data using a log function to make them more normally distributed and reduce variability.

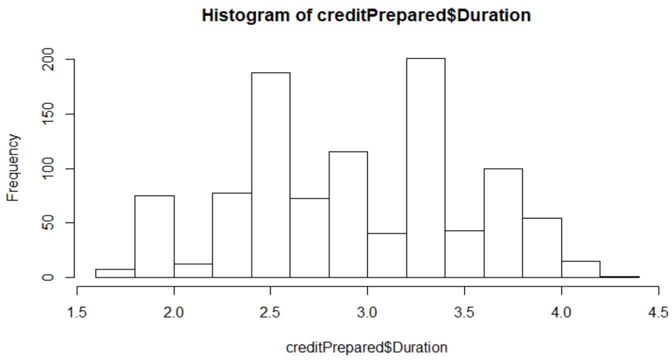
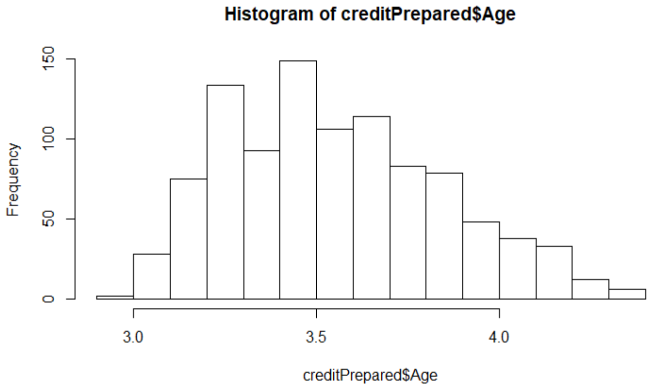


Fig 4. Log transformation of Duration of Credit (months)



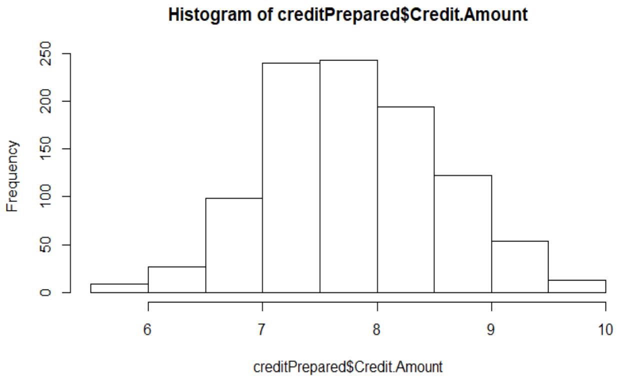


Fig 5 & 6. Log transformation of Age (left) and Credit Amount (right), resulting in a bell-shaped distribution

- Categorical Variables

Similar to numerical variables, certain categorical variables within the dataset were highly skewed. Various treatments were applied and tested throughout our investigation to ensure the data was appropriate for analysis. These include: PCA, manifold learning, and one-hot encoding, and each is explored in greater detail in proceeding sections.

It is also worth noting that the “Foreign Worker” variable initially reported 96% of applicants as foreign and 3% as domestic. However, further research online confirmed that the coding of this variable was incorrect. Therefore, this variable has been re-coded such that the majority of applicants are domestic.

It was also verified that there are no missing values within this data set.

3.0 What is the Optimal Model to Determine Customer Loan Approvals?

3.1 Comparison of Diverse Data Handling Approaches on Predictive Model Performance

To determine which predictive model is most appropriate for determining customer approvals, it is first vital that multiple data handling approaches are tested to identify which provides the best predictive model performance on average.

The three data handling approaches that we tested on the above cleaned data were: no data handling, pre-processing, and data reduction. The process involved in applying these approaches to our data set was as follows:

Approach	Method
No data handling	No action needed
Pre-processing	Normalisation technique applied, and discrete variables one-hot encoded to remove skewness and ensure the reliability of results.
Data Reduction	Manifold learning applied to capture orthogonal components of relatively correlated variables.

Each of these three approaches was then utilised to predict credibility using four different models. The performance of each data handling approach was then measured by finding the average score across the four models for three different performance metrics: F1 score, Precision, and Recall. The results of this analysis are as follows:

Handling Approach	Average Metric Score		
	F1	Precision	Recall
Original Data (No Data Handling)	0.820	0.788	0.857
Pre-Processing	0.822	0.791	0.859
Data Reduction	0.820	0.790	0.855

These results demonstrate that the average performance of predictive models is not significantly different across diverse data handling approaches. Furthermore, as the purpose of data reduction is to minimise variables by aggregating those with high correlations, it is not surprising that this approach has not improved model performance on this dataset (which largely possesses uncorrelated variables). However, pre-processing has produced slightly better performance across all three metrics. Based on this observation, and the improved integrity offered by pre-processing, this data handling approach will be utilised to make comparisons across predictive models.

3.2 Comparison of Predictive Model Performance using Pre-Processed Data

The below charts compare the performance of four diverse predictive models in estimating credibility.

