PREDICTIVE & CLUSTER ANALYSIS ON GERMAN CREDIT DATA

Group 4 - Beatrice Jacinda, Justin Lam, Athena Mosphilis and Malhar Sapre | BUSA3020 | S1 2020

Table of Contents

Background

The German Credit dataset was obtained for this study and contains 20 variables of information about 1000 loan applicants and their creditability classification. This report focuses on performing predictive and cluster analysis on these applications, to assist management on the best way to predict the creditability of an applicant and identify different market segments to provide tailored services to customers. This report will be segmented by the series of steps taken to complete all the requisite analysis.

1.0 Actions taken prior to conducting analysis

Method

Univariate analysis is performed first to review how each variable is accounted for in the data set and if there is any data cleaning and feature engineering needed. Following the univariate analysis and possible cleaning/feature engineering, bivariate analysis is done to determine the association between the variables, and consequently determine whether data reduction should be performed to improve the quality of analysis and interpretation.

Results

Univariate analysis

There was no missing data observed and most of the variables are categorical. Table 1.1 shows a summary of notable categorical variables in the data set in relation to their proportion of responses.

<table>
<tr><th rowspan="2">Creditability (Target variable)</th><th colspan="10">Good</th><th colspan="10">Bad</th></tr>
<tr><td colspan="10">70%</td><td colspan="10">30%</td></tr>
<tr><th rowspan="2">Foreign Worker</th><th colspan="10">Yes</th><th colspan="10">No</th></tr>
<tr><td colspan="10">96%</td><td colspan="10">4%</td></tr>
<tr><th rowspan="2">Sex and Marital Status</th><th colspan="4">Male & Divorced</th><th colspan="4">Female & Divorced/ Married</th><th colspan="4">Male & Single</th><th colspan="4">Male & Married</th><th colspan="4">Female & Single</th></tr>
<tr><td colspan="4">5%</td><td colspan="4">31%</td><td colspan="4">55%</td><td colspan="4">9%</td><td colspan="4">0%</td></tr>
<tr><th rowspan="2">Occupation</th><th colspan="5">Unskilled & Non-resident</th><th colspan="5">Unskilled & Resident</th><th colspan="5">Skilled</th><th colspan="5">Highly Skilled</th></tr>
<tr><td colspan="5">2%</td><td colspan="5">20%</td><td colspan="5">63%</td><td colspan="5">15%</td></tr>
<tr><th rowspan="2">Purpose</th><th colspan="2">Car (new)</th><th colspan="2">Car (used)</th><th colspan="2">Furniture</th><th colspan="2">Radio/ TV</th><th colspan="2">Appliances</th><th colspan="2">Repairs</th><th colspan="2">Education</th><th colspan="2">Retraining</th><th colspan="2">Business</th><th colspan="2">Other</th></tr>
<tr><td colspan="2">24%</td><td colspan="2">10%</td><td colspan="2">18%</td><td colspan="2">28%</td><td colspan="2">1%</td><td colspan="2">2%</td><td colspan="2">5%</td><td colspan="2">1%</td><td colspan="2">10%</td><td colspan="2">1%</td></tr>
</table>

Table 1.11 – Proportion descriptives of key categorical variables

There is a 7:3 ratio of good to bad creditability of respondents in the dataset, which will need to be accounted when performing analysis to prevent bias towards good creditability. Additionally, due to the variables in table 1.1 having few observations within levels of responses, feature engineering needs to be conducted.

Feature engineering

Feature engineering is required to make the dataset more useful for both predictive and clustering analysis. Table 1.21 explains how the numerical variables have been engineered, and table 1.22 presents the categorical variables that have been engineered.

