Data Minig Project

Analysis of loan customers' characteristics

by

Burcu Kalender

Abstract

The target of this study is to provide the marketing division of a young bank with information to set up a new campaign to gain more loan customers. The specific questions of interest are what combination of parameters makes a customer more likely to accept a personal loan and are there any association among special offers as online services, security accounts, credit cards that support cross-selling opportunities. The data mining techniques employed are explanatory data analysis, entropy classification trees, neural networks, 4-means clustering and principal component analysis. The analyses yield three different characteristics of individuals who are likely to take a loan. One group named “new generation of self-made man” contains young people with high income, an undergraduate level of education and a high credit card spending. A second group are the “open minded people” who have a high level of education, a high income and are interested in different facilities of the bank. A third group build the “conservative”, they are highly educated, live in a single household and have no interests in additional bank services.

Table of contents

I Table of Figures and Tables

1. Introduction

The target of this study is to provide the marketing division of a young bank with information to set up a new campaign to gain more loan customers.
This study is about a young bank that is growing rapidly in terms of overall customer acquisition. The customers of the bank are divided into two major groups. The first one is the liability customers, which build the biggest group. A liability customer deposits money into an account at the bank, which the bank has to pay back when requested by the customer. Usually, the bank gives a small amount of interest for deposited money. The second group is the personal loan customers. Those are customers borrowing money from the bank. Under a concluded contract, the customers are obliged to return the money back, with an additional interest rate. This interest rate is greater than the one given on a deposit. Thus, a loan is a source of income for the bank and they are interested in raising the number of loan customers. Moreover, the bank aims to convert there liability customers into loan customers. A campaign the bank ran for liability customers last year showed a conversion rate of over 9% successes. An overall objection is to find a connection between the variables and an enhancement of loan customers based on the data of the previous campaign.

2. Materials and Methods

This chapter describes the features of the data and the variables as categories as well as units. After knowing the data better, the aim is defined more precisely. Moreover, the data mining techniques to be used are presented and an extension to not used techniques is shortly given.

2.1. Data Description

The data set includes 5000 observations with fourteen variables divided into four different measurement categories. The binary category has five variables, including the target variable personal loan, also securities account, CD account, online banking and credit card. The interval category contains five variables: age, experience, income, CC avg and mortgage. The ordinal category includes the variables family and education. The last category is nominal with ID and Zip code. The variable ID does not add any interesting information e.g. individual association between a person (indicated by ID) and loan does not provide any general conclusion for future potential loan customers. Therefore, it will be neglected in the examination.

Table 2.1: Data description

Personal Loan	Did this customer accept the personal loan offered in the last campaign?
Securities Account	Does the customer have a securities account with the bank?
CD Account	Does the customer have a certificate of deposit (CD) account with the bank?
Online	Does the customer use internet-banking facilities?
Credit Card	Does the customer use a credit card issued by Universal Bank?
Age	Customer's age in completed years
Experience	years of professional experience
Income	Annual income of the customer ($000)
Family	Family size of the customer
CC Avg	Avg. spending on credit cards per month ($000)
Education	Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
Mortgage	Value of house mortgage if any. ($000)
ZIP Code	Home Address ZIP code.
ID	Customer ID

After introducing the data variables, the research aim can be defined more specifically:
1) What combination of parameters makes a customer more likely to accept a personal loan?
2) Are there any association among special offers like online services, security accounts, credit cards, etc. for finding cross-selling opportunities?

2.2. Data Mining Methods

This part presents the five data mining techniques used i.e. EDA, classification trees, neural networks, cluster analysis and principal component analysis in detail. The idea as well as the process of those methods is explained and justifications for the use are given.
Exploratory data analysis is a very useful method to get to know the data before delving into analysis that is more advanced. It is important to know the data to avoid mistakes in the analysis e.g. if we know that the data does not fulfil assumptions that a method requires than

this method cannot be used straightforward as the results will not be reliable. The distribution of the variables is examined, in this course the mean, variance, normality and symmetry are important, also possible transformation to maximise normality. Moreover, associations not only between the variables but also between those and the target variable can be discovered. Thus, we try to find any correlation among the variables applying graphical tools.
Next, we use classification trees to predict the target variable personal loan with the independent variables and discover associations. The decision tree utilise the variables to separate the two groups loan and non-loan taker, starting with the variable, which distinguishes the target variable the best. Each division ends in a decision node; there are more and more nodes created until there is no variable left, which is able to significantly separate the response variable. There are three different measurements to compute the separation; those are Gini, Entropy and CHAID. We need to analyse, which of them yields the smallest error rate. For this and the following methods, it is best to partition the data into training, validation and testing sets. Classification trees have some advantages; they are easy to interpret, errors in the data are unlikely to affect the result and they automatically remove unnecessary variables. Furthermore, this method can handle interactions between the variables well.
The idea of neural networks method is based on the functionality of the neuronal networks in the brain. The neurals are connected with each other and learn from experiences (recognized patterns), which the model can use to make future predictions. In this report, we use a network with multiple inputs, one hidden layer and a single output. The hidden layer employs an activation function consisting of a combination function i.e. a weighted sum of input variables and a transformation function. For the transformation function, we choose a logistic function, as the target is a binary variable. The advantage of this method is that it can handle linear relations as well as interactions. Hence, the prediction accuracy is high and the results robust. The disadvantage is the complexity of the model. It is not always easy to understand how the result is evaluated and which variables are important to distinguish between the response variable. Nevertheless, with a few hidden layers and nodes, the results can still be clear to understand.
Cluster analysis is used to group a set of data objects into clusters, where the similarity of individuals in the same cluster and difference to the individuals in other clusters is maximized. The aim is to find a cluster, which includes mostly loan taker, so we get an insight in the characteristics of this group. The results depend on the method of distance measurement. There are different measurements for different kind of variables; our data contains interval, ordinal and binary variables. On the interval and ordinal ones SAS applies the same measurement but a different measurement for the binary variables. Prior the clustering, it is necessary to standardize the data, so that each difference of the variables contributes equally to the overall distance value. For building the clusters we use K-means clustering, the reason for this is given in the analysis part. We first have to decide on a number of clusters on our own, then the initial centres are chosen and the observations are allocated to the closest centre, which is repeated until the clusters do not change anymore. The universal bank data set has 5000 observations and no nominal data; hence, it is of advantage to use this method as it can handle big data sets well but no nominal data. A drawback is that the number of clusters have to be predefined by the analyst.
The principal component analysis tries to reduce the number of variables to a lower dimension to make the model easier to interpret. Those dimensions are new variables called principal components, which are linear functions of weighted original variables. It is important to standardize the data prior conducting the method. Otherwise, the computed weights will be false. Each of those uncorrelated components explains some of the variability in the data; our aim is to reduce the dimension to a few components that explain around 80% to 90% of the variability. Then we can interpret those components and probably find some associations to the target variable.

