PREDICTIVE ANALYTICS

A SOLUTION TO MINIMISING RISK AND MAXIMISING REVENUE FOR FINANCIAL INSTITUTIONS?

BUSA3020 - ASSIGNMENT 4
GROUP 13
RIJWA ABBAS 45697795
Chieh-Ying LO 45561265
HASSAN JAKHURA 45544344
ROSE DEVINE 45414394

Contents

Executive Summary

The report employs various predictive algorithms and models to minimize the risk of giving bad loans and refusing good loans. Logistic regression with Reduced Variables using Python's recursive elimination feature and oversampled bad loans proved to be the most appropriate model with the lowest cost matrix score. Random forest with reduced variables and oversampled bad loans had the highest F1 score overall in 20 Fold cross-validation. However, random forest performed rather poorly for test-on-test data. Given Logistic regression's superiority to other models in 70:30 Random Sampling and repeat testing hundred times, this report concludes Reduced variable Logistic Regression to be the most appropriate model. A 93% Kaggle Score achieved through this model further supports this conclusion.

Background and Situational Review

Nature of Problem

Approving a bad loan and refusing a good loan is a risk, one more costly than the other, financial institutions bear every day. This report seeks to create a predictive model to aid loan managers in minimising the financial risk by predicting bad loans efficiently and accepting good loans.

Current Situation

A German Credit Data set, containing 20 variables (Table 1) and 800 observations, is used to generate a predictive model. The model obtained is employed to predict the credibility of an additional 150 observations containing the same variables, however, without a credibility score.

Table 1 Available Variables for German Credit Data:

Account Balance	Value Savings/Stocks	Duration in Current address	Number of Credits at this Bank
Duration of Credit (month)	Length of current employment	Assets	Occupation
Payment Status of Previous Credit	Instalment percent	Age (years)	Number of dependents
Purpose of Loan	Sex & Marital Status	Concurrent Credits	Telephone
Credit Amount	Guarantors	Housing	Foreign Worker

Method

Research Question

Can a predictive model minimize the cost of giving bad loans and refusing good loans?

Data gathering and Data Treatment

German Credit Data was obtained from iLearn and given meaningful names (Table 2).

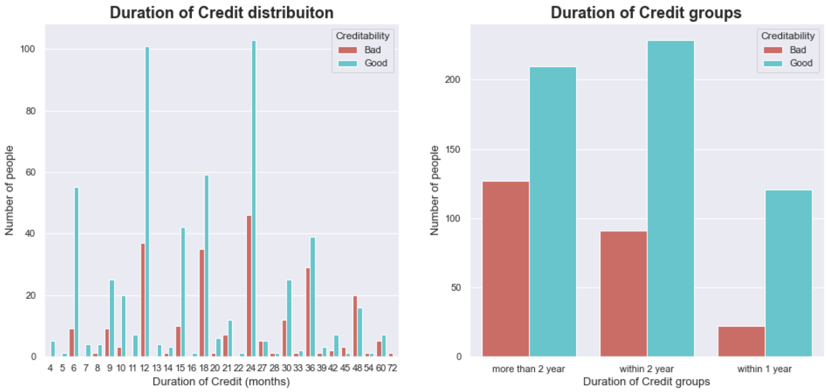
Table 2 Renaming Variable and sub-variable categories:

Variable name	Account Balance	Value Saving/Stocks	Instalment rate percent
Change of category names	<0 → Negative = <200 EU → poor = >200 EU → rich	=<100 EU → poor between → moderate =>1000 EU → rich	<20% → less between → moderate >35% → most

Further, groups of variables were created based on distribution to identify important variables better. For instance, continuous data like Duration of Credit, Age, and Credit Amount were grouped based on Credibility distribution to potentially assist with identifying good or bad credit (Table 3).

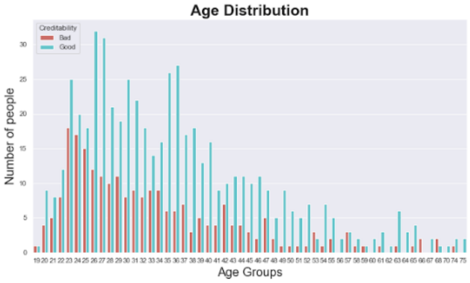
Duration of Credit:

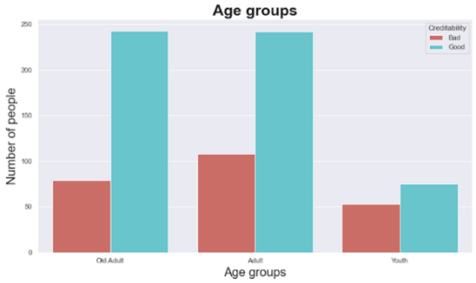
Based on high peaks (12 months and 24 months) of distribution to create groups



Age:

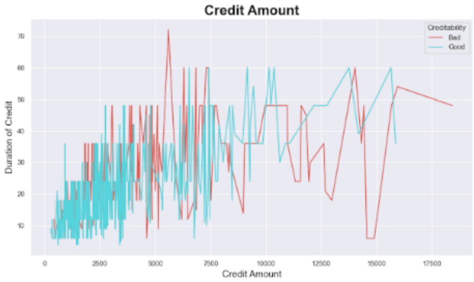
Based on good/bad credibility ratio and number of people in different age (25- and 36-year-old) in distribution to create group





Credit Amount

Based on good/bad credibility ratio and amount (2k and 10k) of distribution to create groups



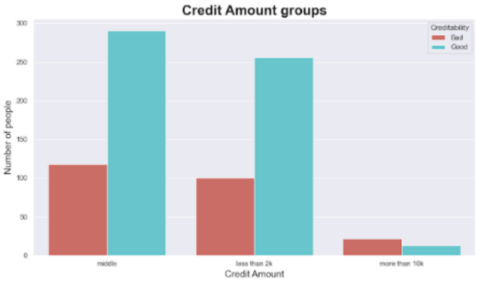
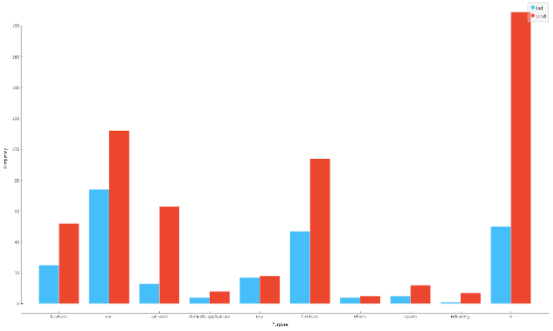
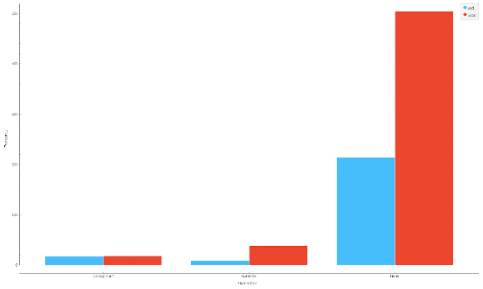
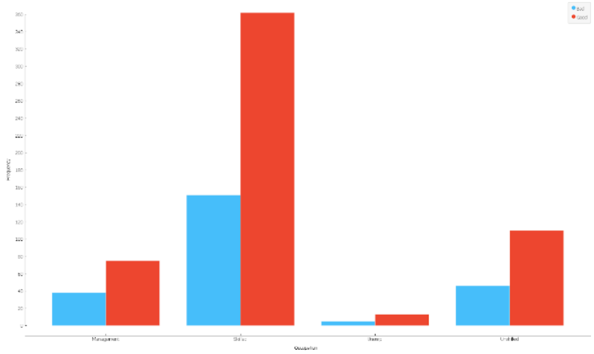
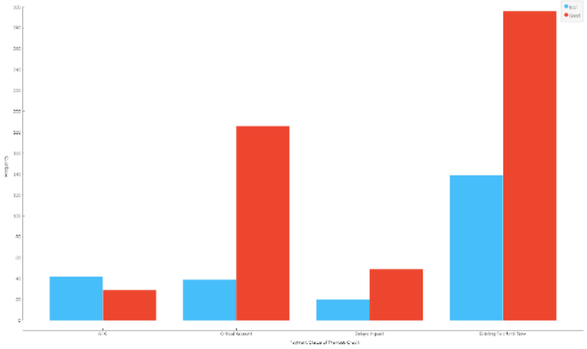
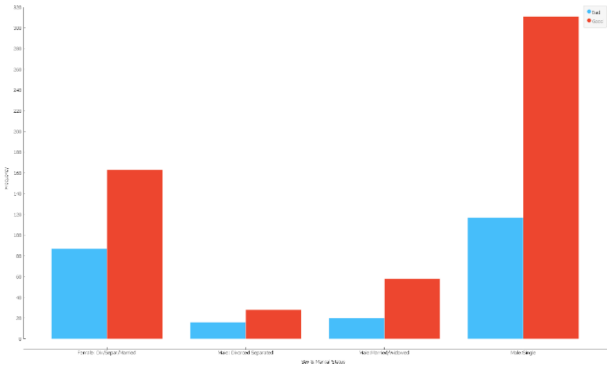


Table 3: Grouping Continuous data

Similar categories or ones with less observations were merged following univariate analysis (Table 4).

Variable Name	**Justification for combining**	**Name for new categories**
Purpose	Similar categories Education and retraining, furniture/equipment and radio/TV are similar categories, so it could be better to combine them for data gathering. 	education and retraining → education furniture/equipment and radio/TV → furniture/equipment
Guarantor	Small observations Both co-applicant and guarantor have less than 100 observations, compared with the non-guarantor category. 	co-applicant and guarantor → guarantor
Occupation	Small observations unemployed only has 18 observations in occupation, which is too low, so combine it with another similar category, which could reduce the impact on analysis .	unemployed and unskilled → unskilled

		
Payment Status of Previous Credit	Small observation No credits taken/ all credits paid back duly and all credits at this bank paid back duly have less observation ,compared with other categories, so they can be combined as one variable for data gathering. 	no credits taken/ all credits paid back duly and all credits at this bank paid back duly → no history
Marital Status	Similar categories Small observation Divorced/separated is similar to single and it has 48 observations in total, so it can be combined with single to differentiate the married status. 	divorced/separated and single → single

Analysis Procedures

Predictive Modelling

Orange and Python were employed for predictive modeling due to ease of use. Treated data was further split into 70% Training and 30% Testing Data. All the treated variables were used to form a baseline score using 20 Fold Cross-Validation (Table 5) to examine the impact of variable engineering, and other steps undertaken to improve the model. Further, the 30% testing data was used to measure the model's out-of-sample predictive (Table 6) and cost-matrix score.