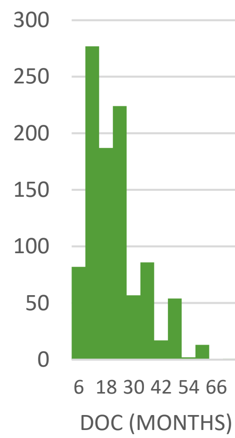
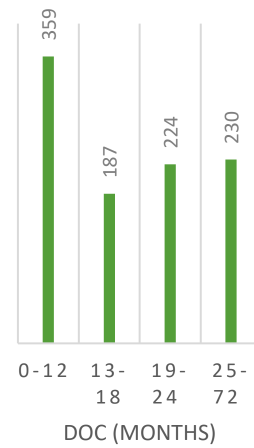
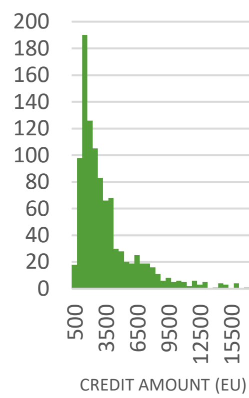
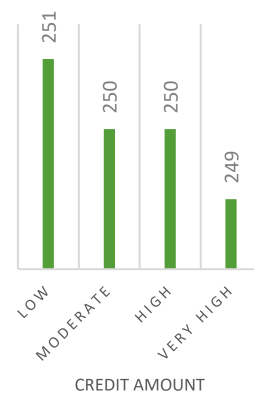
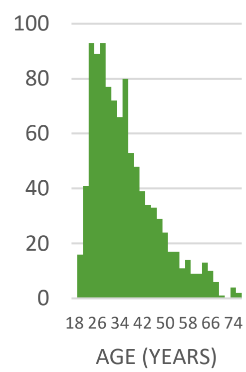
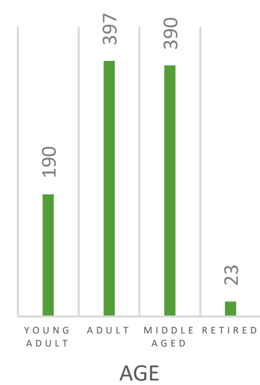
Variable Engineered	New Parameters	Justification	Histogram Before Engineering	Column Chart After Engineering
Duration of Credit (months)	Grouped by the quartile ranges of: 1. 0 - 12 months 2. 13 - 18 months 3. 19 - 24 months 4. 25 – 72 months	Consideration has been made to the different types of credit duration the bank may offer. It also allows us to see which duration period is the most popular by different segments of customers.		
Credit Amount	Grouped by quartile ranges of: 1. Low = 0 - 1366 2. Moderate = 1367 – 2320 3. High = 2321 – 3973 4. Very High = 3974 - 18424	Allows for easier explanation on what level of credit a market segment applies for in cluster analysis.		
Age	Grouped by social classification of ages: • Young Adult = 19 - 25 • Adult = 26 - 35 • Middle Age = 36 - 64 • Retired = 65+	Consideration has been given to the possible different loan approval rates based on the age group an applicant belonged in.		

Table 1.21 – Feature engineering on numerical variables

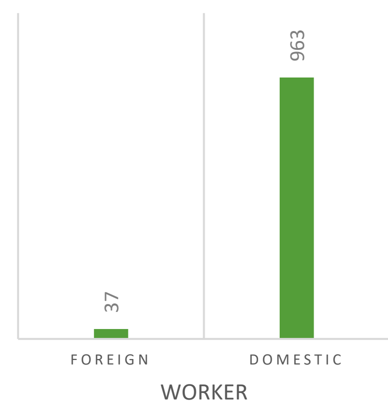
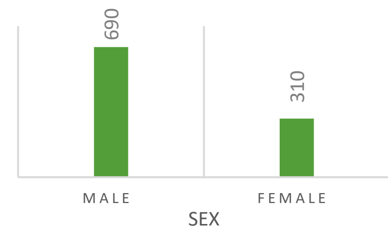
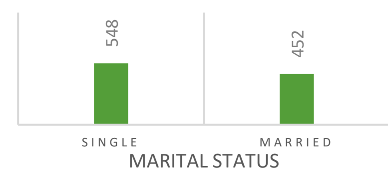
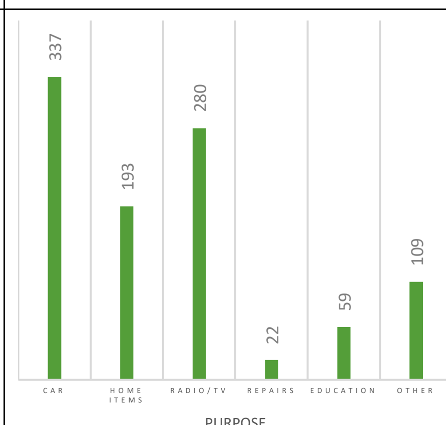
Variable Engineered	New Parameters	Justification	Column Chart After Engineering
Foreign Worker	Responses redefined – Domestic made to be majority statistic	Upon review, it was found that there were more foreign customers than domestic ones, which seemed to be a contradiction to the high approval of applications. It was found through research that these responses should be changed, so that there would be more domestic customers than foreign ones.	
Sex and Marital Status	Variable split into two new variables: Sex: • Male • Female Marital Status: • Single • Been married	Gender and marital status ultimately explain two different things. It allows each variable to be separately determined in predictive and cluster analysis, which may lead to provide stronger results due to the evening of the distribution amongst responses.	
Occupation	Re-grouped by level of skill: • Unskilled • Skilled • Highly skilled	Removes the non-resident response which had only a few observations, allowing for the data to be more simplified – leading to creating more distinct clusters and stronger prediction models.	
Purpose	Re-grouped by the similarities of the categories within the variables. • Car • Home Items • Radio/TV • Repairs • Education • Other	Improves the distribution of responses and allows for the data to be more simplified – leading to creating more distinct clusters and stronger prediction models.	