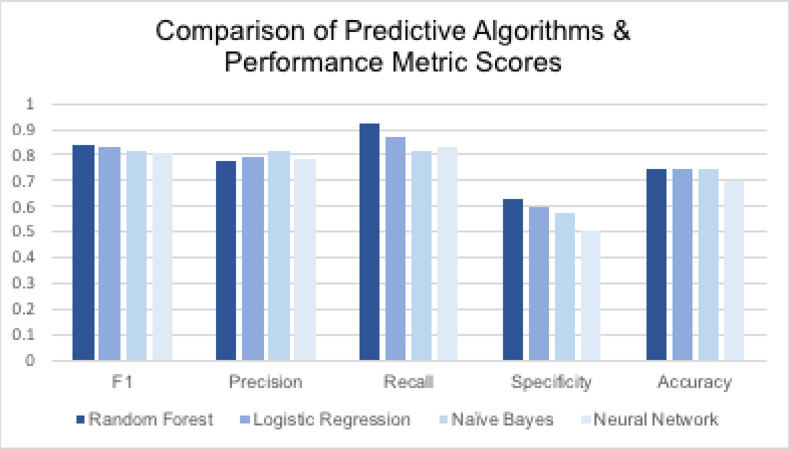


Fig 7. Comparison of Predictive Algorithms

The above graph demonstrates that there are no highly significant differences in performance among models. However, Random Forest seems to perform slightly better than other models on each performance metric (excluding precision).

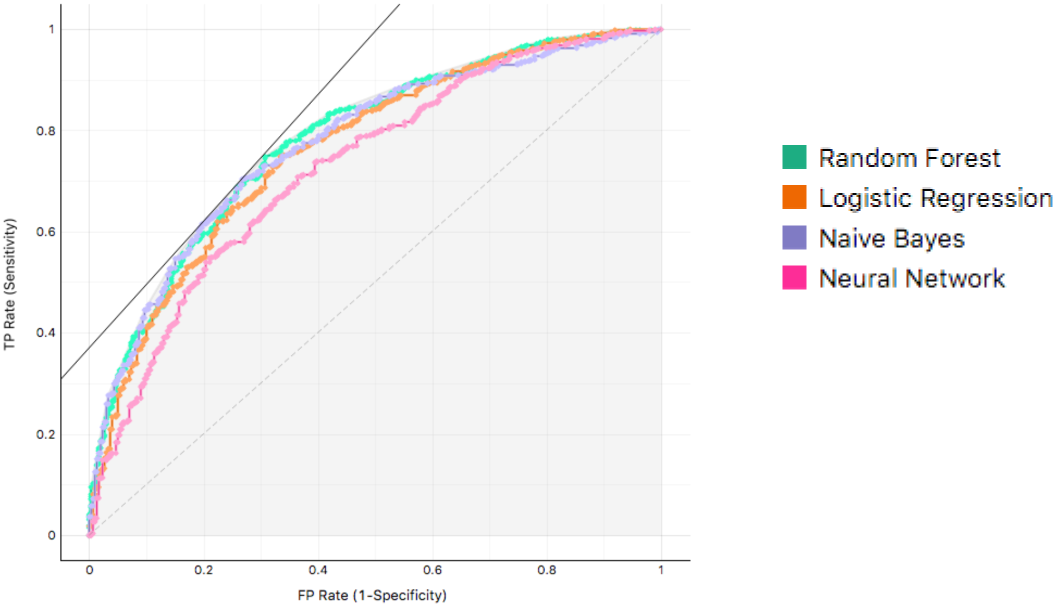


Fig 8. Receiver Operating Characteristics (ROC) Curve

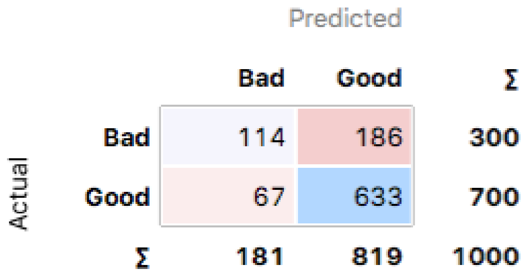
The above ROC graph signifies that the Naive Bayes model performs best at lower levels while a Random Forest triumphs higher along the graph. Based on this observation and the performance scores above, Random Forest is considered the most appropriate model for predicting credibility. It is also worth noting that this curve has been generated using the cost matrix (i.e. FP cost 5 vs FN cost 1). The intersection of the tangent and curve indicates that by using a Random Forest model it is possible to achieve 75% of true positive predictions at a cost of 30% false positives.

3.3 Random Forest Optimisation and Explanation

The Random Forest model selected above was constructed using default parameters. Although these values produced a relatively accurate model, it is worth considering whether any adjustments would improve model performance. Multiple iterations of the model were run with a diverse set of parameters until accuracy was enhanced. The findings of this optimisation are as follows:

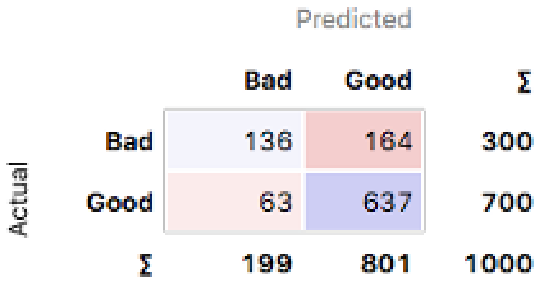
Parameter	Default Value	Optimised Value
No. of Trees	50	47
No. of attributes considered at each split	8	10
Do not split subsets smaller than	3	3

Default Parameters



Actual \ Predicted	Bad	Good	Σ
Bad	114	186	300
Good	67	633	700
Σ	181	819	1000

Optimised Parameters



Actual \ Predicted	Bad	Good	Σ
Bad	136	164	300
Good	63	637	700
Σ	199	801	1000

Fig 9 & 10. Confusion matrices of random forest model before and after parameter optimisation

The accuracy of the Random Forest model has improved from 74.7% with default parameters to 77.3% with optimised values.