Table 5 Baseline Score 20 Fold Cross Validation

Baseline	**AUC**	**CA**	**F1**	**Precision**	**Recall**
SVM	0.740	0.748	0.731	0.732	0.748
Logistic Regression	0.734	0.730	0.721	0.717	0.730
Random Forest	0.715	0.720	0.703	0.700	0.720
Naive Bayes	0.737	0.696	0.703	0.713	0.696

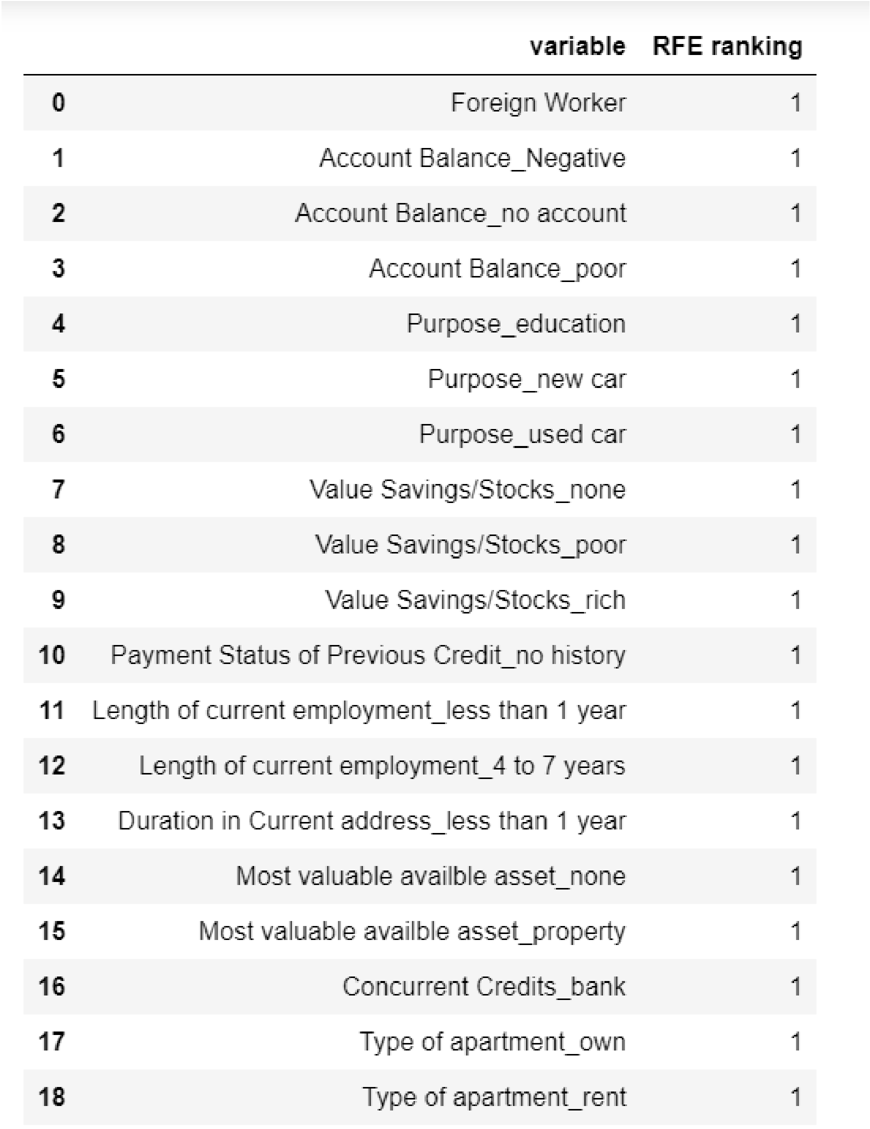
Table 6 Test on Test Baseline Score

TOT Baseline	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Logistic Regression	0.793	0.758	0.749	0.747	0.758
SVM	0.769	0.754	0.748	0.745	0.754
Naive Bayes	0.797	0.742	0.747	0.756	0.742
Random Forest	0.774	0.729	0.714	0.711	0.729

The following Table documents steps undertaken to improve the model. Bad loan predicted as good are penalised with 5 units and good loans predicted as bad are penalised with one to account for higher financial loss and risk bad loan carry.

Improving the model: Recursive Feature Elimination

Variable’s ranked 1 through Python’s Recursive Elimination Feature (Figure 1) were then used to examine the difference.



	variable	RFE ranking
0	Foreign Worker	1
1	Account Balance_Negative	1
2	Account Balance_no account	1
3	Account Balance_poor	1
4	Purpose_education	1
5	Purpose_new car	1
6	Purpose_used car	1
7	Value Savings/Stocks_none	1
8	Value Savings/Stocks_poor	1
9	Value Savings/Stocks_rich	1
10	Payment Status of Previous Credit_no history	1
11	Length of current employment_less than 1 year	1
12	Length of current employment_4 to 7 years	1
13	Duration in Current address_less than 1 year	1
14	Most valuable availble asset_none	1
15	Most valuable availble asset_property	1
16	Concurrent Credits_bank	1
17	Type of apartment_own	1
18	Type of apartment_rent	1

Figure 1 Reduced Variable Score 20 Fold Cross Validation

Table 7 shows improvement in Logistic Regression and Naïve Bays model, however, the test on test data only shows improvement in logistic regression and Random Forest (Table 8)

Table 7 Reduced Variable Score 20 Fold Cross Validation:

Reduced Variables	AUC	CA	F1	Precision	Recall
Logistic Regression	0.764	0.750	0.733	0.734	0.750
Naive Bayes	0.764	0.738	0.732	0.729	0.738
SVM	0.686	0.714	0.691	0.689	0.714
Random Forest	0.707	0.698	0.684	0.679	0.698

Table 8 Test on Test Reduced Variable Score:

TOT Reduced Variables	AUC	CA	F1	Precision	Recall
Logistic Regression	0.794	0.771	0.758	0.759	0.771
Random Forest	0.792	0.758	0.752	0.749	0.758
Naive Bayes	0.804	0.746	0.745	0.745	0.746
SVM	0.697	0.700	0.700	0.700	0.700

Other variables with a significant relationship with bad loans were later included in the model to observe improvement. Variables like "More than 10K credit amount", "Credit Duration more than two years," and age group "youth" seemed to improve the score when added individually; however, adding these variables weakened the model for out-of-sample predictions.

Contrary to the predictability score, random forest and naïve bays do not perform well for cost matrix. However, the Logistic Regression cost matrix score was reduced by 20 points. (Table 11)

Table 9 Baseline Logistic Regression Confusion Matrix and Cost Matrix Score

Baseline Logistic Regression Confusion Matrix	**Predicted Bad**	**Predicted Good**
Actual Bad	32	40
Actual Good	21	147
Cost Matrix Score	40*5+21*1	221

Table 10 Reduced Variable Logistic Regression Confusion Matrix and Cost Matrix Score

Reduced Variable Logistic Regression Confusion Matrix	**Predicted Bad**	**Predicted Good**
Actual Bad	34	38
Actual Good	11	157
Cost Matrix Score	38*5+11*1	201

Table 11 Cost Matrix Baseline vs Reduced Variable Score

Model	Baseline	Reduced Variable
Logistic Regression	221	201
Random Forest	228	246
Naïve Bays	167	184
SVM	265	299

It is noteworthy that Naïve Bays' Baseline score is lowest overall through the Baseline model and worsens with reduction in variables.

Improving the predictive Model: Oversampling

Given the unbalanced data - twice as many good loans as bad loans - the bad loans were oversampled to the size of good loans. Table 12 and 13 show predictability results.

Table 12 20-Folds Cross Validation Oversampled-Reduced Variable Score

Reduced Variables Over Sampled	AUC	CA	F1	Precision	Recall
Random Forest	0.862	0.764	0.763	0.767	0.764
Logistic Regression	0.773	0.708	0.708	0.708	0.708
Naive Bayes	0.768	0.691	0.691	0.692	0.691
SVM	0.688	0.619	0.618	0.620	0.619

Table 13 Test-on-Test Data Over-Sampled Reduced Variable Score

TOT Reduced Variable Over Sampled	AUC	CA	F1	Precision	Recall
Logistic Regression	0.782	0.675	0.688	0.742	0.675
Naive Bayes	0.791	0.671	0.684	0.728	0.671
Random Forest	0.721	0.671	0.680	0.698	0.671
SVM	0.553	0.604	0.608	0.612	0.604

Although the scores are significantly lower, when penalized for bad loan as good loan, the over sampled model out performs with a lowest score of 150 for logistic regression. (Table 15)

Table 14 Oversampled Reduced Variable Logistic Regression Confusion Matrix and Cost Matrix Score:

Over Sampled Reduced Variable Logistic Regression Confusion Matrix	Predicted Bad	Predicted Good
Actual Bad	54	18
Actual Good	60	108
Cost Matrix Score	18*5+60*1	150

Table 15 Cost Matrix Scores Reduced Variable vs Oversampled Reduced Variable:

Model	**Reduced Variable**	**Oversampled Reduced Variable Score**
Logistic Regression	201	150
Random Forest	246	163
Naïve Bays	184	199
SVM	299	275

Improving the predictive model: Clustering

Data was handled using the t-SNE approach as this algorithm tries to separate values into smaller components. Perplexity and exaggeration was set accordingly to identify nearest neighbours. K-means were further used to identify clusters (Figure 2).