Table 1.22 – Feature engineering on categorical variables

Bivariate analysis

Pearson’s chi-square test is done on all variables since they are now all categorical after the feature engineering. Table 1.31 is a summary of the output from the chi-square test on creditability against the most significant predictor variables. Whereas, table 1.32 summarises which variables were not significant in determining creditability.

Variable	P-value	Correlation Coefficient	Interpretation
Account balance	0.000	0.351	Whilst all these variables are significant in determining creditability since p-value is less than 0.05, they do not appear to also be highly correlated to creditability. These variables are more likely to be considered when forming an optimal predictive model for creditability.
Payment status of previous credit	0.000	0.229	
Value savings stocks	0.000	0.179	
Duration of credit month	0.000	-0.178	
Most valuable available asset	0.000	-0.143	
Age	0.000	0.128	
Length of current employment	0.000	0.116	
Concurrent credits	0.001	0.110	
Foreign worker	0.009	-0.082	
Marital status	0.011	-0.081	
Credit amount	0.016	-0.076	
Sex	0.017	-0.075	
Instalment percentage	0.022	-0.072	

Table 1.31 – Variables that are significant predictors of creditability

Variable	P-value	Interpretation
Duration in current address	0.925	These variables all have p - values of more than 0.05, hence deemed insignificant. These variables are less likely to be considered when forming an optimal predictive model for creditability.
No. of dependents	0.924	
Type of apartment	0.567	
Purpose	0.469	
Guarantors	0.427	
Telephone	0.249	
Occupation	0.244	
No. of credits at this bank	0.148	

Table 1.32 – Variables that are not significant predictors of creditability

Bivariate analysis is also done between the independent variables, to determine if variables explain the same information, as shown in table 1.33.

Correlation Matrix	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Account balance (1)																					
Duration of credit (2)	-0.06																				
Pay credit status (3)	0.19	-0.05																			
Purpose (4)	0.04	0.10	-0.08																		
Credit amount (5)	-0.02	0.62	-0.04	0.00																	
Savings/stocks (6)	0.22	0.06	0.04	-0.03	0.05																
Employment duration (7)	0.11	0.06	0.14	0.03	0.00	0.12															
Instalment percentage (8)	-0.01	0.08	0.04	0.07	-0.28	0.02	0.13														
Sex (9)	-0.03	-0.08	-0.07	-0.03	-0.11	-0.03	-0.20	-0.09													
Marital status (10)	-0.05	-0.11	-0.09	0.01	-0.18	-0.06	-0.24	-0.12	0.74												
Guarantors (11)	-0.13	-0.02	-0.04	-0.01	-0.03	-0.11	-0.01	-0.01	-0.01	-0.01											
Current address duration (12)	-0.04	0.04	0.06	-0.05	0.02	0.09	0.25	0.05	0.01	-0.06	-0.03										
Most valuable asset (13)	-0.03	0.29	-0.05	0.00	0.30	0.02	0.09	0.05	-0.05	-0.15	-0.16	0.15									
Age (14)	0.09	-0.03	0.17	0.01	0.02	0.09	0.26	0.06	-0.24	-0.27	-0.03	0.21	0.07								
Concurrent credits (15)	0.07	-0.08	0.16	-0.09	-0.06	0.00	-0.01	0.01	0.02	0.05	-0.04	0.02	-0.11	-0.05							
Apartment type (16)	0.02	0.12	0.06	0.02	0.09	0.01	0.12	0.09	-0.22	-0.26	-0.07	0.01	0.34	0.32	-0.10						
No of credits (17)	0.08	0.03	0.44	0.06	0.02	-0.02	0.13	0.02	-0.09	-0.12	-0.03	0.09	-0.01	0.17	-0.06	0.05					
Occupation (18)	0.03	0.23	0.01	-0.02	0.29	0.02	0.05	0.08	-0.06	-0.07	-0.07	0.01	0.30	0.07	0.00	0.11	-0.01				
No of dependents (19)	-0.01	-0.02	0.01	-0.03	0.03	0.03	0.10	-0.07	-0.20	-0.28	0.02	0.04	0.01	0.20	-0.08	0.12	0.11	-0.10			
Telephone (20)	0.07	0.18	0.05	0.05	0.24	0.09	0.06	0.01	-0.08	-0.08	-0.08	0.10	0.20	0.18	-0.03	0.10	0.07	0.40	-0.01		
Foreign worker (21)	0.04	0.14	-0.03	0.11	0.06	-0.01	0.02	0.09	0.07	0.06	-0.14	0.04	0.13	-0.03	-0.01	0.08	0.02	0.09	-0.08	0.08	