The importance of each variable within the optimised model is as follows:

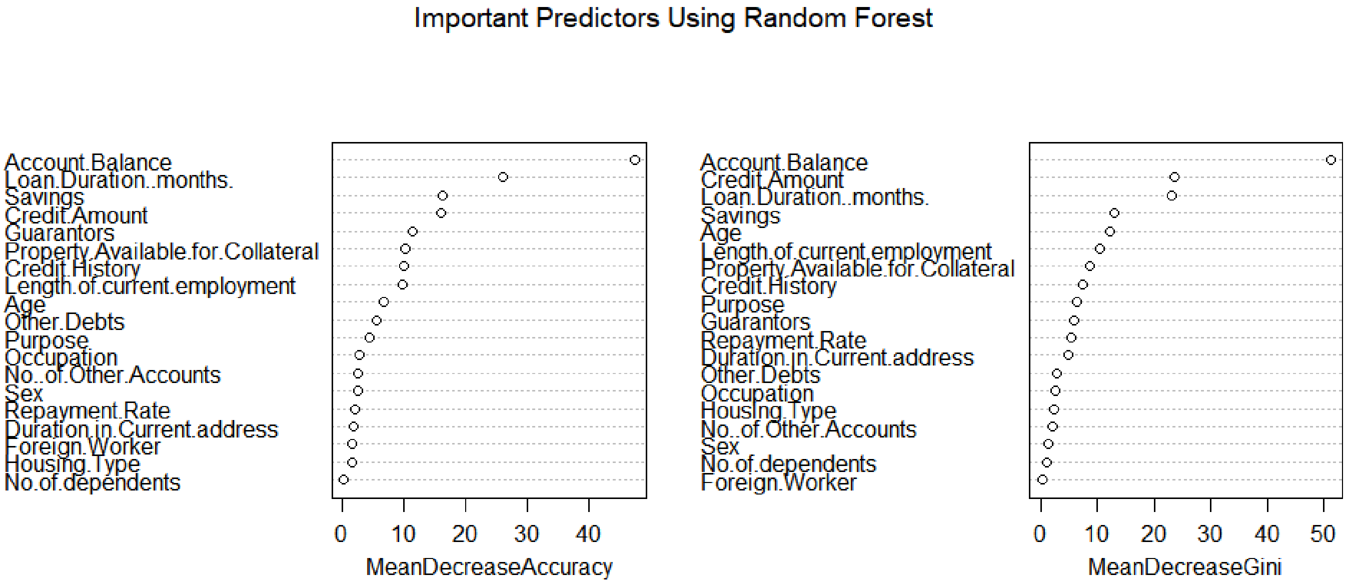


Fig 11. Important Predictors Using Random Forest, Without Cost Matrix

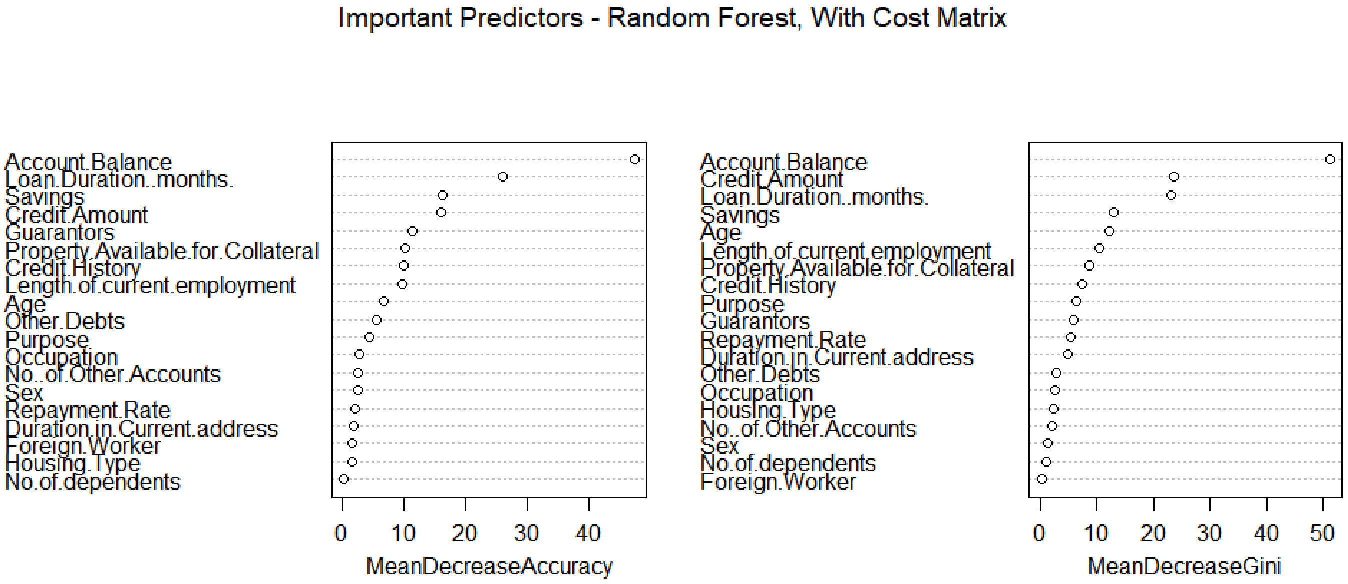


Fig 12. Important Predictors Using Random Forest, With Cost Matrix

<u>Conclusion</u>: In future analysis, the variables “Foreign Worker”, “Housing Type”, and “No. of Dependents” could be removed as they do not contribute significantly to explaining credibility. Therefore, the optimum model is Random Forest with 10 parameters.

4.0 Are There any Distinct Market Segments Among the Customers?