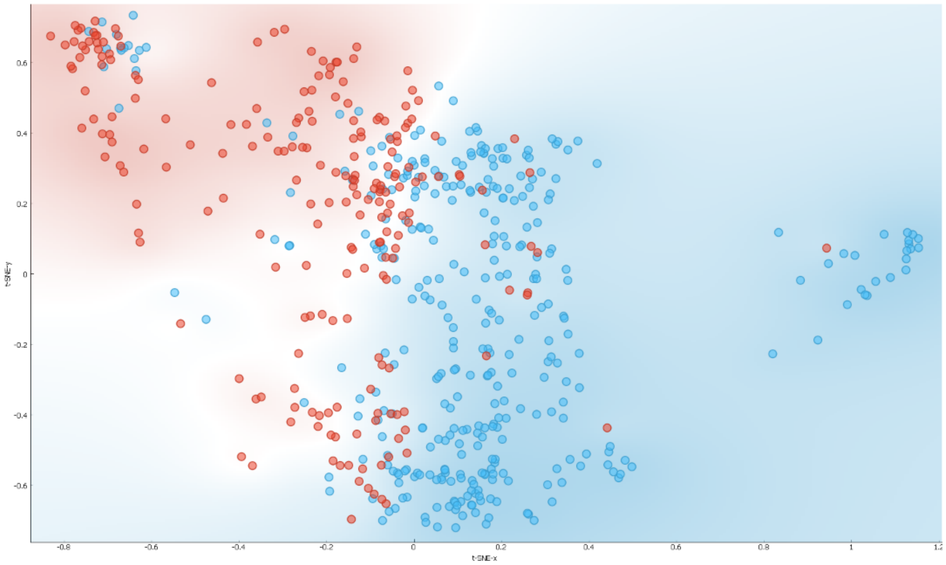


Figure 2 K-Means Scatter plot showing two clusters

These two clusters were profiled using the help of Sieve Plots (Appendix 2). It was discovered that Cluster 1 consists of people likely to be either rich or with a negative bank account, in possession of property and highly likely to ask for an amount less than 2k. People in cluster 2 are likely to be married and pay back credit within 2 years. They are also likely to be female and unskilled, and often require loans for furniture/equipment, repairs, car or education etcetera, living in their own or rented home. People in cluster 1 are more likely to pay back loans compared to people in cluster 2.

People in cluster 2 tend to have a poorer account balance, no property, require a larger loan, often for more than two years. They also tend to be single, male and have a management job, requiring loans for business but mostly for buying used cars. People in cluster 2 are more likely to default.

Given this significant association between people in the two different clusters and defaulting, Clustering is used to examine if the model predictability can be improved. A Baseline model was first to compare scores of other modifications.

Baseline Score

A noticeable improvement can be witnessed over both Baseline and Reduced variables models with a F1 score of 0.749 (Table 16) compared to .731 or .733 for 20 Fold Cross Validation. However, Reduced Variable Model performs better than Clustering Model for Test on Test data with a score of .752 compared to .749 (Table 17). Further, the cost matrix score for logistic regression in the clustering model, too, is higher than the reduced variable version.

Table 16 Clustering 20 Fold Cross Validation Baseline Score:

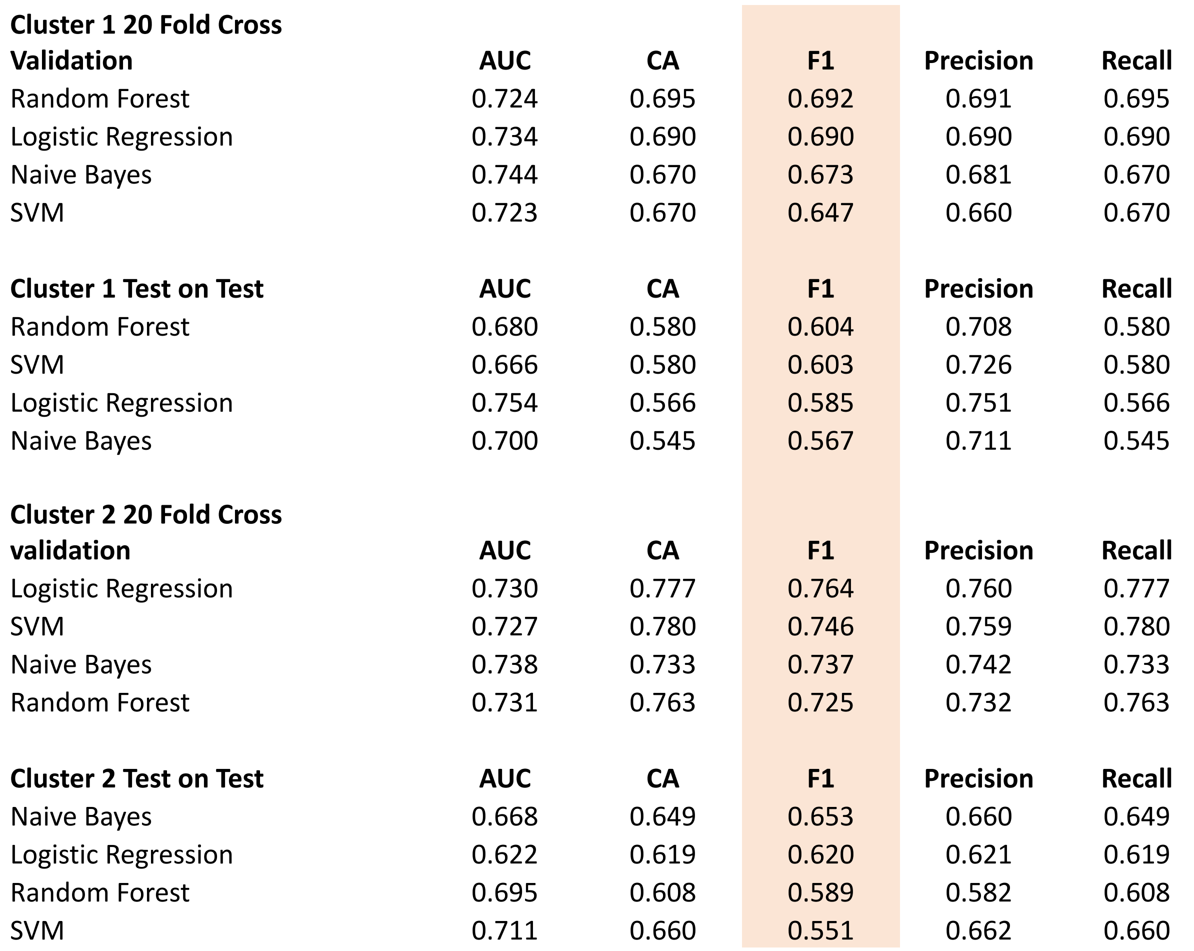
Clustering Cross Validation	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Logistic Regression	0.797	0.755	0.749	0.746	0.755
Random Forest	0.733	0.720	0.704	0.700	0.720
SVM	0.754	0.730	0.720	0.717	0.730
Naive Bayes	0.769	0.725	0.727	0.730	0.725

Table 17 Clustering Test on Test Score

Clustering Test on Test	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Logistic Regression	0.753	0.763	0.749	0.750	0.763
Naive Bayes	0.767	0.721	0.723	0.727	0.721
Random Forest	0.705	0.713	0.688	0.686	0.713
SVM	0.689	0.708	0.679	0.679	0.708

Clustering Model Based on Separate Cluster:

The predictive algorithms were applied on two clusters separately to observe improvement in score. The following figure shows the result.



Cluster 1 20 Fold Cross Validation	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Random Forest	0.724	0.695	0.692	0.691	0.695
Logistic Regression	0.734	0.690	0.690	0.690	0.690
Naive Bayes	0.744	0.670	0.673	0.681	0.670
SVM	0.723	0.670	0.647	0.660	0.670

Cluster 1 Test on Test	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Random Forest	0.680	0.580	0.604	0.708	0.580
SVM	0.666	0.580	0.603	0.726	0.580
Logistic Regression	0.754	0.566	0.585	0.751	0.566
Naive Bayes	0.700	0.545	0.567	0.711	0.545

Cluster 2 20 Fold Cross validation	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Logistic Regression	0.730	0.777	0.764	0.760	0.777
SVM	0.727	0.780	0.746	0.759	0.780
Naive Bayes	0.738	0.733	0.737	0.742	0.733
Random Forest	0.731	0.763	0.725	0.732	0.763

Cluster 2 Test on Test	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Naive Bayes	0.668	0.649	0.653	0.660	0.649
Logistic Regression	0.622	0.619	0.620	0.621	0.619
Random Forest	0.695	0.608	0.589	0.582	0.608
SVM	0.711	0.660	0.551	0.662	0.660

It can be deduced that the Cluster 2 predictive model performs comparatively well with an F1 Score of .764. However, the test score is rather low.

Oversampling Clustering Model:
The treated data was oversampled to balance the observations and to examine improvement in clustering model. The overall highest F1 score for the Clustering model, .752 (Table 18), was lower than Reduced - Variable Oversampled Random Forest's F1 score, .763. However it was an improvement on the logistic regression's score in the Reduced variable model, 0.752 compared to 0.708. This improvement however, was not shared by the testing model and resulted in a score of 0.650 (Table 19) for logistic regression and a cost matrix score of 241 (Table 20) compared to 150 in Oversampled Reduced Variable Score.