2.3. Further Data Mining Methods

This section is about further data mining methods that are not used in this study. Those are the five techniques association analysis, logistic regression, bundling techniques, memory based reasoning and text mining. They are briefly presented and we explain why those methods are not selected for this study.

The general term market basket analysis covers the two methods association and sequence analysis. Both are useful to find frequent patterns among the variables. The association method is useful to identify, which variables occur together and accordingly creates a rule. The rule is developed by counting how often a variable emerge alone and in combination in the data. In addition to the connection of variables and their probability, sequencing also considers the order in which the relationships occur. Thus, it includes a timing element in the analysis. Overall, the market basket analysis is useful to find out the probability that variables appear together. Unfortunately, this analysis does not give us any results out of two reasons. First, no significant association at a confidence level of 5% could be created for unknown reasons. Second, there is no time element, which is necessary for performing a sequence discovery.

Another useful method to predict the response variable is a regression model. As the target variable loan is binary, we need to use logistic regression, which predicts the odds that the event loan will happen against the probability that it will not occur. This method is based on the assumption of normal distributed variables; thus, it is probably necessary to transform the variables prior building the model. Furthermore, it does not discard variables automatically. Therefore, it is necessary to use a variable selection node or a variable selection method e.g. backward elimination. An advantage of regression models is that it can handle linear relationships between variables well. In general, it is a more precise method than e.g. classification trees because each individual receives an individual output. The reason for not using this method is that the classification trees have a lower error rate. Furthermore, the model does not give any further information that is not already included in the tree model.

The general procedure of bundling techniques is combining results of several models to an averaged output. This average is more accurate i.e. the total error is reduced as the individual errors cancel out and the result is more stable as differences are small for different sets of measurements. The advantage of this technique is improved predictions when scoring new data. The disadvantage is that the results are harder to interpret. Regarding our aim to identify important variables to decide who is likely to be a loan-taker, we need to be able to interpret the results of an analysis and are less interested in scoring new data. Thus, the benefit of this method to score new data well is not useful for our purpose and we do not use bundling as an analysing tool.

Memory based reasoning uses the K-nearest neighbour method to make prediction for new data. For binary target variables, this method searches a local area of predefined K numbers of neighbours and allocates the new object to the closest neighbour. In terms of our target, the disadvantage is again that this is a predictive method and does not help to find important variables to explain the characteristics of loan taker.

Obviously, text mining is used to detect patterns in articles or other written documents and therefore is of no use for the universal bank data set.

NOTE: The parts 2.2 and 2.3 refer mostly to the “Practical Data Mining” study guide.

3. Results and interpretation

This part shows the results from the SAS analysis for the five techniques. Moreover, the results are interpreted primarily in terms of characteristics of loan taker.

3.1. EDA

This section provides a first close look at the data so we get to know it better. We want to know how the variables are distributed, what the average values are, the proportions for the ordinal variables, if there is something in the data that we should be aware of and what kind of relationships between the variables exist.
The histograms in the appendix A show the distributions and the transformed distributions maximising normality for the interval variables: age, experience, income, cc avg, mortgage as well as the ordinal variables family and education. Each of those variables shows a high skewness, kurtosis or both, therefore a transformation improves the variables.
The variable age ranges between 23 and 67 with approximately same percentage of people for the different ages. The square root transformation improves the kurtosis but the skewness gets worse. In the histogram for years of experience it appears that, there are negative values; -1, -2 and -3. This could be a data input error as in general it is not possible to measure negative years of experience. However, the proportion of those values in the data is below 1% and as we cannot detect the reason for those negative values, we should delete them. Regarding the variable income, the minimum value is $8000 and the maximum value is $224000. The majority of individuals have an income between $20000 and $90000. As typical for financial data it is right skewed, thus a transformation is able to improve the level of skewness from 0.84 to -0.08. The average credit card spending has a wide range of $0 to $10000 per month; the majority spends less than $2500. A log transformation improves the skewness level. Concerning the mortgage, 70% of the individuals have a mortgage of less than $40000; however, the maximum mortgage is as high as $635000. A square root transformation can improve the skewness from 2.1 to 1.2 and the kurtosis from 4.76 to only 0.03. The variables family and education are ordinal variables but for the EDA we treat them as interval variables so we can use them in the EDA. Thirty percent of the people in this study have a single household, 27% count for two people, 20% for three person and 23% for four person families. Thus, the distribution of families is evenly distributed. Regarding education, 40% are at undergraduate level, graduate and professional levels each count for 30% approximately.
The Box plots in Fig. 3.1 show the relation between the target variable personal loan and the explanatory variables family and cc avg. The plot on the left hand side indicates that families with a median of size three are more likely to take a loan (loan-taker = 1, non loan-taker =0). Single households or families with two people are less likely to take a loan. This might be a useful association when considering a future campaign e.g. the campaign could aim on families with children. We should mention that the mean is not useful for the comparison because of the skewness in both distribution the mean values of family size are similar. Concerning the influence of average credit card spending, the loan and no loan distributions show a clearer distinction as the distribution do not overlap much. In general, a higher average credit card spending with a median of $3800 indicates a higher probability of personal loan. Lower credit card spending with a median of $1400 is less likely to take a loan. This could be useful information when selecting people for example to mail loan advertisements. For the remaining variables, the box plots do not give distributions that are helpful to distinguish between loan takers or non-loan taker; the box plots overlap too much within each variable.