4.1 Method

Data Reduction

The team ran a Principal Component Analysis in Orange Data Miner and found that only 8 dimensions explain 35% of the data. Hence, all variables in the dataset will be used for cluster analysis in RStudio.

After scaling the data, Hierarchical Clustering and K-means Clustering were used to find groups of observations that share similar characteristics and attributes.

Hierarchical Clustering

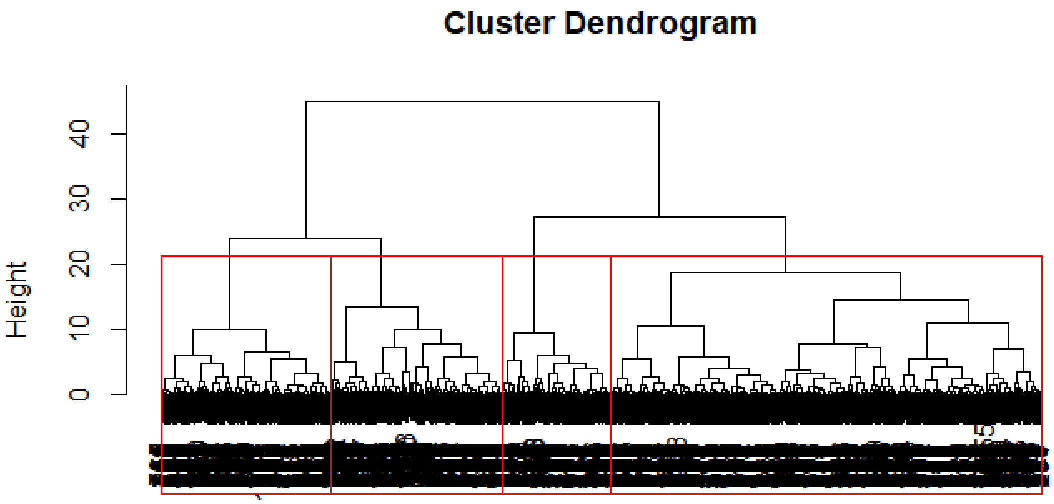
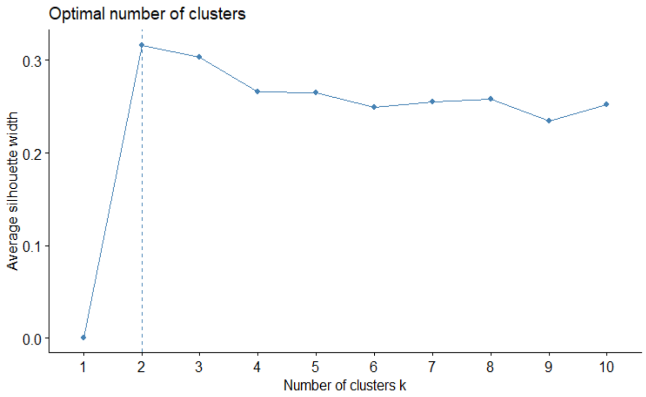


Fig 13. Result of Hierarchical Clustering

The hierarchical clustering resulted in 4 clusters.