Table 18: Oversampled 20 Fold Cross Validation Clustering Model

Clustering Oversampling Cross Validation	**AUC**	**CA**	**F1**	**Precisio n**	**Recall**
Logistic Regression	0.792	0.757	0.752	0.750	0.757
Random Forest	0.766	0.738	0.720	0.720	0.738
Naive Bayes	0.771	0.716	0.720	0.726	0.716
SVM	0.738	0.725	0.712	0.709	0.725

Table 19 Oversample Test on Test Clustering Model

Clustering Oversampling Test on Test	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Naive Bayes	0.743	0.704	0.712	0.727	0.704
SVM	0.672	0.704	0.693	0.688	0.704
Random Forest	0.664	0.700	0.686	0.681	0.700
Logistic Regression	0.661	0.646	0.650	0.656	0.646

Table 20 Oversampled Clustering Logistic Regression Confusion and Cost Matrix

Logistic Regression	**Predicted Bad**	**Predicted Good**
Actual Bad	33	39
Actual Good	46	122
Cost Matrix Score	39*5 + 46 *1	241

Results

Predictive Model Of Choice

Reduced-Variable, Oversampled Logistic Regression outperforms all other models with lowest cost matrix score, however, Reduced-Variable, Oversampled Random Forest has the highest F1 Score. However, when data is randomly sampled and split in 70-30 a hundred times, logistic regression outweighs all other models (Table 21). It is worth mentioning that reduced variable logistic regression achieved a score of 93 on Kaggle whereas oversampled, reduced variable was a close second with 91 score. This difference, however, can be attributed to sample differences.

Table 21 70-30 Randomly sampled and tested hundred times

70 Train – 30 Test Split Randomly Sampled and tested hundred Times	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Logistic Regression	0.764	0.754	0.736	0.738	0.754
Naive Bayes	0.766	0.726	0.726	0.726	0.726
Random Forest	0.718	0.725	0.714	0.710	0.725
SVM	0.668	0.694	0.684	0.678	0.694

It is also worth noticing that Model generated through Cluster 2 when randomly sampled and tested hundred times, out performs the aforementioned model. (Table 22).

Table 22 70-30 Randomly sampled and tested hundred times

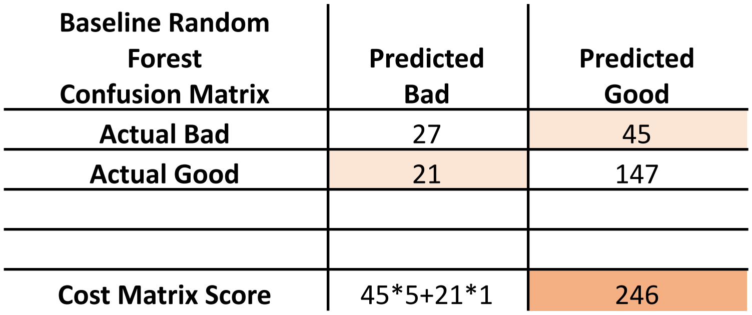
70 Train – 30 Test Split Randomly Sampled and tested hundred Times	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Logistic Regression	0.735	0.771	0.758	0.754	0.771
Naive Bayes	0.739	0.725	0.731	0.739	0.725
Random Forest	0.724	0.755	0.721	0.723	0.755
SVM	0.735	0.773	0.716	0.770	0.773

Conclusion

To conclude, the Oversampled, Reduced-Variable Logistic Model was found to be superior to all other models generated. However, it was second to Reduced-Variable Logistic Model in obtaining a high Kaggle score (91 compared to that of 93). Cluster 2, defined above, seems to outperform the aforementioned models with a significantly higher F1 score, however, its superiority is inconclusive owing to inability to measure Cost Matrix score given comparatively less number of observations. Regardless, this score provides an opportunity for further experimentation and research, and enables us to safely conclude that predictive models can reduce cost matrix scores to a great extent.

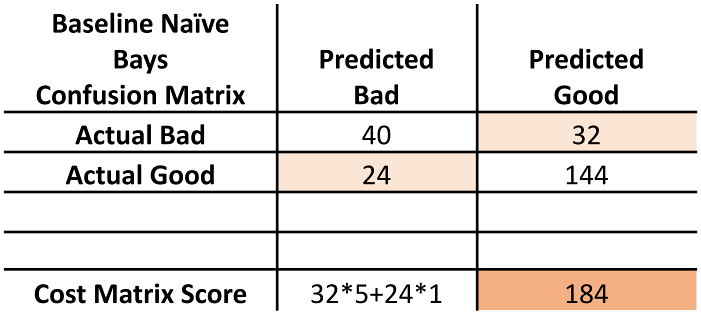
Lastly, the users of this report are recommended to use Reduced Variable Logistic Model owing to its reliability. Users are also encouraged to gather new data given the outdated nature of data. Further research might provide more insights given the changed dynamics and gender since the data was acquired.

Appendices:



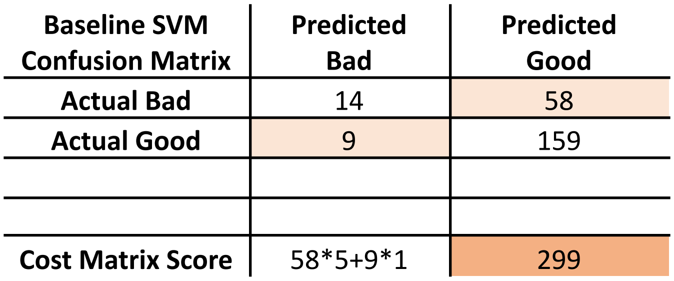
Baseline Random Forest Confusion Matrix	Predicted Bad	Predicted Good
Actual Bad	27	45
Actual Good	21	147
Cost Matrix Score	45*5+21*1	246

Baseline Random Forest Confusion Matrix and Cost Matrix Score



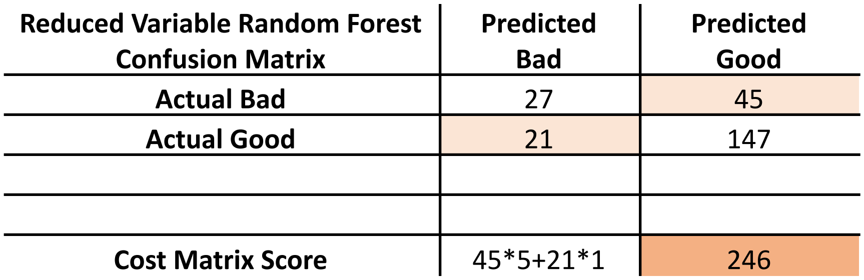
Baseline Naïve Bays Confusion Matrix	Predicted Bad	Predicted Good
Actual Bad	40	32
Actual Good	24	144
Cost Matrix Score	32*5+24*1	184

Baseline Naive Bays Confusion Matrix and Cost Matrix Score



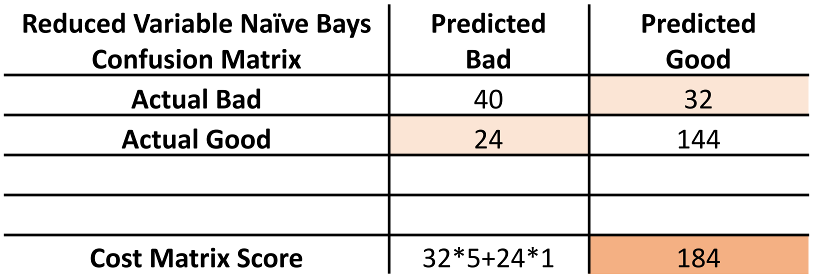
Baseline SVM Confusion Matrix	Predicted Bad	Predicted Good
Actual Bad	14	58
Actual Good	9	159
Cost Matrix Score	58*5+9*1	299

Baseline SVM Confusion Matrix and Cost Matrix Score



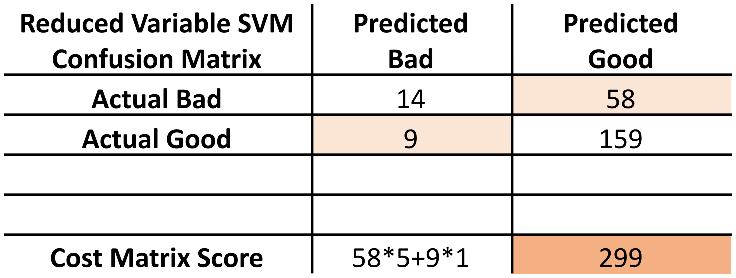
Reduced Variable Random Forest Confusion Matrix	Predicted Bad	Predicted Good
Actual Bad	27	45
Actual Good	21	147
Cost Matrix Score	45*5+21*1	246

Reduced Variable Random Forest Confusion Matrix and Cost Matrix Score



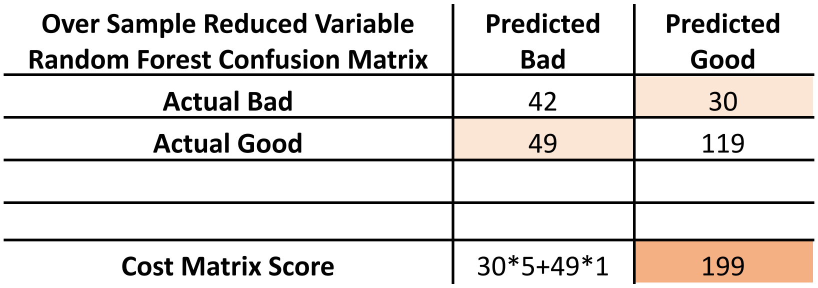
Reduced Variable Naïve Bays Confusion Matrix	Predicted Bad	Predicted Good
Actual Bad	40	32
Actual Good	24	144
Cost Matrix Score	32*5+24*1	184

Reduced Variable Naive Bays Confusion Matrix and Cost Matrix Score



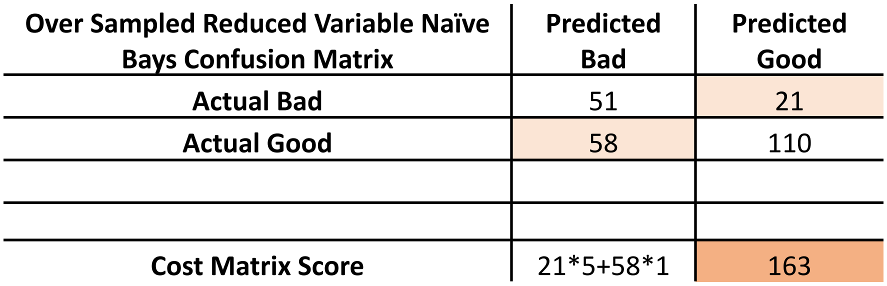
Reduced Variable SVM Confusion Matrix	Predicted Bad	Predicted Good
Actual Bad	14	58
Actual Good	9	159
Cost Matrix Score	58*5+9*1	299