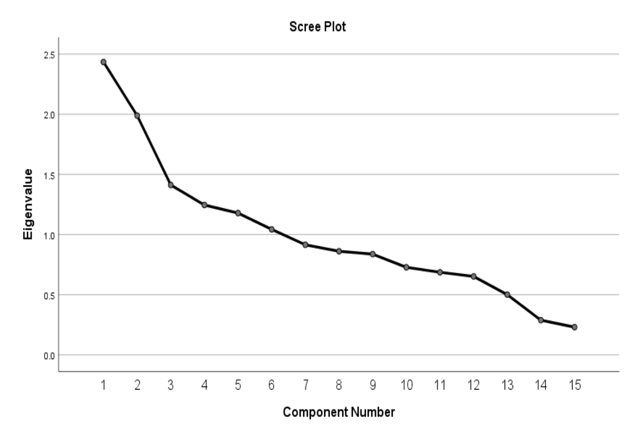
Table 1.33 – Correlation matrix heatmap

As highlighted in the correlation matrix heatmap, there are several variables that have a significant relationship between them. Therefore, based on all the output on bivariate analysis, it would be feasible to perform data reduction.

Data reduction

The German Credit dataset contains 21 possible predictors of creditability. In order to reduce the dimensionality of the dataset whilst being able to still define the reduced data, Exploratory Factor Analysis (EFA) will be conducted. However, we can see from the bivariate analysis that there are certain predictors that do not contribute significantly to creditability *and* have little association with other variables. Therefore, these predictors (top 6 variables in table 1.32) will not be included in the EFA.

Reference to the detailed results of the EFA can be seen in the appendix (A.1), however, graph 1.41 outlines the scree plot which was observed to determine the number of factors to use, and table 4.2 summarises and compares how each factor is defined when reduced in two ways.



Based on the scree plot, the line starts to flatten when it reaches 3 factors but based on the criterion of eigenvalues more than 1, 6 factors will be selected. The 6 factors will explain 62% of the total variance in the 16 predictors. It is important to note that by reducing the dataset to 6 factors, there will be around 38% loss of information. Whereas if only the first 3 factors are selected, there would be around 63% loss of information.

Graph 1.41 – Scree plot from exploratory factor analysis

	Main variables captured	
Factor ID	*6 factor results*	*3 factor results*
EF1	• Credit amount • Duration of credit • Most valuable available asset • Occupation This factor appears to look at the *financial criteria* of an applicant, and could be the most important factor to look at when predicting creditability, as it contains the most variance in the data.	• Credit amount • Duration of credit • Most valuable available asset • Occupation • Foreign worker

<table>
<tr><td><u>EF2</u></td><td>Sex
Marital status
Age
This factor also contains a high amount of variance within the data and points towards the demographics of an applicant.</td><td>Marital status
Sex
Age
Length of current employment
Instalment percentage</td></tr>
<tr><td><u>EF3</u></td><td>No. of credits at this bank
Payment status of previous credit
This factor seems to be looking at the credit status of an applicant.</td><td>Payment status of previous credit
No of credits at this bank
Account balance
Value of savings/stocks
Concurrent credits</td></tr>
<tr><td><u>EF4</u></td><td>Account balance
Value of savings/stocks
This factor is looking at how much cash in hand an applicant holds.</td><td rowspan="3">Whilst each factor in the 3 factor results is less definable than in the 6 factor results, they are worth taking into consideration when performing cluster and predictive analysis.</td></tr>
<tr><td><u>EF5</u></td><td>Instalment percentage
Credit amount
Foreign worker
This factor mainly captures the remaining variables.</td></tr>
<tr><td><u>EF6</u></td><td>Concurrent credits
This factor identifies that concurrent credits should remain to be independent amongst other variables</td></tr>
</table>

Table 1.42 – Factor composition summary & comparison from EFA

Inference

Through reviewing the dataset prior to analysis, more detailed information about each variable was able to be obtained. Undertaking univariate analysis allowed feature engineering to be done. The bivariate analysis then allowed us to see the relationship between variables and reduce the number of variables in the dataset, to help make future analysis simpler and with better results. Both EFA results will be tested when doing both cluster and predictive analysis.

2.0 **Determining market segments using cluster analysis**

Method

Two-step clustering analysis on SPSS will firstly be performed, to determine the optimal variables and number of clusters to use, based on the highest mean silhouette score. The higher the silhouette score (can range from -1 to 1), the more applicants are matched to its allocated cluster and poorly matched to neighbouring clusters. Once the optimal number of clusters is determined, then it will be tested against other clustering methods; k-means clustering and hierarchical clustering, to determine if a different method can provide a higher mean silhouette score (with the same number of clusters). Finally, the optimal clustering method, with the optimal number of clusters, will be analysed to determine how each cluster/market segment is defined.

Results

Table 2.1 summarises the results of mean silhouette scores through experimentation of different variable inputs and number of clusters.