K-means Clustering



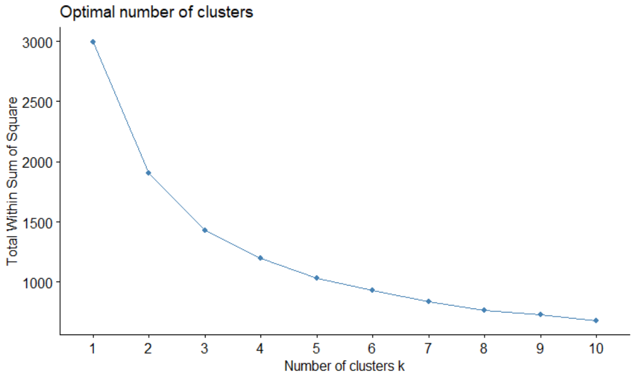


Fig 14. Silhouette

Fig 15. Elbow Method

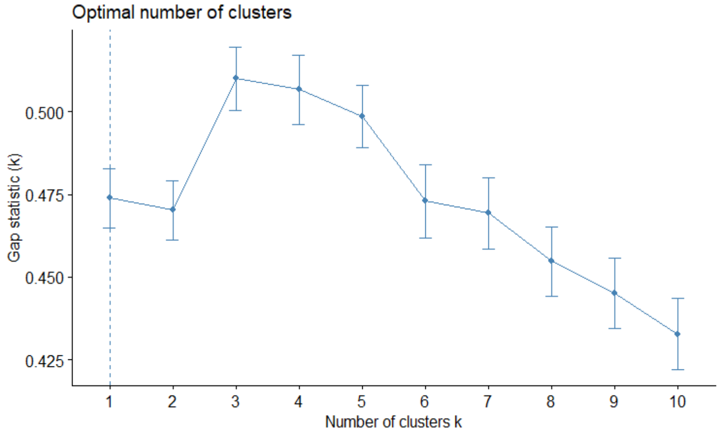


Fig 16. Gap Statistic

From the plots we can see that the best number of clusters for: silhouette is 2, elbow method is 3, gap statistic is 3. As a result, 3 clusters were chosen.

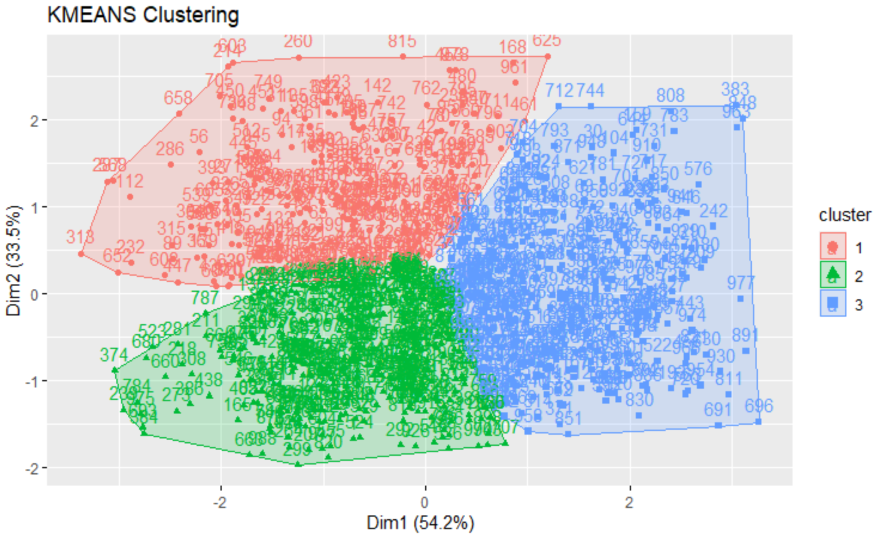


Fig 17. K-means Clustering (3 Clusters)

4.2 Results

The clustering analysis shows that K-means dealt better with the data. It also provided better classification than the Hierarchical Clustering method, which is more understandable for the interpretation.

Market Segmentation

3 market segments were obtained - the first one has 367 objects, 371 in the second, and 262 in the third.

