

BUSA3020 - Advanced Analytics Techniques
Assignment 4: Group Project
by John Allagiannis (45410399), David Kozul (45410429), Adina Varkey
(45342148) & Quan Dang (45240221) (Group_19)

Table of Contents

Executive Summary	2
Introduction	2
Exploratory Data Analysis	2
Predictive Analysis	3
<u>Which model is the most accurate that would minimise financial loss?</u>	3
Data Treatment	3
Predictive Model Accuracy	3
Risk Percentage Based on Proportion of Predicted	4
Conclusion	4
Cluster Analysis	5
<u>Are there any distinct market segments among the bank's customers?</u>	5
Data Reduction	5
Clustering Algorithms	5
k-Means Clustering Results	5
Conclusion	6
Segmentation	7
<u>Can predictive analysis be improved by using the 8 market segments identified in the cluster analysis?</u>	7
Data Sampling	7
Conclusion	7
<u>Can predictive analysis be improved by using the 3 groups of segments identified in the cluster analysis?</u>	7
Data Sampling	7
Predictive Model Accuracy	8
Conclusion	9
Recommendation	9

Executive Summary

The German bank should use the Neural Network predictive analysis model to minimise the business loss risk and financial loss risk when approving applicant loans. The bank should also aim to maximise its profits by targeting 3 main groups: young and employed married women, older single men with established careers and young to mid 30s adult males who have paid off their education debt and own their homes. Furthermore, clustering does not improve prediction as the predictive models for the whole sample were more accurate than the individual cluster predictive models. In the future, it may be best to explore different clustering models as they may provide better results or improve prediction accuracy.

Introduction

The German bank wants to maximise their profits while minimising loaning risks such as loaning to a customer who will default on the loan or not loaning to a customer who will repay the loan in a timely manner.

We are going to use demographic and socio-economic data of 1000 loan applicants to predict and minimise risks to aid the bank's management in approving loans. We are also going to discover market segments that the bank can target for better services and maximising their profits.

Exploratory Data Analysis

Univariate Analysis was used initially to analyse the German credit data. We found no missing values in the dataset and made adjustments to certain variables. These adjustments included dividing the "Sex & Marital Status" variable into "Sex" and "Marital Status" and combining some of the loan purposes from the "Purpose" variable. The "Foreign Worker" variable was also relabelled as it makes more sense that the bank is dealing with more domestic than foreign customers.

We then used Bivariate Analysis and discovered that "Credit History", "Other Debts", "Housing", "Foreign Worker", "Loan Amount" and "Age" are variables in the dataset that influence the response variable "Creditability".

Predictive Analysis

Which model is the most accurate that would minimise financial loss?

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. To minimise the bank's losses, we used predictive analytics to produce the optimal model the bank should use to accurately class their customers.

Data Treatment

The data was randomly split in a 70-30 ratio where 70% is the training data and 30% is the test data. The models that we used were Decision Tree, Random Forest, Logistic Regression, Naïve Bayes and Neural Network in Orange Data Miner.

Predictive Model Accuracy

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.809	0.776	0.769	0.772	0.776
Random Forest	0.819	0.762	0.741	0.773	0.762
Logistic Regression	0.822	0.752	0.740	0.748	0.752
Naïve Bayes	0.787	0.736	0.734	0.733	0.736
Decision Tree	0.652	0.696	0.696	0.695	0.696

The F1 score should be considered as it is a weighted average of Precision and Recall as we also want to minimise business loss. According to the F1 comparison, **Neural Network** and **Random Forest** are the best models for predictive analysis.

Risk Percentage Based on Proportion of Predicted

Model	Losing Business Risk	Financial Loss Risk
Neural Network	32.3%	20.6%
Random Forest	30.1%	23.2%
Logistic Regression	35.3%	21.4%
Naïve Bayes	35.7%	19.4%
Decision Tree	47.6%	24.6%

Conclusion

The aim of the predictive analysis is to find the model which reduces the financial loss risk. The financial loss risk comes from the predictor above when the applicant is a bad credit risk but the bank model predicted them to be good.