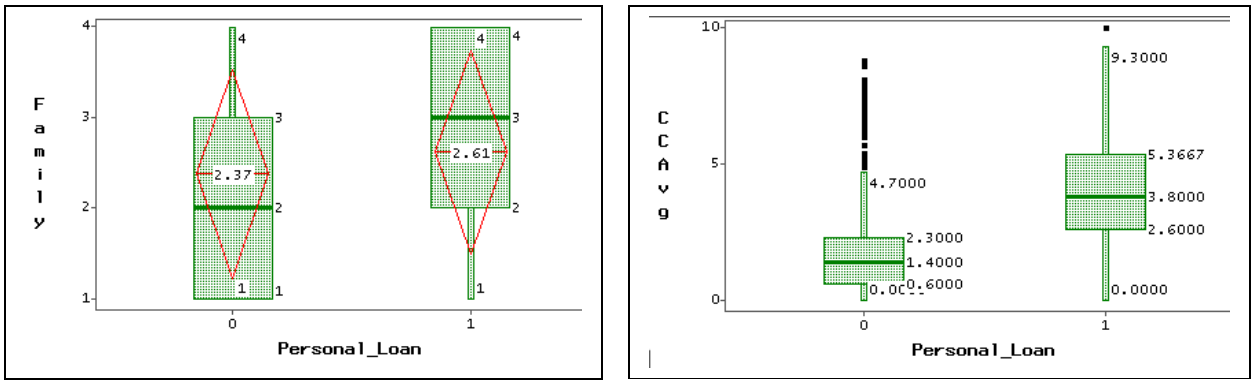


Figure 3.1: a) Box plot of family and personal loan b) CC avg and Personal loan

Fig. 3.2 shows a scatter plot, which present the relationship and correlation among the variables experience and age. It indicates that years of work experience and age have a positive correlation, which seems to be reasonable. Moreover, we recognise some kind of grouping, education level three (professional, in black) distinguishes itself from education level two (graduate, in red) and one (undergraduate, in green). Level three has the same positive correlation but overall fewer years of experience. Maybe this group spend more years on education and thus has a shorter period of working experience. Furthermore, there is a gap for the professionals in their mid-forties, probably those people are not included in the study and therefore missed out. The majority of individuals in the study are from undergraduate level, they have the lowest education level but the most years of experience.

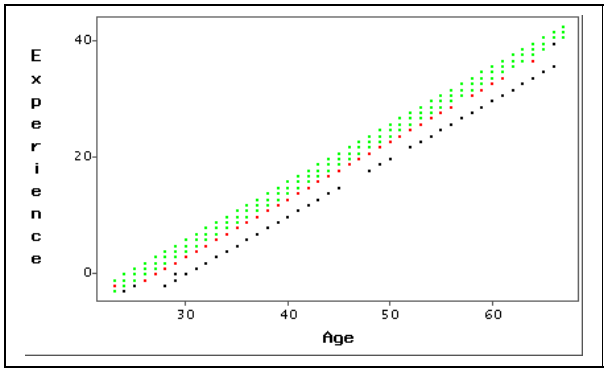


Figure 3.2: Scatter plot of age and experience

The next Fig.3.3, the scatter plot for income, cc avg and mortgage respectively is shown. The relation between credit card average and income varies between no relations to positive relation. The general statement for the positive correlation is that a higher average credit card spending tends to indicate a higher income. Individuals who earn less have a limited credit card spending. However, a high income does not necessarily suggest a high credit card spending. Unfortunately, no explanation could be found, which factor has an influence on the spread of no use and high use of credit card.

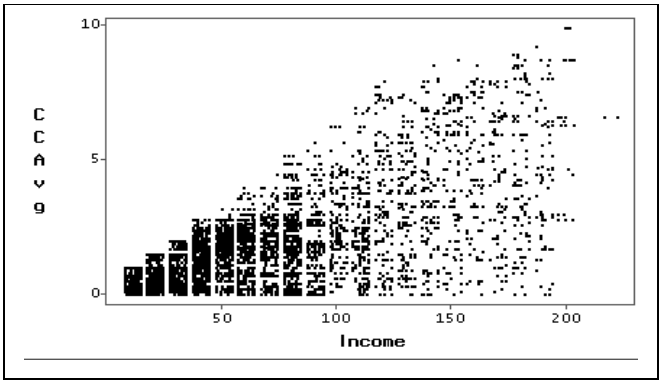


Figure 3.3: Scatter plot of income and CC avg

If people with a high income tend to spend more with their credit cards and a high credit card, spending indicates a higher probability of being a loan taker than there might be an indirect relationship between income and loan. When we group the scatter plot for loan (loan taker are indicated in red), we receive an interesting result. Fig. 3.4 suggests that people with a high credit card spending above $4000 and an income of about $100000 take a loan. Above an income of $100000, people take a loan independent of how much they are spending with there credit card.

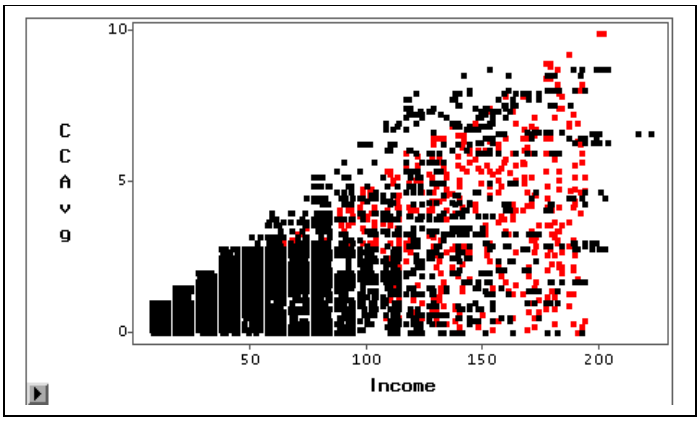


Figure 3.4: Scatter plot for income and CC avg grouped for personal loan

Maybe it is reasonable to check whether we get similar results between family, loan and income. The box plots in Fig. 3.5 show again an interesting result. It seems like that not the size of family is really influencing the likeliness of a loan but it is more the amount of income. Families with low income below $100000 are less likely to take a loan than families with high income regardless the size of family. We obtained similar result for credit card average.
Nevertheless, we do not know which of those two results (family size has or has no influence) is the more appropriate one, but maybe in the further analysis we will find out more about this association.