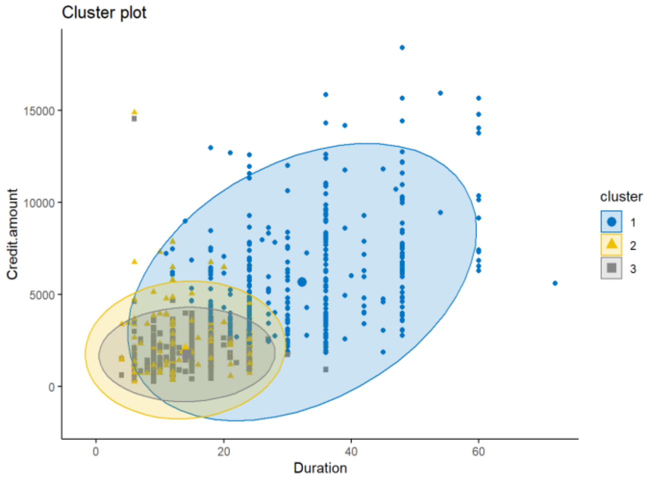


Fig 18. Cluster Plot by Credit Amount & Duration

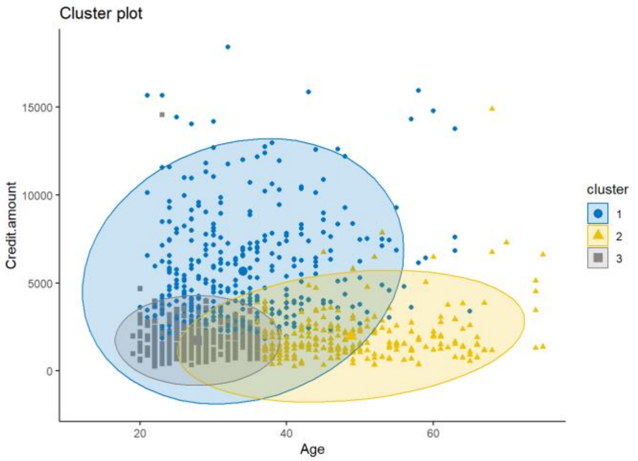


Fig 19. Cluster Plot by Credit Amount & Age

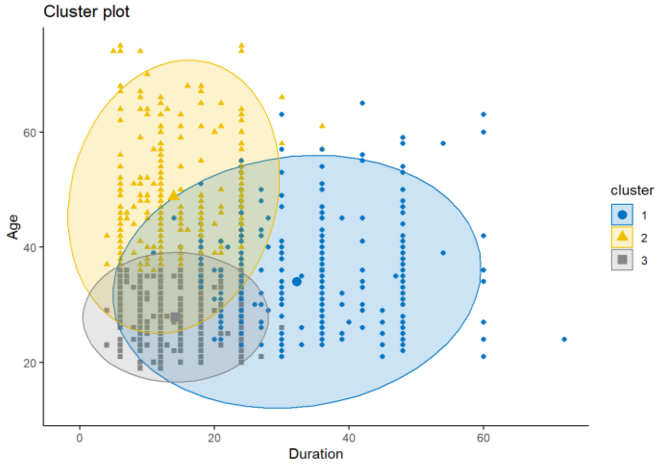


Fig 20. Cluster Plot by Duration & Age

Graphical Representation of Each Cluster Breakdown

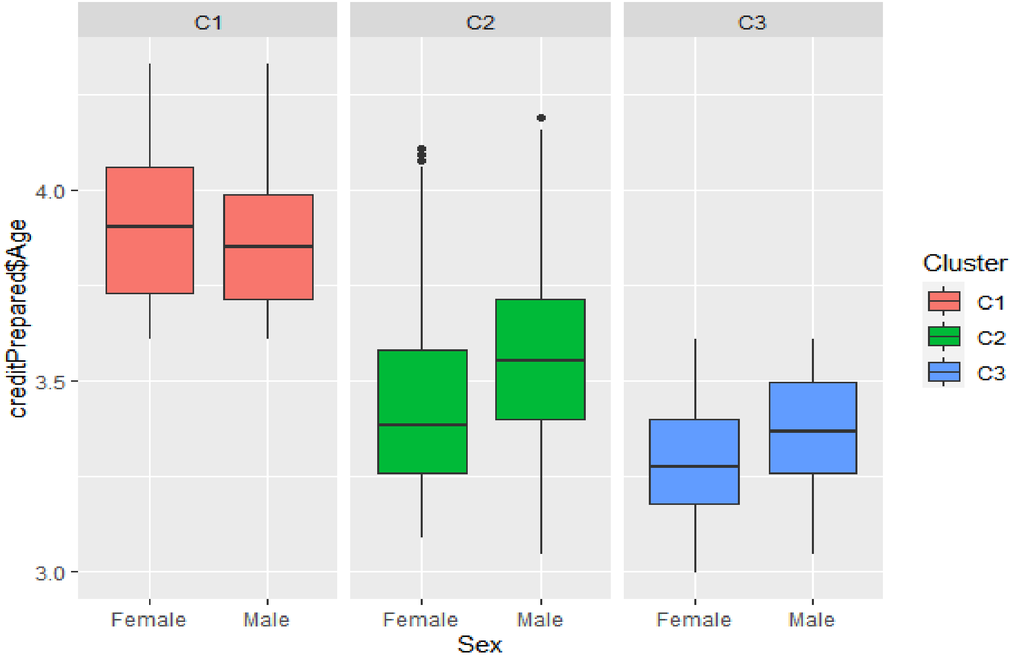


Fig. 21 Box Plot of each cluster by Age and Gender. The females in Cluster 1 are older than the males. In contrast to Clusters 2 and 3, the females are younger, and the males are marginally older.

- By Duration:

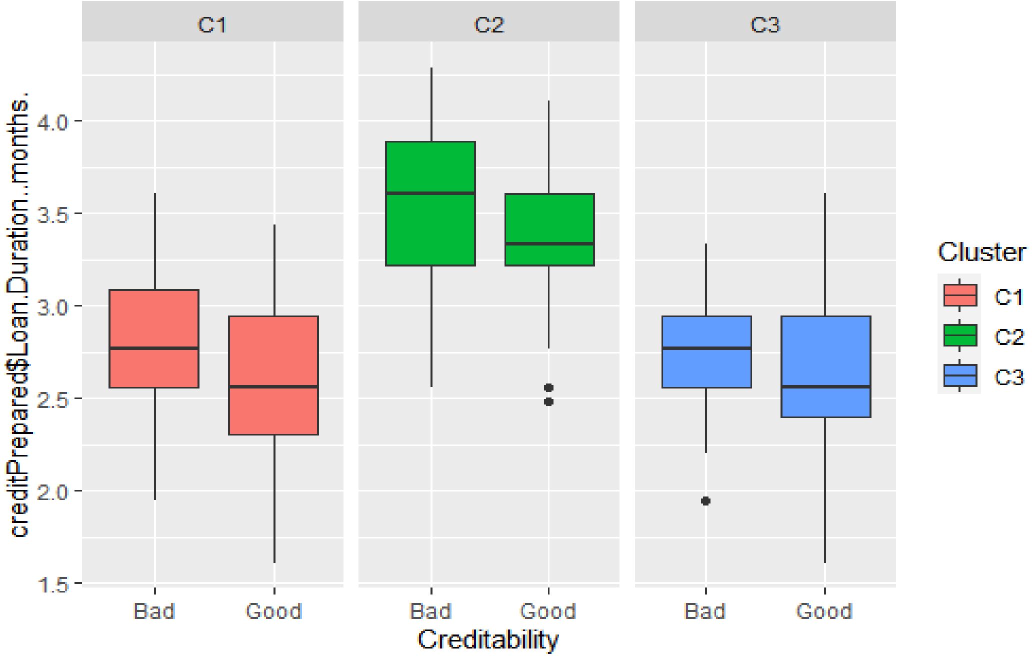


Fig 22. Box Plot representing each cluster by Duration (months) and Credibility.