Variable Input	Number of clusters	Mean Silhouette Score
3 factors from EFA	**7 (automatically determined by SPSS)**	**0.307**
3 factors from EFA	3	0.285
3 factors from EFA	10	0.284
3 factors from EFA	2	0.284
3 factors from EFA	5	0.283
6 factors from EFA	3 (automatically determined by SPSS)	0.172
12 variables (excluding all variables in table 1.32 and foreign worker)	2 (automatically determined by SPSS)	0.051
15 variables (excluding variables not included in data reduction)	2 (automatically determined by SPSS)	0.049
All 21 original variables	2 (automatically determined by SPSS)	0.043

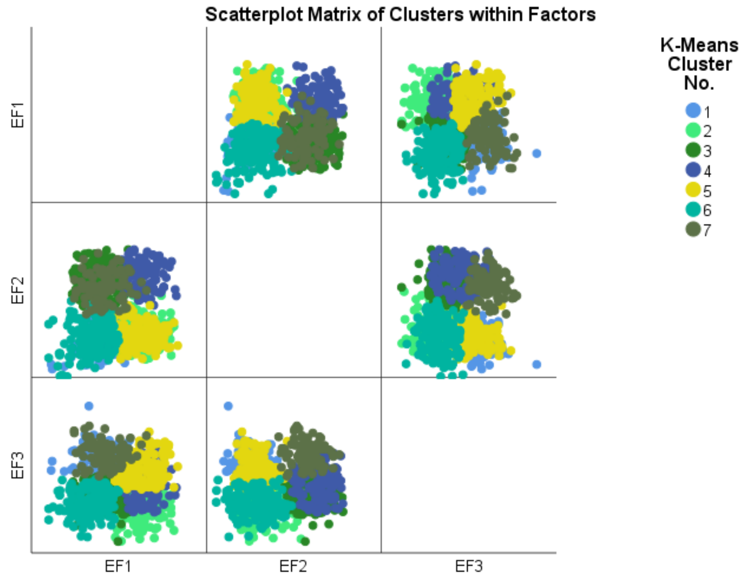
Table 2.1 – Silhouette score comparison summary from two-step clustering experimentation

Table 2.2 summarises the results of mean silhouette scores through experimentation of different clustering methods using the strongest number of clusters in table 2.1; 7 clusters.

Clustering Method	Mean Silhouette Score
K-Means Clustering	**0.314**
Two-Step Clustering	0.307
Hierarchical Clustering (Ward's method)	0.268
Hierarchical Clustering (Between groups linkage)	0.267
Hierarchical Clustering (Within groups linkage)	0.264
Hierarchical Clustering (Furthest neighbour)	0.196
Hierarchical Clustering (Centroid method)	0.184
Hierarchical Clustering (Median method)	0.172
Hierarchical Clustering (Nearest neighbour)	-0.25

Table 2.2 – Silhouette score comparison summary from clustering experimentation with predetermined 7 clusters

Graph 2.3 visually displays how each cluster was determined against the 3 factors used as input, and table 2.4 summarises how each of the 7 clusters resulting from k-means clustering are mainly differentiated. Reference to detailed results from k-means analysis can be found in the appendix (A.2).



Graph 2.3 – Scatterplot matrix of how each cluster visually differentiates from one another

Market Segment	Key descriptors (in rank of differentiation/significance)
1 - (11.8%) – Older male emergency loan applicants	• **Critical credit payment issues / existing credits outside of bank** • **Middle aged / retired and current employment over 7 years** • Moderate credit loan or less for under 18 months • Male and single • Real estate being their most valuable asset • Lower skilled occupation • Loan for car or radio/TV
2 – (16.5%) – Privileged male loan applicants	• **Owns a telephone** • **Holds 1 existing credit loan** • Car being their most valuable asset • Male and single • High credit loan for more than 18 months • Loan for car/other
3 – (17.3%) – Young housewives	• **Loan for radio/TV or home items** • **Young adult** • Female & Married • Moderate credit loan or less for under 18 months • Real estate being their most valuable asset • Been employed for between 0 – 4 years in a lower skilled occupation

4 – (12.1%) – Working females	• **Domestic worker employed for between 0 – 4 years** • Hold some money (under 200 EU) in a cheque account • High credit loan for more than 18 months • Female and married • Younger range of adults • Loan for car
5 – (16.7%) – Older high-flying males who can't manage their money	• **Has some sort of credit payment issue** • **Holds 2 to 3 credit loans** • **Owns a telephone** • **Working for over 4 years in higher skilled occupation** • High credit loan for more than 18 months • Male and single • Middle aged
6 – (16.6%) – Working class males	• **More likely to be a foreign worker than in other market segments** • **Hold one credit loan** • Moderate credit loan or less **for under 12 months** • Been employed for between 1 – 4 years in a lower skilled occupation • Male and single • Loan for car or radio/TV
7 – (9%) – Mature aged female students	• **Critical credit payment issues / existing credits outside of bank** • **More likely to want loan for education than in other market segments** • Female and married • Moderate credit loan or less for under 18 months • Real estate being their most valuable asset

Table 2.3 – Key features of clusters using k-means analysis

Inference

The results gathered in this section have shown that it is feasible to segment applicants in the database through cluster analysis - to provide more tailored services to the bank's different customers. Demographic, employment status, and credit status variables were the most significant drivers in differentiating clusters. However, it still needs to be determined whether clustering can also improve predictive analysis results.