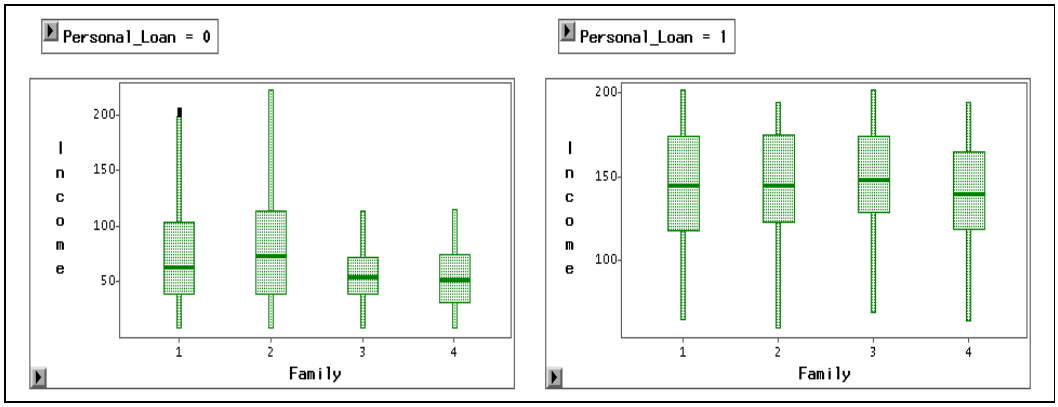


Figure 3.5: Box plot of family and income grouped for personal loan

Concerning the correlation between mortgage and income, we observe the same pattern as for the credit card average (Fig. 3.6). The relation fans out between no relations to highly positive correlation between income and mortgage. The positive relation suggests that a high mortgage implies high income. We could not find an explanation for the fanning out. In addition, there is a gap visible between zero dollar and $75000; this is because the minimum value of house mortgage is about $75000. The red dots again show the loan taker, there is no relationship between mortgage and loan visible but again we can see the same association between income and loan as before.

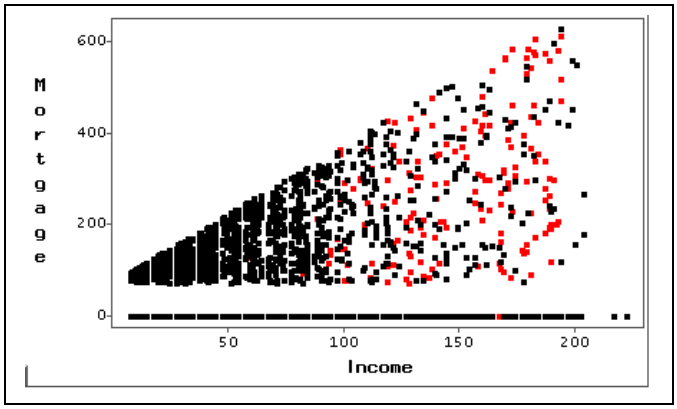


Figure 3.6: Scatter plot of income and mortgage grouped for personal loan

In summation, the EDA gives following results. Overall, the normality transformation improves the skewness and the kurtosis. This will be useful for further investigation when there is the assumption of normally distributed data. From the box plot graphs, we conclude that big families (more than three members) are more likely to take a loan. On the opposite, the other box plot indicates that there is no relationship between loan and family size but a positive association between income and loan. Moreover, we know that a high credit card spending combined with low income is likely to take a loan, whereas the credit card spending is not important when the income is high. In addition to these results, we know that most of the people in this data set have undergraduate level education and most of the people earn between $20000 and $90000.

3.2. Classification models

In this part, we try to predict the target variable personal loan based on the independent variables using entropy classification tree. First, we try to combine different classification trees and logistic regression with variable selection and variable transformation nodes. Then we decide which model to use based on the misclassification rate. Using model diagnosis methods, we see how well the selected model performs.
Before deciding on the use of a classification tree, we examined different combinations of trees (and logistic regression) with variable selection and transformed variable nodes. The result is that not using any variable selection and no transformation yields the smaller misclassification rates.
Next, we want to decide which specific model to use. The Tab. 3.1 shows the results the misclassification rates for four different models. Overall, the entropy tree with the untransformed variables and without any variable selection node performs the best. It has the lowest misclassification rate of 1.8% (for test data) compared to the Gini and chi-square trees with an error of 2.2% and the logistic regression with an error of 3.2%.

Table 3.1: Misclassification rates for classification trees

Description	Target	Target Event	Misclassification Rate	Valid:Misclassification Rate	Test:Misclassification Rate
Entropy	PERSONAL_LOAN	1	0.015015015	0.0133333333	0.0186418109
Gini	PERSONAL_LOAN	1	0.019019019	0.014	0.0219707057
Tree	PERSONAL_LOAN	1	0.019019019	0.014	0.0219707057
LogReg	PERSONAL_LOAN	1	0.044044044	0.0413333333	0.0326231691

Now, we use the model diagnosis methods confusion matrix and lift chart to check how well the entropy tree performs and where its weaknesses are.
The confusion matrices for the training and validation data sets in Tab. 3.2 show almost the same results. The entire non-loan taker from the data set could be classified correctly; the confusion rate is zero for this group. Regarding the confusion rate for the loan taker, the

value is 12% in the training data set and 11% in the validation set. This rate is not too high and acceptable for this kind of loan data.

Table 3.2: confusion matrix

SOURCE	STAT	PERSONAL_LOAN	==> 0	==> 1	TOTAL
TRAIN	Row%	0	100	0	100
TRAIN	Row%	1	12	88	100
TRAIN	Row%	+	91	9	100

SOURCE	STAT	PERSONAL_LOAN	==> 0	==> 1	TOTAL
VALID	Row%	0	100	0	100
VALID	Row%	1	11	89	100
VALID	Row%	+	91	9	100

The Fig. 3.7 shows the lift chart, which gives information about the profit gained when using the entropy tree. The area between the tree and the baseline is the additional gain we make for the prediction when using the entropy model. For the top 50% of the loan taker in the data, the entropy tree has a profit of 0.2 and above, which is a high value. In conclusion, this fitted model is a good at making predictions for the response variable.