Reduced Variable SVM Confusion Matrix and Cost Matrix Score



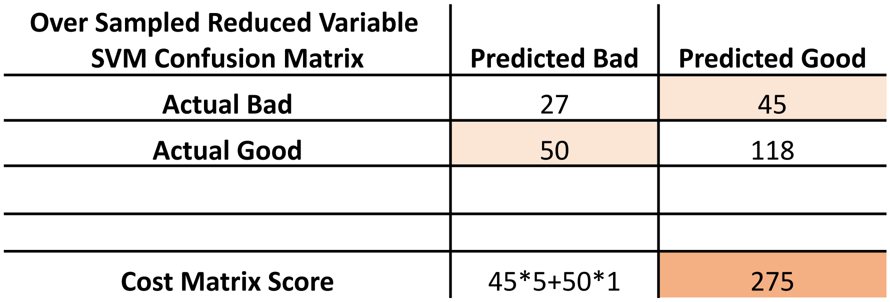
Over Sample Reduced Variable Random Forest Confusion Matrix	Predicted Bad	Predicted Good
Actual Bad	42	30
Actual Good	49	119
Cost Matrix Score	30*5+49*1	199

Oversampled Reduced Variable Random Forest Confusion Matrix and Cost Matrix Score



Over Sampled Reduced Variable Naïve Bays Confusion Matrix	Predicted Bad	Predicted Good
Actual Bad	51	21
Actual Good	58	110
Cost Matrix Score	21*5+58*1	163

Oversampled Reduced Variable Naive Bays Confusion Matrix and Cost Matrix Score

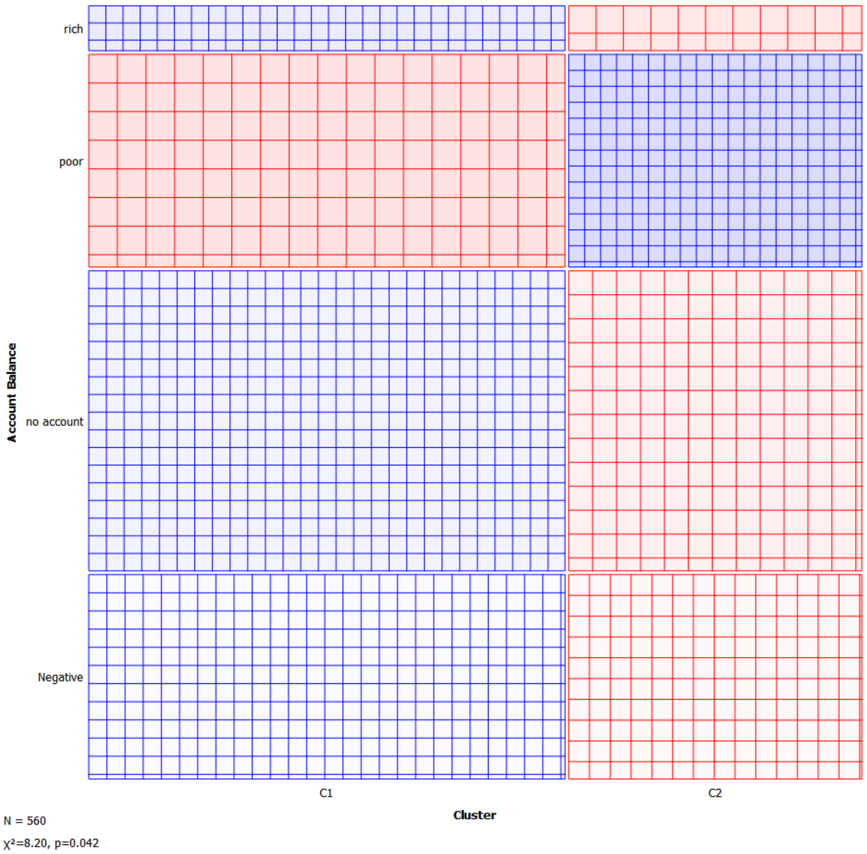


Over Sampled Reduced Variable SVM Confusion Matrix	Predicted Bad	Predicted Good
Actual Bad	27	45
Actual Good	50	118
Cost Matrix Score	45*5+50*1	275

Oversampled Reduced Variable Naive Bays Confusion Matrix and Cost Matrix Score

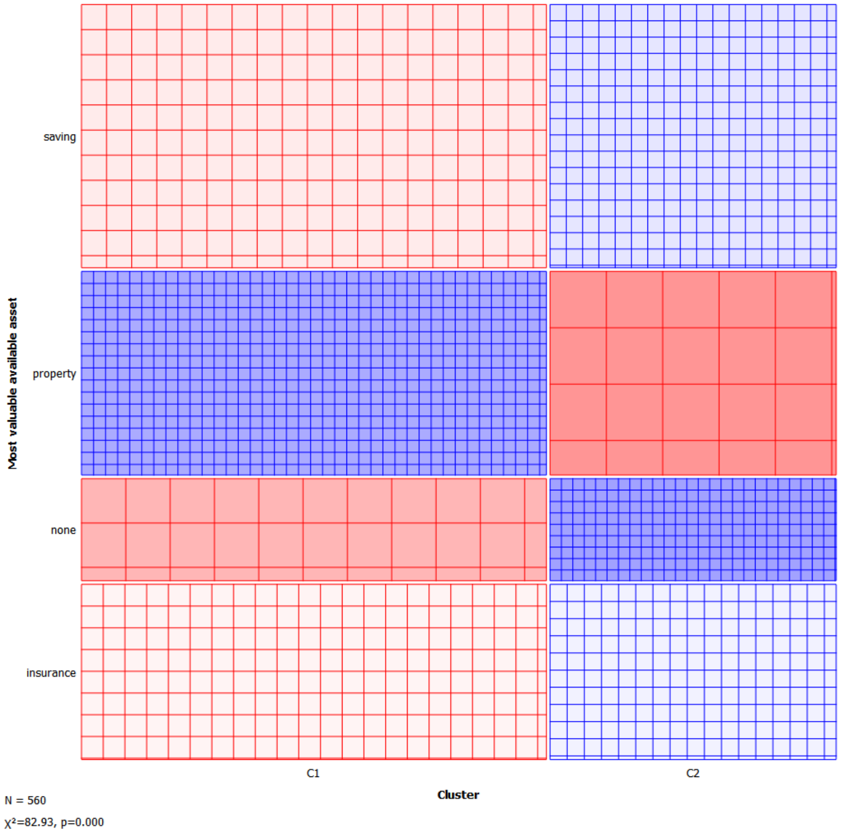
Appendix 2: Sieve Diagram

Account Balance



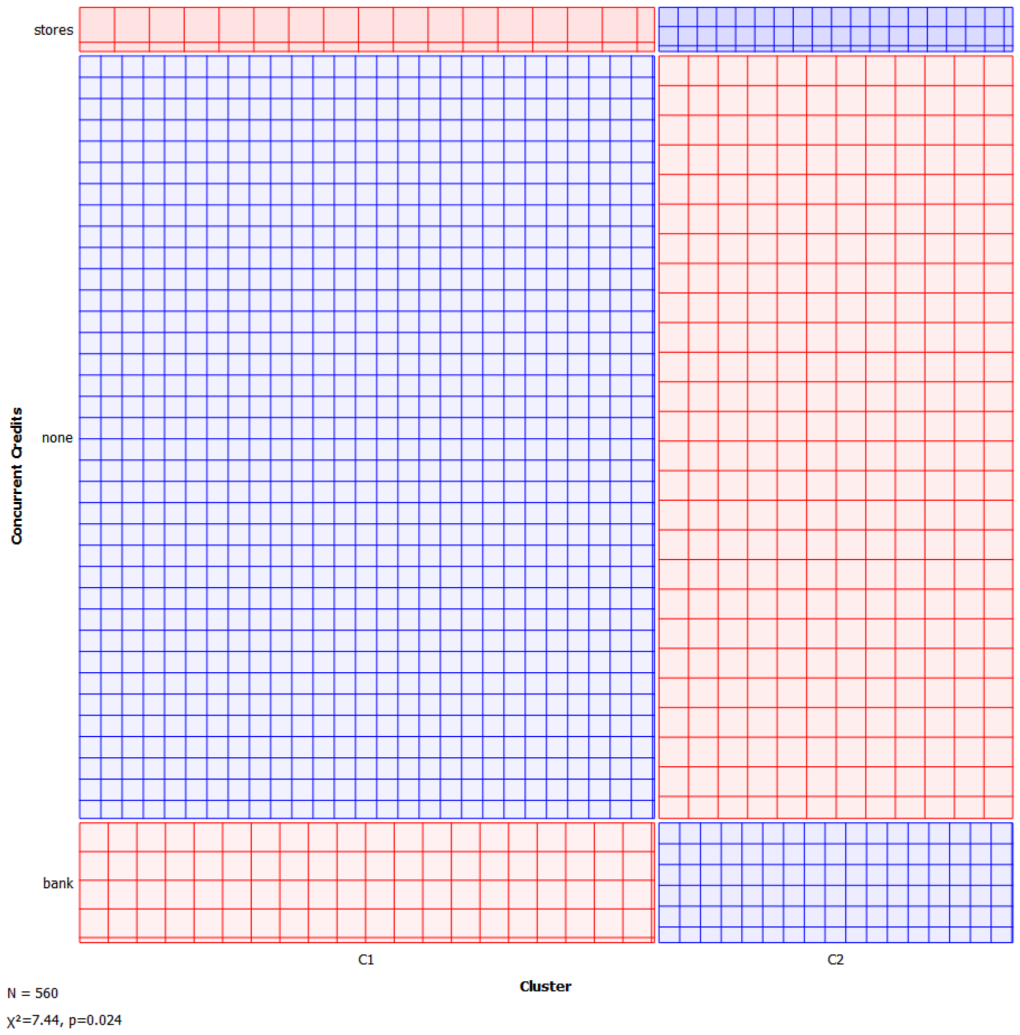
Cluster to consists of people with a poorer Account Balance whereas people in Cluster 1 tend to have either a negative balance or are extremely rich or have no account.