3.0 Finding the best predictive algorithm to determines a loan applicant's creditability

Method

As part of the predictive modelling process, the data needs to split into training (the sample of data to fit the model) and testing (the sample of data to assess the performance of the model) subsets. In this instance, the dataset has been split into 70% training and 30% testing. Additionally, the process will be randomly initialised 42 times to ensure validity. A series of predictive models will then be used against the two new datasets to determine which model yields the highest F1-score. F1-score is used to evaluate the performance of the model, as it is a combined measure of both precision and recall. A range of data input combinations will also be tested to find the optimal model, including:

- All original variables - excluding variables not included in data reduction (Normal)
- 6 factor results from EFA as variables (6F)
- 3 factor results from EFA as variables (3F)
- Normal + two-step clustering (TSC) results with 6 factors as variables (Normal & TSC)
- 6F + TSC as variables (6F & TSC)
- Separate datasets of applicants only from one cluster within TSC with Normal as variables
- Separate datasets of applicants only from one cluster within TSC with 6F scores as variables

Once the optimal data input is found, each algorithm result will be tested against the cost matrix in table 3.1, which accounts for the trade-off between the two types of errors, to find the lowest cost method (where cost = [0*TP] + [1*FN] + [5*FP] + [0*TN]).

Cost Matrix		Predicted	
		Good	**Bad**
Actual	**Good**	0 (True Positive)	1 (False Negative)
	Bad	5 (False Positive)	0 (True Negative)

Table 3.1 – Cost matrix of errors in predictions – worse to approve applicant but in the end they're a bad debt to the bank

Results

Table 3.2 summarises the F1 score results from experimentation with different data input and predictive algorithms, with the best data input to predict creditability highlighted.

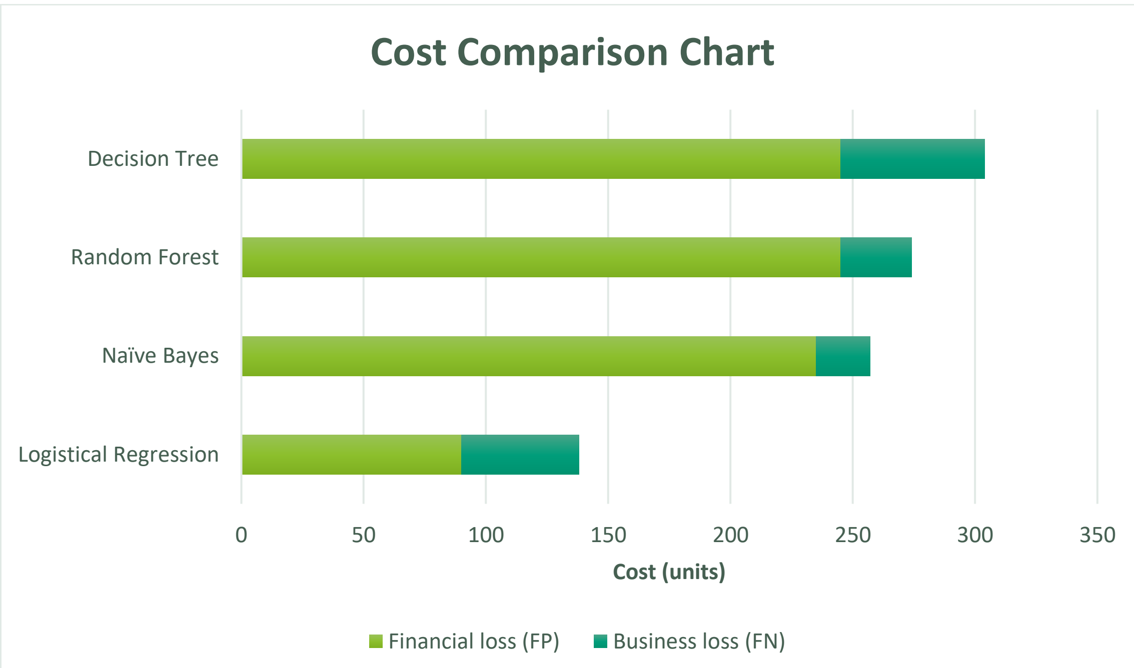
F1 Scores	**Predictive Algorithm**			
Data Input	**Logistical Regression**	**Naïve Bayes**	**Random Forest**	**Decision Tree**
Normal	0.842	0.792	0.827	0.781
6F	0.852	0.844	0.821	0.734
3F	0.804	0.806	0.752	0.752
Normal & TSC	0.839	0.770	0.808	0.796
6F & TSC	0.846	0.840	0.806	0.711
NormalC1	*0.836*	*0.143*	*0.807*	*0.802*
NormalC2	*0.677*	*0.698*	*0.610*	*0.724*
NormalC3	*0.843*	*0.029*	*0.815*	*0.656*
NormalC (Weighted average)	0.814	0.192	0.779	0.743