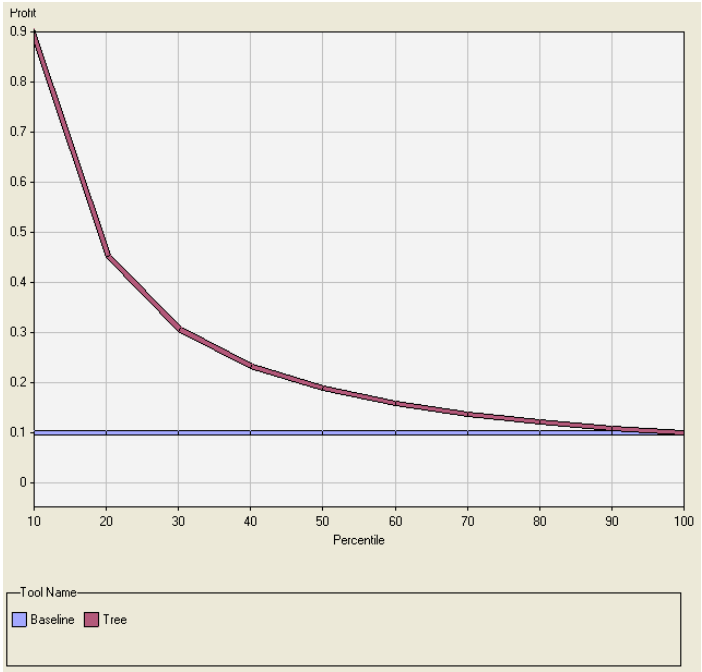


Figure 3.7: Lift chart for entropy tree

Finally, the tree is shown in Fig. 3.8. The most important variable to distinguish between loan and non-loan taker is income. If the annual income is below $94500, the chances of not being a loan taker are 99%. If the income is above that level, the chances of not being a loan taker are still high with 66%. To be able to distinguish between the two levels of loan, we also need to consider education level. If the education level is 1 i.e. undergraduate then the person is likely to be a non-loan taker. However, if the education is at graduate of professional level, the person has a chance of 74% to be a loan taker. In addition, if the person has an income above $116500 then this probability increases to 100%. There are two further possible ending nodes with loan taker as a result. The second path is a person with an income between $94500 and $116500, with an education level of graduate of professional and an average monthly credit card spending above $2725. This combination is a loan taker with a chance of 70%. The third path is a person earning more than $113500, an

undergraduate education level and a family size bigger than 3 do have a probability of 100% to be a loan.

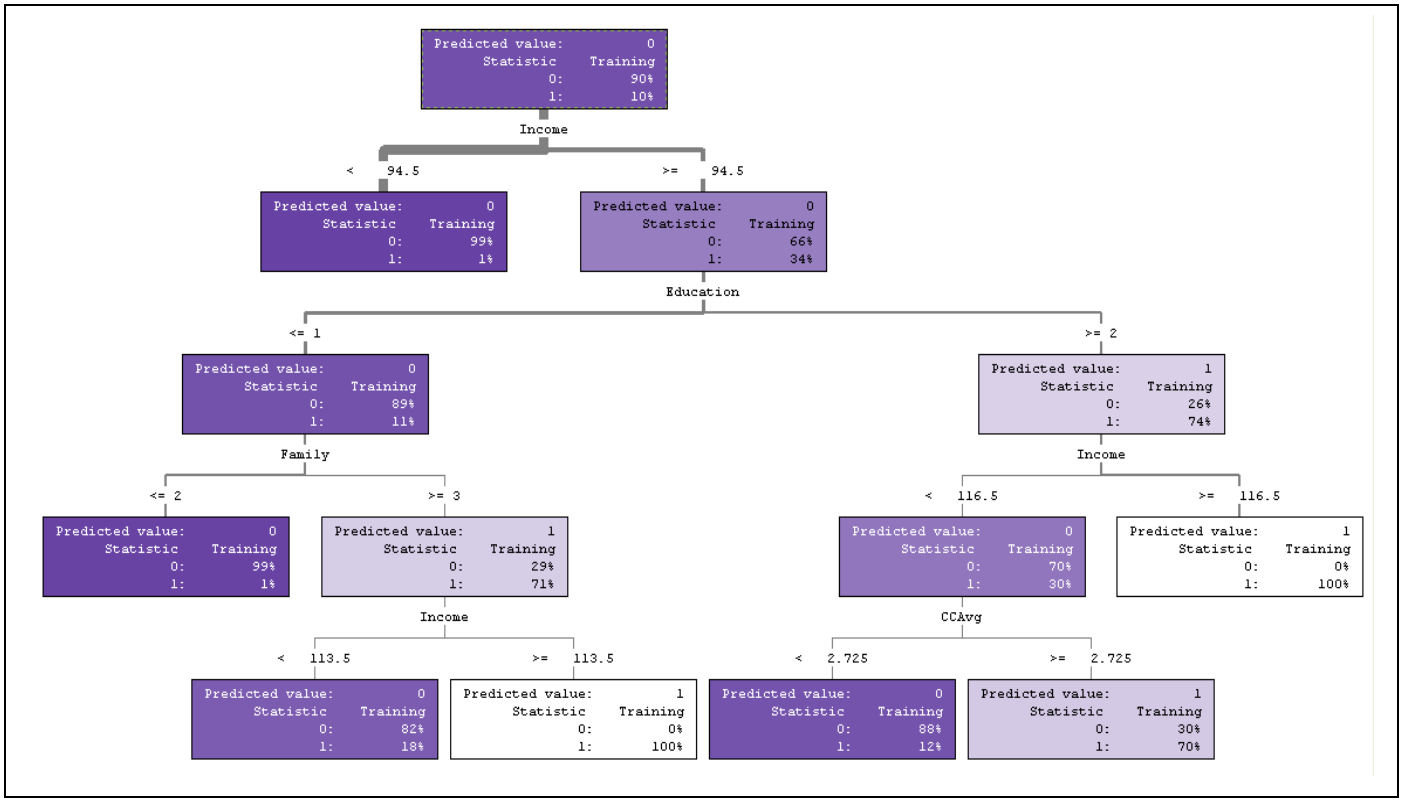


Figure 3.8: entropy tree model

In summary, we used the entropy tree because it has the lowest error rate compared to other possible methods and shows a good prediction power for the two loan cases. The most important variable in the tree model is income. Only individuals with an income above $94500 have a chance to be a loan taker. The variable education is also important. Individuals with a high education level of graduate or professional level are more likely to be loan taker especially in combination with a high average credit card spending. However, individuals with undergraduate education, an income of above $113500 and a big family size are also likely loan taker.

3.3. Neural Network

In this part, we fit a neural network model to the data and try to figure out, which variables are important to predict the response variable personal loan. As the response is binary, we choose a neural network with two nodes and a logistic activation function. Prior the analysis we try to use a variable selection model to discard unnecessary variables. After deciding on a final model, we interpret the nodes in terms of the response personal loan.
The neural network method does not inhibit any build-in variable selection function; hence, we use a variable selection node and a classification tree as variable selection methods prior to the neural network. However, the result in Tab.3.3 suggests that the neural network with the lowest misclassification rate of about 2% is the one using all the input variables i.e. without any prior variable selection.