According to the two tables, Decision Tree is the worst model to use for predicting loan risks because it has the highest financial loss risk at 24.6%. Hence, we should prioritise the other models. Naïve Bayes would be the best model to minimise the financial loss risk, with a risk of 19.4%. However, the trade-off is the higher risk of losing business, which is the highest of the four usable predictive models, at 35.7%. Therefore, the **Neural Network** model would be the best overall model as it has the best balance and lowest values for the risks. The risk of financial loss is 20.6% and the risk of losing business is 32.3%.

Cluster Analysis

Are there any distinct market segments among the bank's customers?

To determine if there are any distinct market segments in the German bank's credit database, we used a cluster analysis to find different clusters of people. Based on the characteristics of the people in each cluster the bank can deliver services that better provide for their needs.

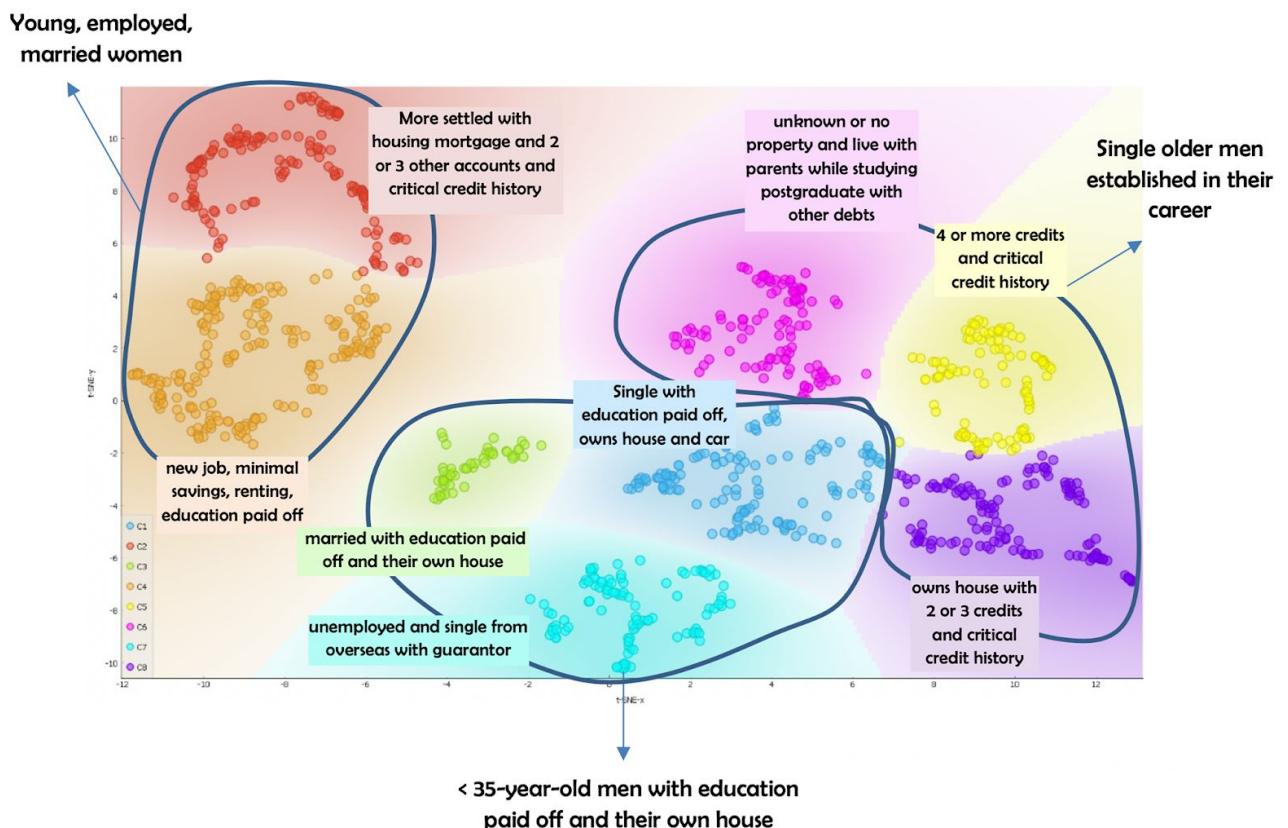
Data Reduction

Before running the cluster analysis, we decided to reduce the data to a smaller number of variables to simplify the analysis by using a data reduction technique called **t-SNE**. It is a Manifold Learning technique which can be used with both categorical and numerical variables. We also decided to standardise the variables to give equal value to each of the variables and remove noise.