6FC1	*0.849*	*0.841*	*0.817*	*0.761*
6FC2	*0.727*	*0.677*	*0.557*	*0.567*
6FC3	*0.826*	*0.803*	*0.734*	*0.672*
6FC (Weighted average)	0.823	0.804	0.750	0.703

Table 3.2 – Comparison of F1 score results between different data inputs and predictive algorithms

The optimal market segment cluster results in section 2.0 were not tested, due to the poorer prediction results when using 3 factors (which was the basis of forming these clusters), as demonstrated in table 3.2.

Table 3.2 shows that logistic regression yields the highest F1-Score using 6 factors as input. However, to validate if this is the best model for the bank to use, cost of errors was calculated using the cost matrix in table 3.1. Graph 3.3 summarises the costs of each predictive method. The detailed confusion matrices of each algorithm can be found in the appendix A.3.



Graph 3.3 – Cost of errors comparison between different predictive algorithms

Inference

Whilst each predictive algorithm had similar F1 scores, logistical regression using the 6 factors in EFA is the best model to reduce the risk of the bank incorrectly predicting the creditability of an applicant and consequently minimise hefty financial loss by having more potential business loss. Using clusters allocation as a variable and doing separate prediction models on each cluster did not improve prediction modelling results.

Conclusion

In summary, the following are the key findings of this report:

- After appropriate action taken prior to analysis, including feature engineering and data reduction, the German credit dataset was optimised to provide higher quality clusters of applicants and predictions of creditability.
- Financial criteria were the key driver of predicting creditability whereas demographics, employment and credit status were the key differentiators in determining market segments.
- Seven clusters using k-means clustering provided the most distinct market segments for the German bank.
- Logistical regression with data reduction of variables to 6 factors provides the strongest prediction model to determine an applicant's creditability whilst minimising financial loss on errors.

Management at the bank should ultilise these key findings to improve services for their customers and maximise profit.

Appendix

A.1 Detailed exploratory factor analysis results

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.434	16.227	16.227	2.434	16.227	16.227	2.095	13.966	13.966
2	1.989	13.259	29.485	1.989	13.259	29.485	2.000	13.330	27.297
3	1.412	9.411	38.896	1.412	9.411	38.896	1.529	10.193	37.490
4	1.245	8.303	47.199	1.245	8.303	47.199	1.315	8.767	46.257
5	1.178	7.855	55.054	1.178	7.855	55.054	1.273	8.485	54.741
6	1.043	6.953	62.007	1.043	6.953	62.007	1.090	7.265	62.007
7	.915	6.100	68.106						
8	.861	5.741	73.847						
9	.837	5.582	79.429						
10	.728	4.853	84.282						
11	.686	4.576	88.858						
12	.652	4.347	93.205						
13	.501	3.337	96.543						
14	.288	1.921	98.464						
15	.230	1.536	100.000						

Extraction Method: Principal Component Analysis.

Rotated Component Matrix - 6 Factors

	Component					
	1	2	3	4	5	6
Duration of Credit	0.783	-0.044	-0.004	-0.025	-0.071	-0.047
Credit Amount	0.782	-0.106	0.017	0.018	-0.455	-0.013
Most Valuable Available Asset	0.619	-0.068	-0.041	-0.003	0.199	-0.177
Occupation	0.590	-0.035	-0.007	0.046	0.210	0.158
Sex	-0.055	0.891	-0.008	0.039	0.023	-0.056
Marital Status	-0.133	0.890	-0.038	-0.008	-0.005	-0.003
Age	-0.023	-0.421	0.291	0.251	0.117	-0.235
No. Credits at Bank	0.016	-0.065	0.843	-0.060	-0.013	-0.153
Payment Status of Previous Credit	-0.045	-0.056	0.799	0.127	0.018	0.250
Value of Savings/Stocks	0.047	-0.041	-0.129	0.787	-0.040	-0.070
Account Balance	-0.031	0.017	0.173	0.704	-0.008	0.160
Instalment Percentage	-0.024	-0.149	-0.027	-0.041	0.823	0.051
Foreign Worker	0.328	0.252	0.055	0.020	0.443	-0.036
Length of Current Employment	0.048	-0.348	0.200	0.327	0.293	-0.155
Concurrent Credits	-0.051	0.013	0.042	0.054	0.029	0.910