Table 3.3: Misclassification rate for neural network

Fit Statistic	Training	Validation	Test
[TARGET=Personal_Loan]	.	.	.
Misclassification Rate	0.017017017	0.0153333333	0.0206391478

The two nodes we decided to use for this data have different signs for the variables age, credit card average, certificate of deposit account, owner of a universal bank credit card, securities account and years of experience income. The Tab.3.4 shows the weights for each of those variables for H11 and H12. The other variables are neglected as they have the same sign and do not contribute to the separation of H11 and H12. The node H11 is high for older people with less experience and low income. They use extra services of the bank as a certificate of deposit as well as a securities account and a universal bank credit card, although there average credit card spending is low. This group seems to be risk averse and tends to save money.
On the opposite, H12 is high for a young person with many years of experience and a high income. Moreover, this person does not have securities or a certificate of deposit account, but a high average credit card spending although no universal bank credit. This group is less risk averse as they do not invest in relative secure investments and have a high credit card spending.

Table 3.4: Estimated weights for variables and neurons

From	To	Weight
Age	H11	1.284797485
Age	H12	-0.53730081
CCAvg	H11	-1.598074976
CCAvg	H12	0.0599556001
CD_Account0	H11	1.0999273257
CD_Account0	H12	-0.511651061
CreditCard0	H11	-0.331022723
CreditCard0	H12	0.1743873199
Experience	H11	-0.946177179
Experience	H12	0.4166464132
Income	H11	-3.621653409
Income	H12	2.0138793882
Securities_Account0	H11	-0.298914945
Securities_Account0	H12	0.2185337644
H11	Personal_Loan1	-11.85376809
H12	Personal_Loan1	9.8259307382

The box plots in Fig.3.9 show that each of the two nodes represents one level of the response variable. The node H11 describes non-loan taker, as the median is high for this response group. The other node H12 has a high median for loan taker, thus the description of this node illustrates loan taker.

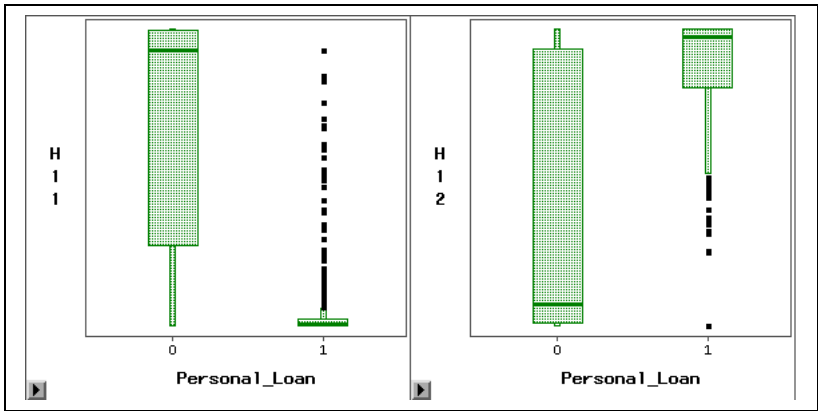


Figure 3.9: Box plot of personal loan and neurons H11, H12

From the neural network analysis we conclude that young individuals with many years of experience and high income, no securities or certificate deposit account, a high credit card spending but no universal bank credit card are likely to be a loan taker. The model has a low error rate of about 2%; hence, the predictions made about the target variable by this model are reliable.

3.4. Cluster Analysis

This part applies a cluster analysis to gain further information about the associations in the data. First, we have to decide which cluster method to use and especially choose a reasonable number of clusters. Second, we have a close look at the clusters and interpret them in terms of loan taker.
After trying different numbers of cluster analysis for the k-means cluster analysis and the Kohonen cluster analysis, we decided to use a 4-means cluster analysis. This evolved to be the one, which distinguishes the data best and is reasonable in terms of interpretation. The distance plot in Fig.3.10 shows that cluster 4 and 2 lie the furthest away, cluster 1 and 3 lie closer together. However, it is better not to merge 1 and 3 into one cluster. The reason is that the distinction between loan and non-loan taker is greater if we use four clusters. The most important variables presented in Tab.3.5 are (in descending order): online baking with a value of 1, family, personal loan, CD account, universal bank credit card, years of experience, age and securities account.

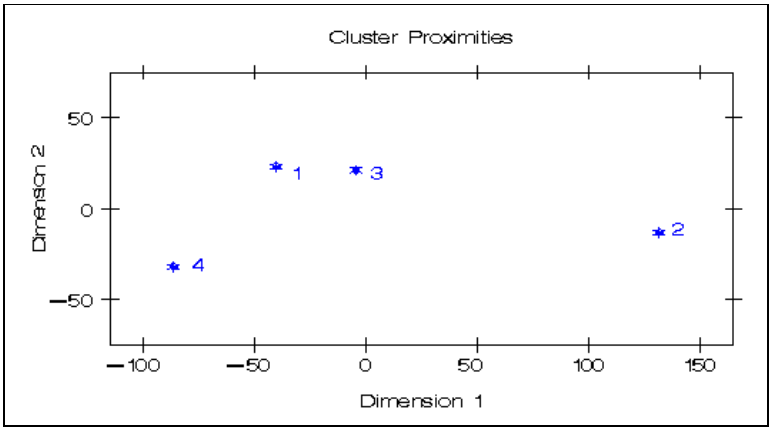


Figure 3.10: Distance plot

Table 3.5: Important variables for 4-means cluster

Name	Importance
AGE	0.4249818228
EXPERIENCE	0.4287737511
INCOME	0
ZIP_CODE	0
FAMILY	0.7539488124
CCAVG	0
EDUCATION	0
MORTGAGE	0
PERSONAL_LOAN	0.7242508621
SECURITIES_ACCOUNT	0.2500777621
CD_ACCOUNT	0.7086547555
ONLINE	1
CREDITCARD	0.600573304

The pie graph in Fig.3.11 shows that cluster 2 and 4 contain high proportions of loan takers. Cluster 2 contains 61 of the in total 196 loan-takers in the data; in the cluster 2 those 61 individuals account for 52.13% of the individuals. Cluster 4 has 132 loan-takers of the 196 in the data, which accounts for 46.8% of the individuals in cluster 4. The other two clusters 1 and 3 almost only contain non-loan taker. As our purpose is to find out, which characteristics loan-takers have, we focus on the attributes of cluster 2 and 4. According to the distance plot above, those two clusters lay far apart from each other, thus have different features and can be distinguished well from each other. It is likely that loan-takers belong to two separate groups with diverse characters. To find out more about this is an interesting task regarding our research aim.