Results in all 3 clusters consistently show that good credibility is associated with shorter credit durations in comparison to bad credibility. Cluster 2 has the longest credit duration compared to the other clusters.

- By Credit Amount:

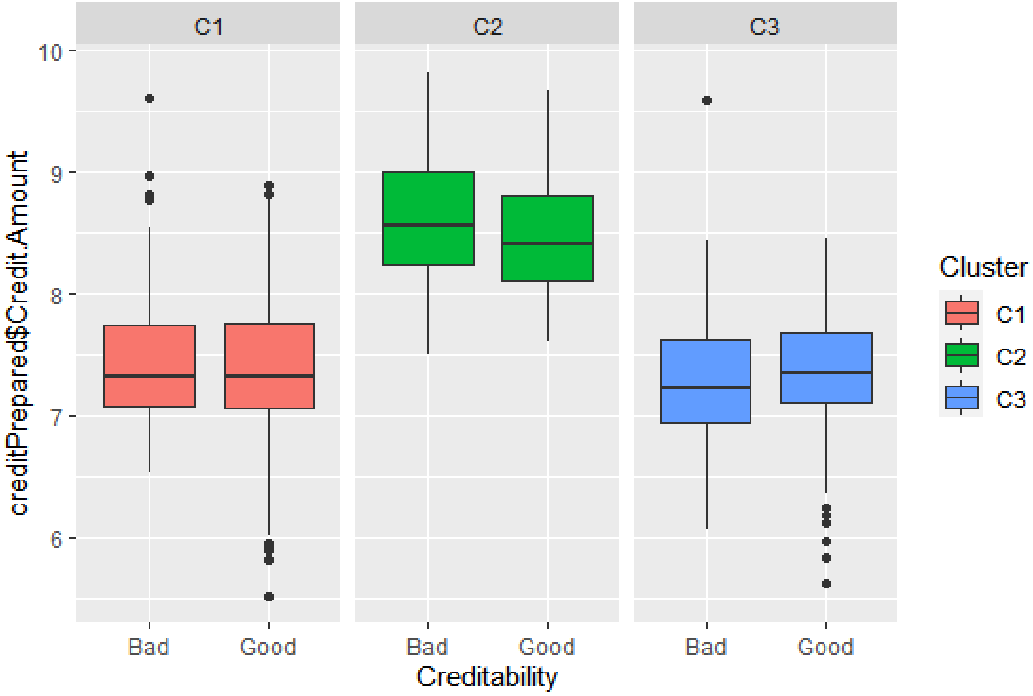


Fig 23. Box Plot representing each cluster by Credit Amount (€) and Credibility.

Cluster 2 has the largest credit amount compared to the other clusters. It seems that the higher the credit borrowed, the more likely it is that the applicant is bad.

- By Age:

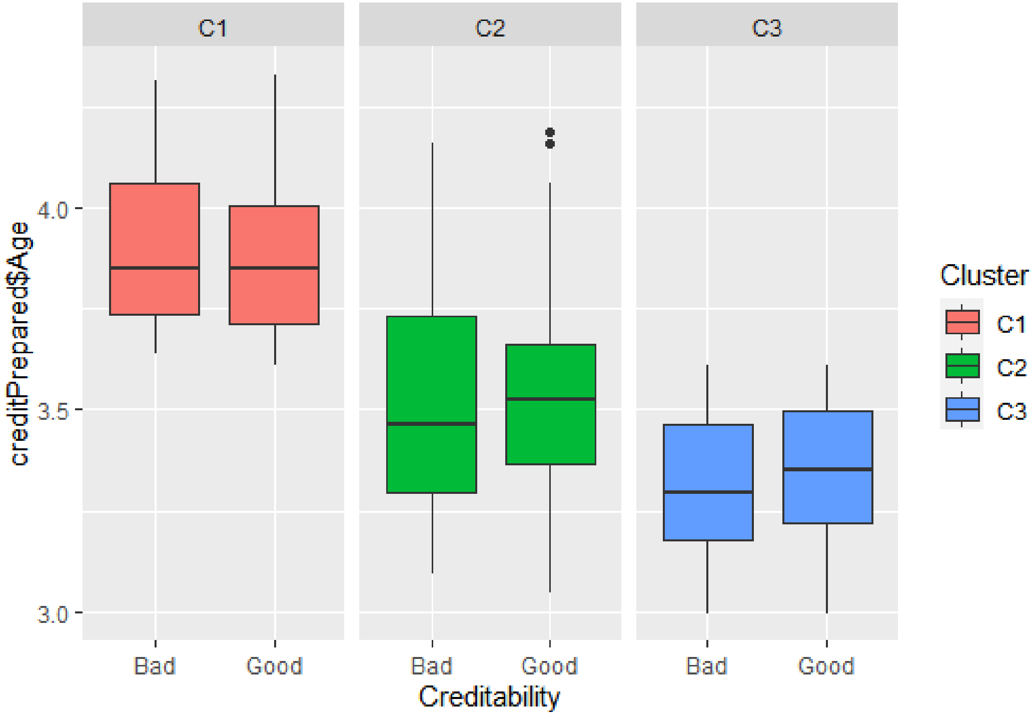


Fig 24. Box Plot representing each cluster by Age (years) and Credibility.