Clustering Algorithms

Using the t-SNE coordinates, we ran the cluster analysis using different clustering algorithms. **k-Means Clustering** is the clustering algorithm that produced the most distinguishable and meaningful clusters, followed by Hierarchical Clustering which also produced similar results. DBSCAN and Louvain Clustering were both unable to form easily interpretable or useful clusters.

k-Means Clustering Results



Conclusion

We were able to find **8** distinct market segments among the bank's customers which can be divided into 3 groups:

1) Young and employed married women

2) Men younger than 35 years old who have paid off their education and have their own house

3) Older single men who are established in their career

Thus, the bank can aim to maximise its profits and minimise risk by targeting and providing customised services to different groups of loan applicants.

Segmentation

Can predictive analysis be improved by using the **8 market segments** identified in the cluster analysis?

The single predictive model used for the whole sample is the Neural Network which poses a 20.6% risk of financial loss and a 32.3% risk of losing business. We want to determine if these risks can be further minimised by using separate predictive models for each of the 8 distinct market segments.

Data Sampling

Each of the 8 clusters were sampled using stratified sampling with a proportion of 70% being training data and the remaining 30% being test data. Stratified sampling was used so that the test data would represent the training data well. The models used were the same as above. Unfortunately, due to **insufficient data**, the test and score was not feasible and produced an error.

Conclusion

Segmentation groups need to have more profiles in order to determine if predictive analysis can be improved using separate predictive models.

Can predictive analysis be improved by using the **3 groups of segments** identified in the cluster analysis?

Since some of the 8 clusters did not have enough profiles for the predictive analysis to accurately run, alternatively, we could use the 3 groups identified above.

Group 1: Cluster 2 and 4

Group 2: Cluster 1, 3, and 7

Group 3: Cluster 5, 6, and 8

Data Sampling

Each group was sampled using stratified sampling with a proportion of 70% being training data and the remaining 30% being test data. Stratified sampling was used so that the test data would represent the training data well.

Predictive Model Accuracy

The test and score settings were set the same for all 3 groups: cross validation with 10 folds and stratified, the target class was the average over classes. Since we want to minimise both the financial risk and business loss risk, **F1 scores** are ideal for model comparison because they are a weighted average of Precision and Recall. **Precision** is also helpful because we want to prioritise minimising our model's false positives for good loan credibility since financial loss is a worse risk.

Group	Model	F1	Precision
1	Naïve Bayes	0.692	0.715
	Random Forest	0.677	0.678
	Logistic Regression	0.672	0.679
	Neural Network	0.654	0.652
	Decision Tree	0.621	0.623

Group	Model	F1	Precision
2	Logistic Regression	0.755	0.754
	Neural Network	0.732	0.728
	Random Forest	0.716	0.718
	Decision Tree	0.692	0.709
	Naïve Bayes	0.677	0.718

Group	Model	F1	Precision
3	Neural Network	0.734	0.733
	Logistic Regression	0.719	0.715
	Random Forest	0.694	0.695

	Naïve Bayes	0.688	0.709
	Decision Tree	0.681	0.687

Conclusion

The most effective model for each group is different from one another, however, the predictive accuracies were inferior to the initial overall predictive model.

Recommendation

The final recommendation for the bank to minimise the financial loss risk and business loss risk is to use the Neural Network predictive analysis model. The bank should focus on 3 main groups for customised services: young and employed married women, young to mid 30s adult males who own their homes and have paid off their education debt and older single men with established careers. The individual cluster predictive models were not as accurate as the overall models so the bank should use the single overall Neural Network model for predictive analysis.

For future research, we recommend exploring different clustering models for better analysis.