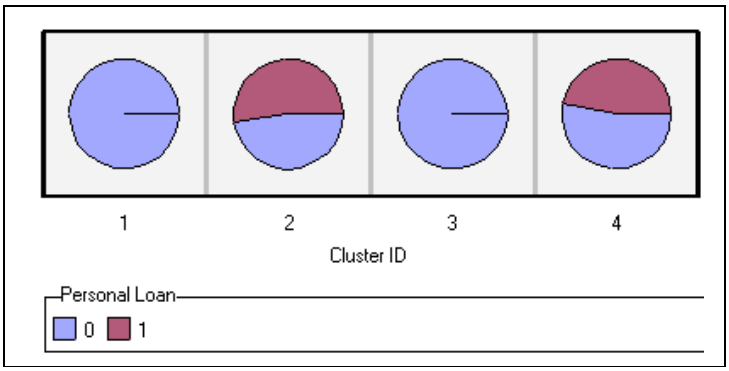


Figure 3.11: Pie graph of personal loan

The pie graph in Fig.3.12 shows the distinguishing variable online banking for the clusters. Cluster 2 mostly consists of people who use online banking; while cluster 4 is characterised by people who do not use online banking.

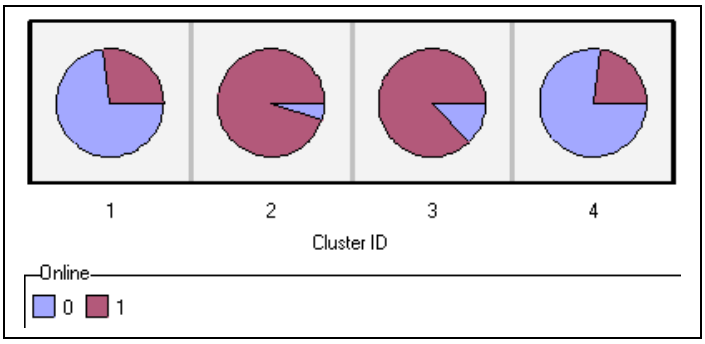


Figure 3.12: Pie graph of online banking

Regarding the variable family (Fig.3.13), cluster 2 does have equal proportions for each family size. Hence, the variable family size does not clearly distinguish loan-taker in cluster 2. In cluster 4, the trend is more obvious, as this cluster mostly contains people with a single household.

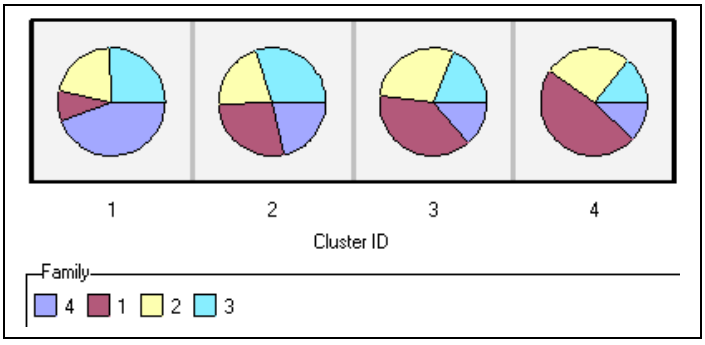


Figure 3.13: Pie graph of family

The variable CD account presented in Fig.3.14 distinguishes well between cluster 2 and 4. In cluster 2, each individual has a CD account. Cluster 4 contains solely people who do not have a CD account.

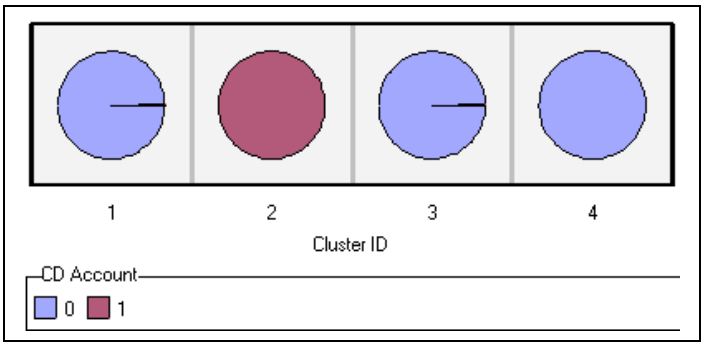


Figure 3.14: Pie graph of CD account

The variable credit card (Fig.3.15) distinguishes cluster 2 well from the other ones. In general, most of people in cluster 2 hold a credit card from the universal bank. For cluster 4, we see that most of the individuals do not hold a credit card. However, this is similar to cluster 1 and 3.

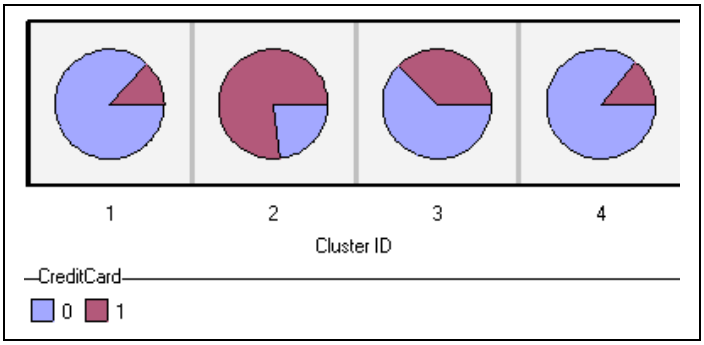


Figure 3.15: Pie graph of credit card

Concerning the variable securities account in Fig.3.16, cluster 2 has equal proportions of both groups while in cluster 4 most clients do not have a securities account.

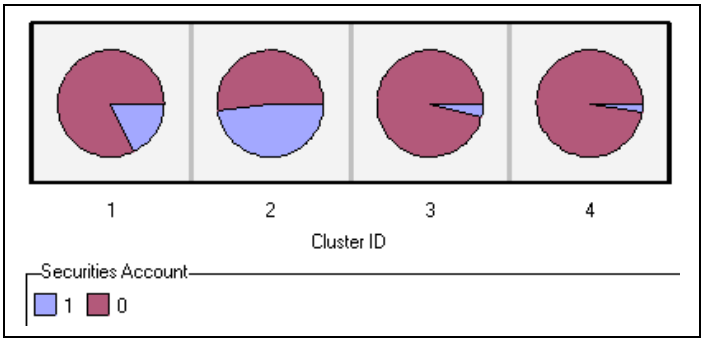


Figure 3.16: Pie graph of securities account

Regarding the education level shown in Fig.3.17, cluster two shows same proportion for each category, thus this variable is not helpful to distinguish cluster 2. However, cluster 4 does contain mostly education level 3, which indicates advanced or professional level.