Most Valuable Asset



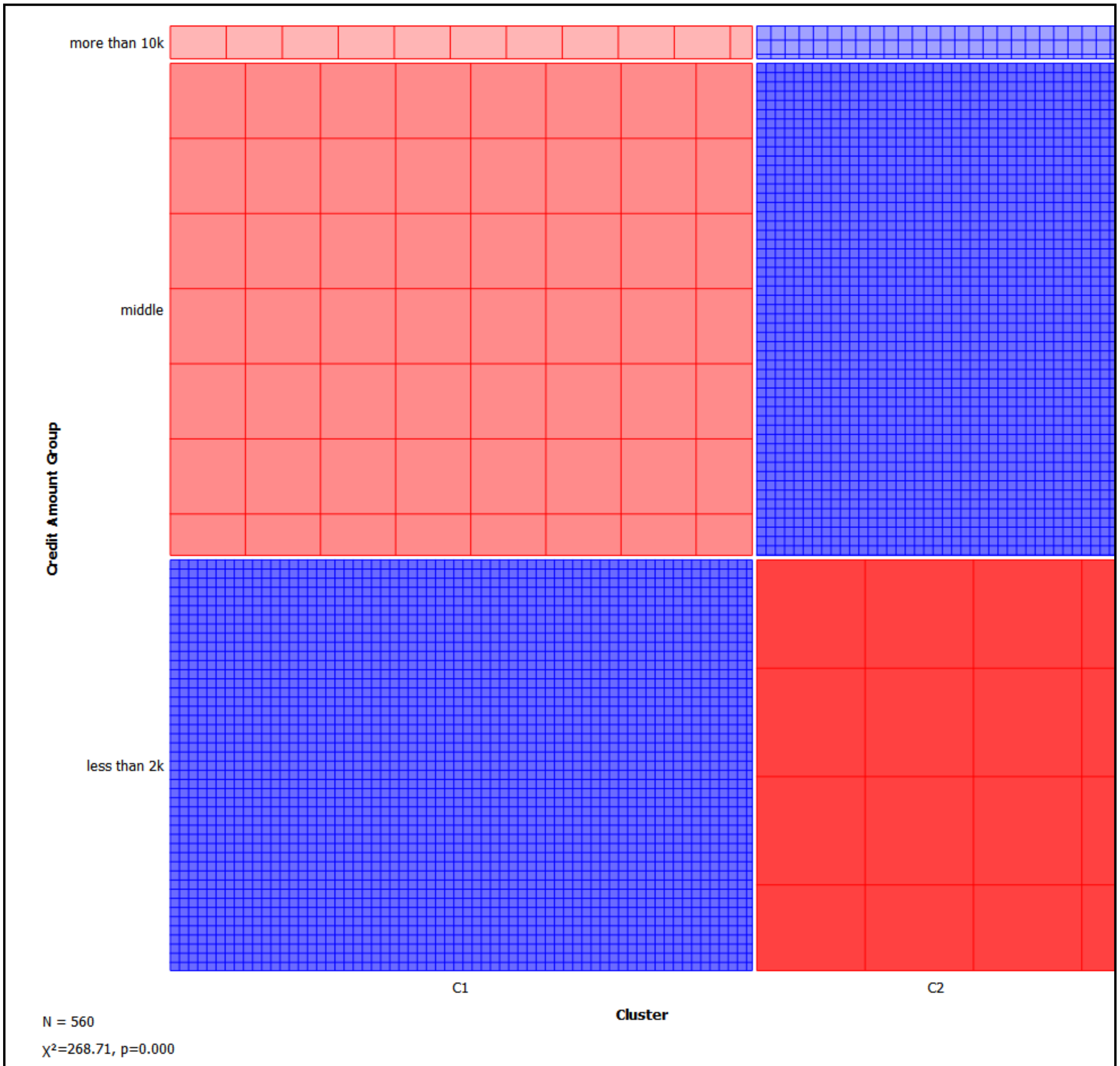
Cluster 1 seems to have property as the most valuable asset in contrast to cluster 2 which does not have any. Though the variables are highly correlated because of high X^2.

Concurrent Credits



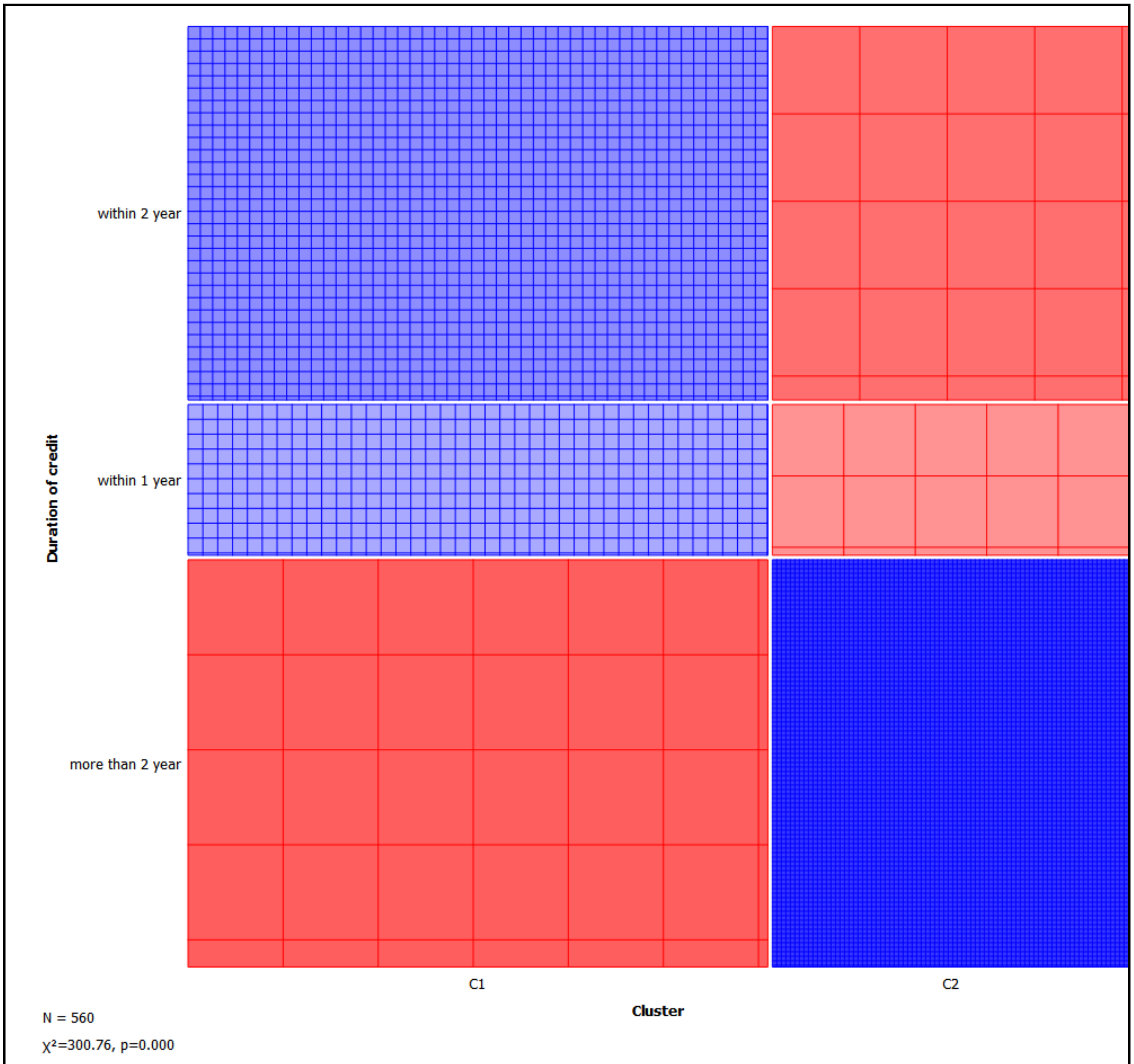
Cluster 1 incorporates people with no instalment plans in regard to people in cluster 2 who have some sort of plan with banks and stores. The difference is significant as p-value is less than 0.05