Rotation Method: Varimax with Kaiser Normalization.
Rotation converged in 6 iterations
Highlighted if main variable(s) captured in factor

Rotated Component Matrix - 3 Factors

	Component		
	1	2	3
Credit Amount	0.805	-0.039	-0.071
Duration of Credit	0.784	-0.020	-0.048
Most Valuable Available Asset	0.619	-0.093	-0.052
Occupation	0.563	-0.024	0.077
ForeignWorker	0.302	0.193	0.145
Marital Status	-0.145	0.871	0.065
Sex	-0.061	0.857	0.107
Age	0.014	-0.503	0.288
Length of Current Employment	0.072	-0.433	0.306
Instalment Percentage	-0.075	-0.225	0.084
Payment Status of Previous Credit	-0.057	-0.121	0.758
NoofCreditsatthisBank	0.028	-0.172	0.605
Account Balance	0.003	-0.023	0.540
Value of Savings/Stocks	0.109	-0.076	0.279
Concurrent Credits	-0.128	0.122	0.258

Rotation Method: Varimax with Kaiser Normalization.
Rotation converged in 5 iterations
Highlighted if main variable(s) captured in factor

A.2 Detailed k-means clustering results

ANOVA						
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
EF2	135.326	6	0.188	993	718.436	0.000
EF1	116.913	6	0.300	993	390.209	0.000
EF3	113.242	6	0.322	993	351.899	0.000

K-Means Cross-Tabulation								
	Cluster Number	1	2	3	4	5	6	7
Account Balance	<0 EU		48%	36%			36%	
	0<=...<200 EU		33%	30%	39%		31%	
	>=200 EU							
	No account	63%				63%		66%
Duration of Credit (months)	0-12	58%		54%			73%	54%
	13-18	22%		27%				28%
	19-24		26%		33%	35%		
	25-72		52%		54%	44%	0%	
Payment Status of Previous Credit	No credits taken / all credits paid back duly	0%				0%		0%
	All credits at this bank paid back duly	0%				0%		
	Existing credits paid back duly till now		65%	86%	62%	30%	74%	
	Delay in paying off in the past					22%		
	Critical account / other credits existing	77%				48%		78%
Purpose	Car	37%	32%		36%	40%	36%	32%
	Homeware			28%				
	Radio / TV	35%		33%			34%	
	Repairs							
	Education							12%
	Other		19%			19%		
Credit Amount	Low	40%		46%			46%	44%
	Moderate	37%		33%			36%	40%
	High		33%		34%	36%		
	Very High		54%		55%	49%		
Length of Current Employment	Unemployed							
	< 1 yr			38%	26%			
	1 <= ... <4 yrs		38%	39%	32%		42%	36%
	4<=...< 7 yrs		22%			22%		
	>= 7 yrs	55%				44%		
Sex	Male	100%	99%			100%	99%	
	Female	0%		76%	85%	0%		81%
Marital Status	Single	90%	92%		0%	92%	81%	0%
	Married			99%	100%			100%
Most Valuable Available Asset	Real estate	47%		43%			50%	36%
	Savings agreement / life insurance	25%		29%			25%	29%
	Car / Other		41%		54%	46%		
	Unknown / none		35%		19%	26%		
Age	Young Adult			45%	34%			
	Adult		49%	40%	45%	39%	38%	39%
	Middle-Aged	68%	34%			53%	45%	46%
	Retired	6%						
No of Credits at this Bank	1		84%	94%	68%		84%	
	2 or 3	67%				55%		69%
	4 or 5			0%	0%		0%	
	Above 6		0%	0%	0%		0%	
Occupation	Unskilled	30%		31%			46%	
	Skilled	67%	66%	65%	68%	63%	51%	64%
	Highly Skilled					32%		
Telephone	No	66%		75%	53%		80%	56%
	Yes		53%			62%		
Foreign Worker	Yes				0%	0%	16%	
	No	96%	99%	99%	100%	100%	84%	99%

A.3 Confusion matrices of predictive algorithms

Logistical Regression Confusion Matrix		Predicted	
		Good	Bad
Actual	Good	190	48
	Bad	18	44

Naïve Bayes Confusion Matrix		Predicted	
		Good	Bad
Actual	Good	186	22
	Bad	47	45

Random Forest Confusion Matrix		Predicted	
		Good	Bad
Actual	Good	179	29
	Bad	49	43

Decision Tree Confusion Matrix		Predicted	
		Good	Bad
Actual	Good	149	59
	Bad	49	43