Loan applicants in Cluster 1 are in the middle-aged group as compared to the other two clusters. Cluster 2 are working adults while Cluster 3 are young adults. It seems that older applicants have better credibility than younger applicants in Clusters 2 and 3.

- Mean of each variable:

Cluster	Age (Years)	Credit Amount (€)	Duration (Months)
1	48.300	1,961.184	13.639
2	34.255	5,753.437	32.489
3	27.635	1,771.141	14.727

4.3 Conclusion

The following characteristics in each of the clusters are found:

- Cluster 1: Contains older people with the shortest duration credit for a lower amount of money.
- Cluster 2: Middle-aged working people who borrowed the highest loan for the longest period.
- Cluster 3: Represents young adults who borrowed the smallest amount of money for a rather short credit duration period.

5.0 Does a Predictive Model by Segment Perform Better Than a Single Predictive Model for the Whole Sample?

5.1 Method

Part 3.0 confirmed that Random Forest was the most accurate and optimal model to predict customer credibility. Evaluation results were then cross-validated by Cluster to determine whether it performed better than the full model.

5.2 Results

The Random Forest produced an accuracy of 69.9% with optimised values.

Model	Specificity	Accuracy	F1	Precision	Recall
Random Forest	0.681	0.699	0.804	0.738	0.883

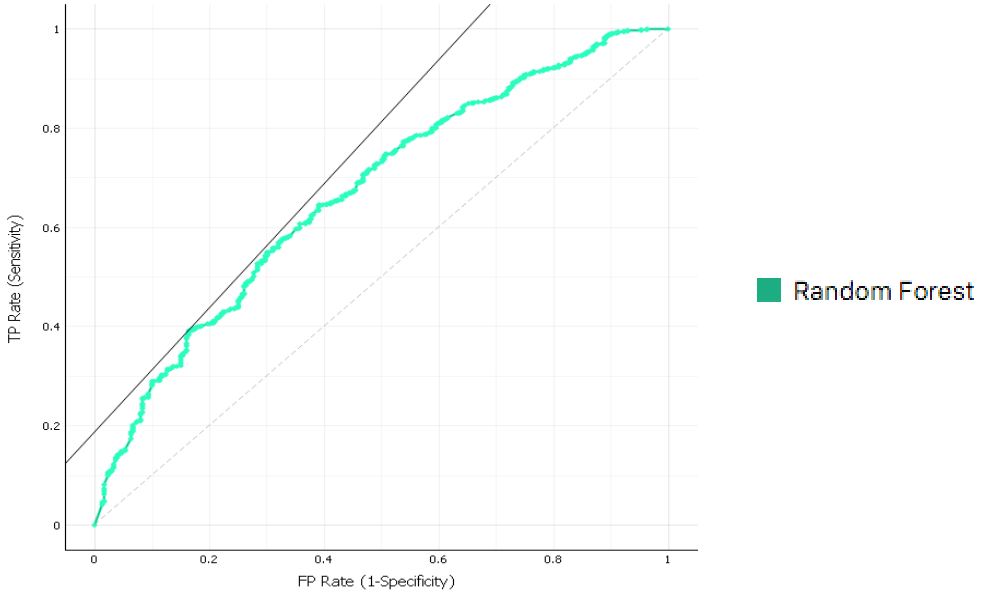
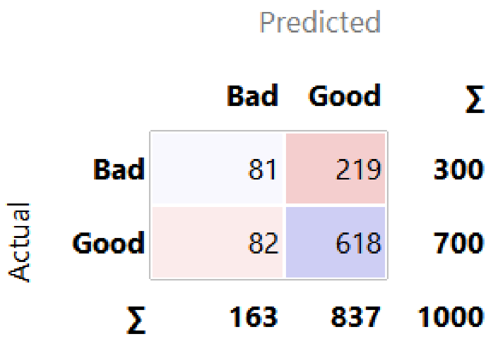


Fig 25. ROC Curve of random forest (by cluster)



		Predicted		
		Bad	Good	∑
Actual	Bad	81	219	300
	Good	82	618	700
	∑	163	837	1000

Fig 26. Confusion matrix of random forest after parameter optimisation

5.3 Conclusion

However, the single predictive model with an accuracy of 77.3% performed better than the model by cluster with only 69.9%.

Therefore, a predictive model by segment does not perform better than the single predictive model.

6.0 Overall Conclusion

Although the K-means clustering method provided us with useful information in distinguishing between customers with three market segments, it is not worth segmenting our database since clustering did not result in a better predictive model.

Therefore, we recommend the single predictive model in Part 3.0 to be used in determining customer loan approvals.

The predictive model showed Random Forest to be the most appropriate and optimal model for predicting credibility, with optimised values producing an accuracy of 77.3%. In addition, it was found that adding a cost matrix does not change the relative importance of predictors. Variables such as "Foreign Worker", "Housing Type", and "No. of Dependents" can be removed for future research as they do not significantly contribute to the model. The four most important criteria for assessing the 'credit-worthiness' of a customer are: Account Balance, Loan Duration, Credit Amount, and Savings.