Credit Amount



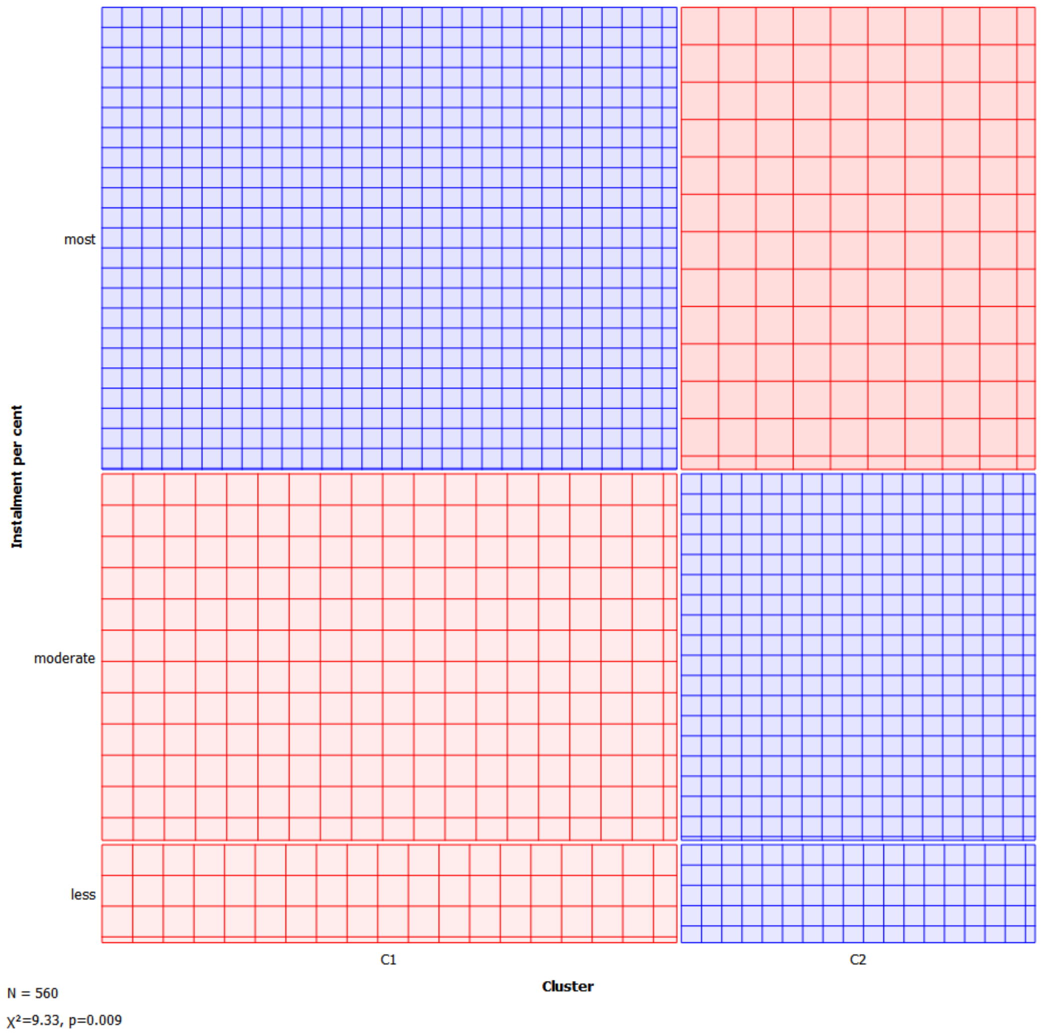
A strong relationship exists between cluster and credit amount group. People with credit less than 2k are likely to be a part of cluster 1 as opposed to those who have credit more than 2k, likely to fall in cluster 2.

Duration of Credit



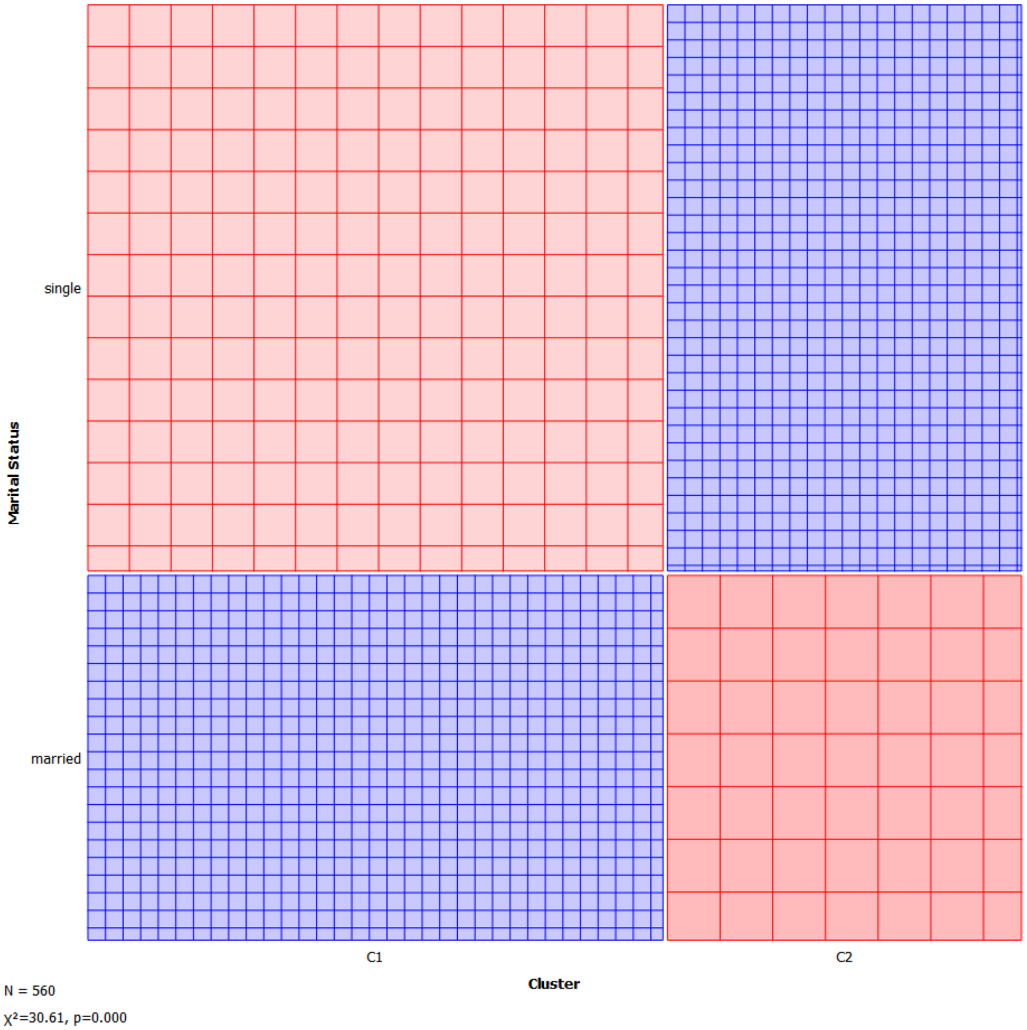
There is a significant correlation between cluster and duration of credit. Cluster 2 consists of people who took credit for more than 2 years and people in cluster 1 likely to payback within 2 years.

Instalment per cent



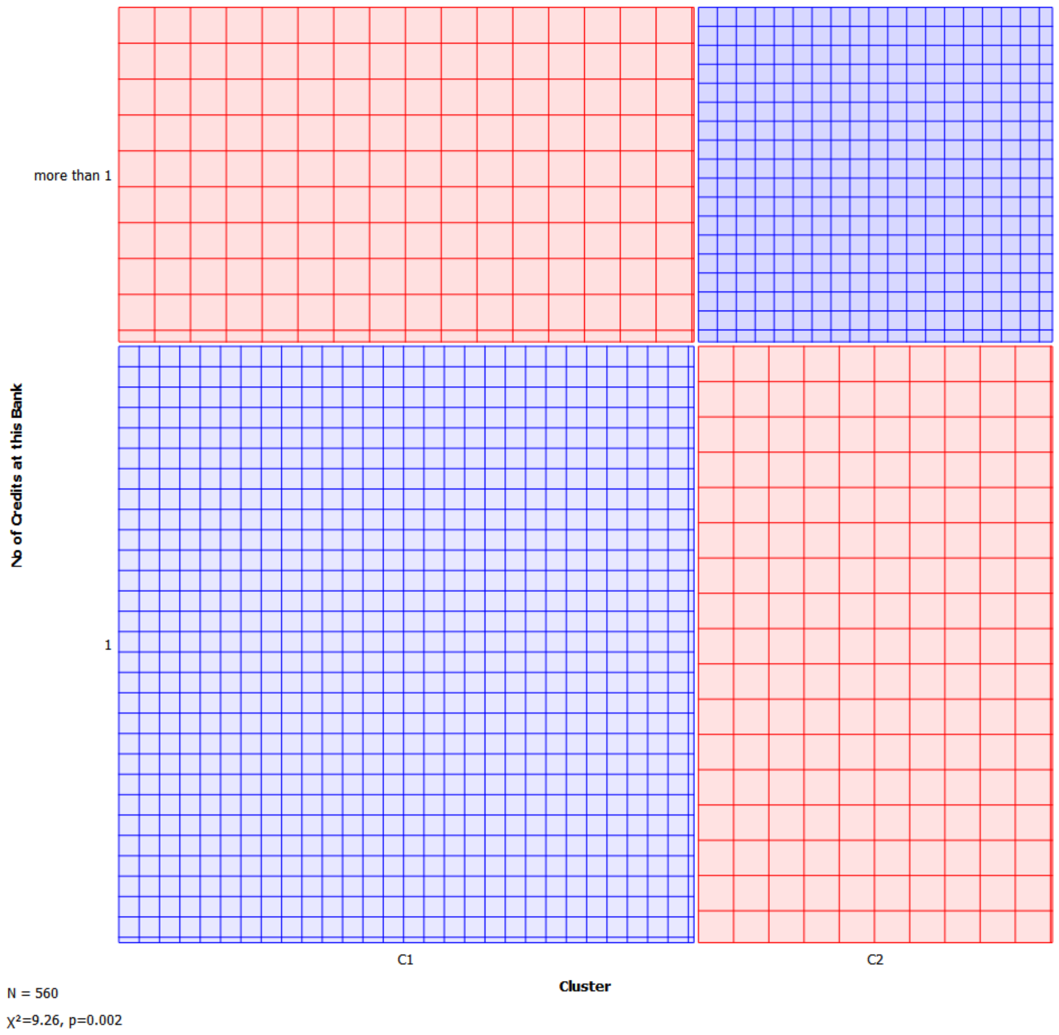
Cluster 1 contains people who have to pay a high amount of their disposable income as interest as cluster 2 contains people who do not.

Marital Status



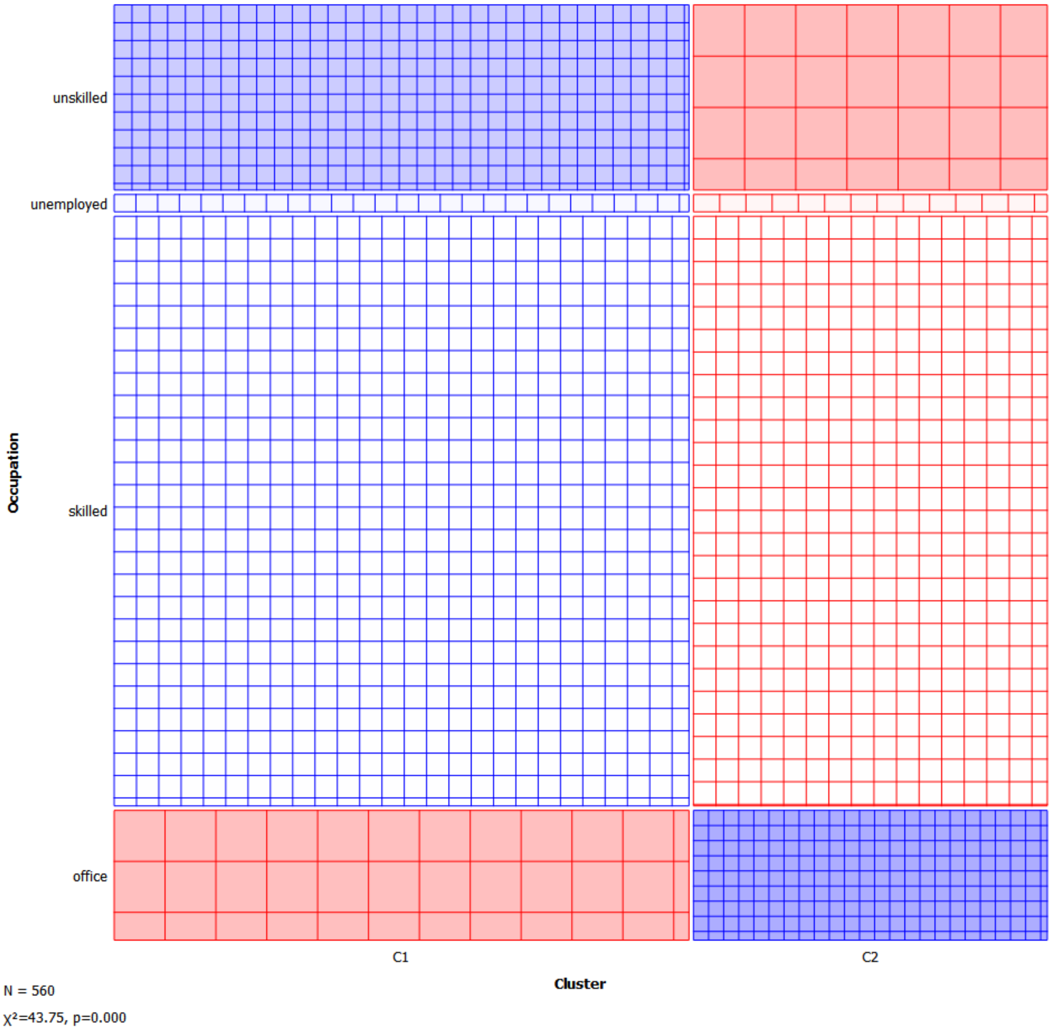
Cluster 1 mostly consists of married people whereas Cluster 2 mostly consists of single people.

Number of Credits



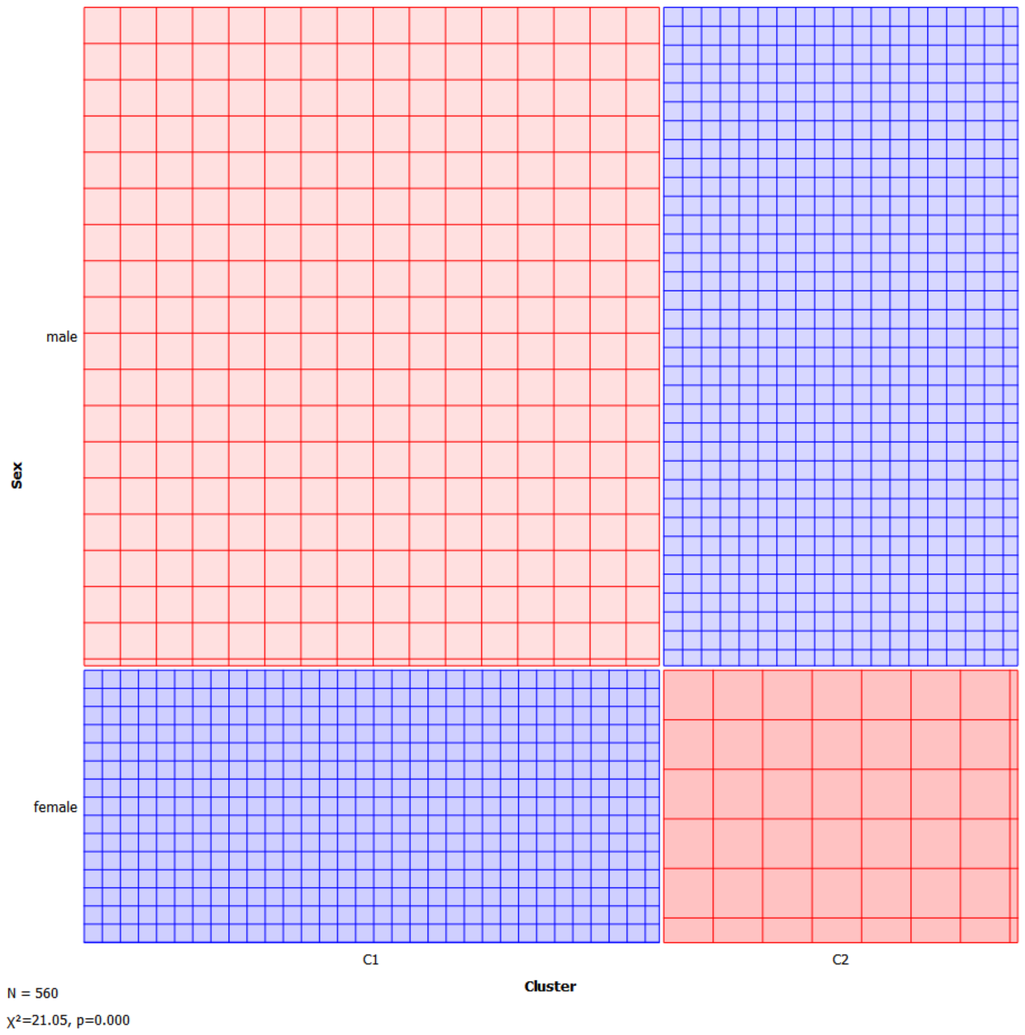
People who already have a credit at this bank lie in cluster 1 in contrast to people of cluster 2 having more than one credit.

Occupation



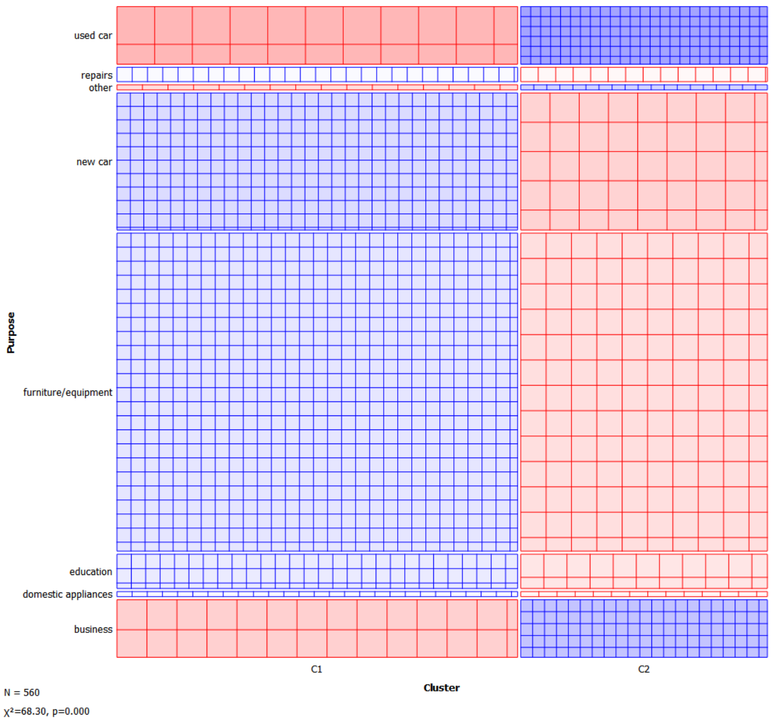
Cluster 1 comprises people who are skilled and unskilled both as opposed to cluster 2 which only have office workers

Sex



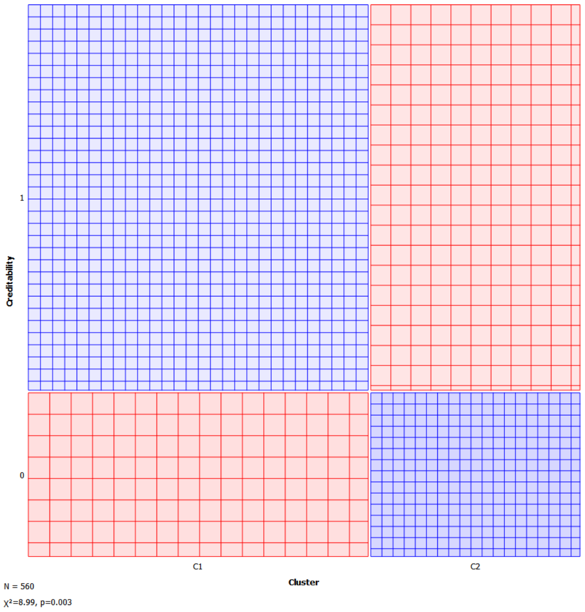
Relatively, cluster 1 has more females and cluster 2 has more males in them.

Purpose



Cluster 1 includes a broad category of people who want to acquire a loan for domestic appliances , education, furniture/equipment, new car or repairs whereas cluster 2 includes people in demand of loan to buy a used car or for business purposes.

Credibility


Credibility
C1
C2
Cluster
N = 560
χ²=8.99, p=0.003