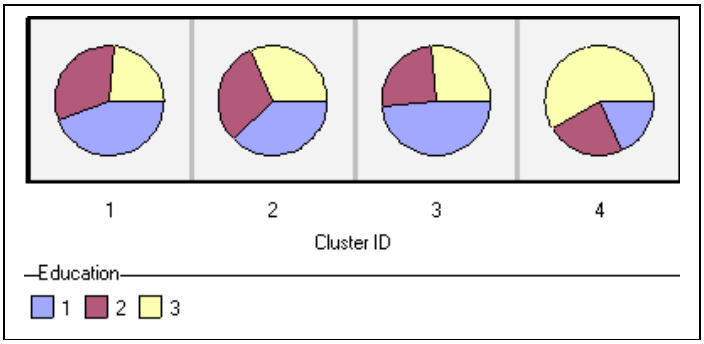


Figure 3.17: Pie graph of education

Overall, we decide that cluster 4 gives an informative description of potential loan taker. Those people do not use online banking, mostly have a single household, do not have a CD or securities account, do not have a credit card of the universal bank and have a high level of education. In other words, well-educated single people not interested in extra services of the bank.
Cluster 2 also describes potential loan taker. However, it gives less information about there character. Those people are online bank user, have a CD account and a universal bank credit card. In other words, people who use extra services of the bank. Unfortunately, this cluster does not give any specific demographic description.

3.5. Principal Component Analysis

Using the principal component analysis, we try to reduce the dimension of the explanatory variables and build some PCs, which sum up the information of the data variables. First, SAS produces a default number of PCs of which we select a few using a cut-off rule. Second, we try to interpret the components regarding our aim to characterise individuals who tend to take a loan.
SAS creates 16 principal components as the default setting where eleven of them have an eigenvalue of greater than one, which is the cut-off point when selecting a number of PC to use (Fig.3.18). The proportion of the total variation explained by each PC is small, beginning at 15% for PC1 going down to 5% for PC11 and adding up to a cumulative proportion of 87% (Tab.3.6). Overall, each principal component explains only a low proportion of variation in the data. To explain an appropriate amount of variability we need 11 principal components. Considering that the data itself only contains 13 variables, the method seems not to give a useful reduction of dimension. However, we will try to find out if there are some useful results in the output.

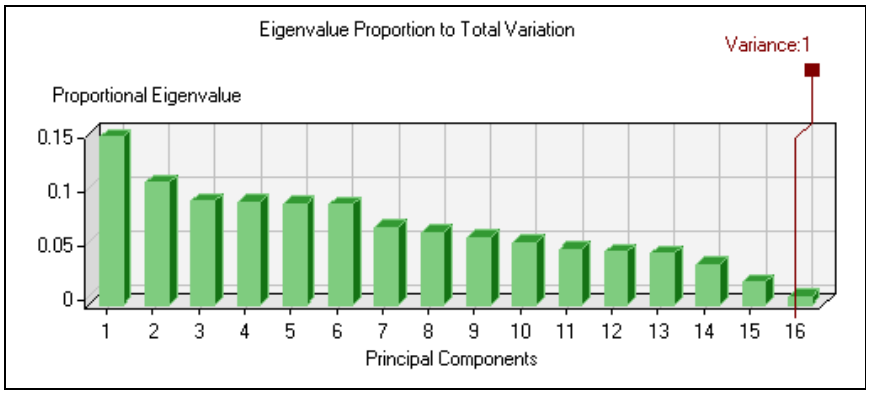


Figure 3.18: Eigenvalue proportion

Table 3.6: Eigenvalues

NAME	EigenValue	ProportionalEigenvalue	CumProportionalEigenvalue
PRIN1	3.4099316568	0.1482578981	0.1482578981
PRIN2	2.4398100464	0.1060786977	0.2543365958
PRIN3	2.0425694828	0.0888073688	0.3431439646
PRIN4	2.0228905683	0.0879517638	0.4310957284
PRIN5	1.9749092129	0.085865618	0.5169613464
PRIN6	1.9555717168	0.0850248573	0.6019862037
PRIN7	1.4779673164	0.0642594485	0.6662456522
PRIN8	1.3719222486	0.0596487934	0.7258944456
PRIN9	1.2671715423	0.0550944149	0.7809888605
PRIN10	1.1462988717	0.0498390814	0.8308279419
PRIN11	1.0013592434	0.0435373584	0.8743653003

Using the important coefficients estimates of the variables for each PC, we can interpret them. In Tab.3.7, the coefficients for the first five principal components are shown. The first PC indicates that a loan taker has a CD account, a universal bank credit card, a high income, a mortgage and a high average credit card use. The second PC describes a loan taker as somebody who does not use extra bank offers as securities of CD account, online banking and universal bank credit card. Moreover, this person has a high income, is young, has only a few years of experience, a mortgage and a high average credit card usage. The third PC is already explained by the first PC and it does not contribute any further information. As we go down in the order of components, we conclude that all of them are already explained by one of the first two. Furthermore, there explanation of variability proportion descends, thus it is not worth to interpret them.

Table 3.7: Principal component coefficient estimates

NAME	_LABEL_	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5
Family_1_	1	-0.015277167	0.0332404006	0.040148536	0.0285910867	-0.01550048
Family_2_	2	0.0153131686	0.0796229538	0.1043029192	-0.070658195	0.0272299881
Family_3_	3	0.0200452887	-0.0463464	-0.044694539	0.1032446926	0.0206414411
Family_4_	4	-0.018139902	-0.073145242	-0.107181425	-0.054733941	-0.030611586
Education_1_	1	-0.004105549	0.071005289	0.1332244768	-0.189935493	-0.071389147
Education_2_	2	-0.000108296	-0.066296905	-0.133123107	0.0712042786	-0.025453185
Education_3_	3	0.0045259593	-0.011452766	-0.012928068	0.134675866	0.1018040083
Personal_Loan_1_	0	-0.38640956	-0.310823005	0.0274603996	-0.135281722	-0.031553759
Personal_Loan_2_	1	0.3864095599	0.3108230051	-0.0274604	0.1352817217	0.0315537593
Securities_Account_1_	0	-0.204869055	0.285928243	0.3963560477	0.095496208	-0.279963804
Securities_Account_2_	1	0.2048690548	-0.285928243	-0.396356048	-0.095496208	0.2799638042
CD_Account_1_	0	-0.428252912	0.2274757355	-0.00678212	0.0222583006	-0.016349651
CD_Account_2_	1	0.4282529119	-0.227475736	0.0067821195	-0.022258301	0.0163496513
Online_1_	0	-0.116237762	0.2037843785	0.0932050853	-0.213153158	0.5930271257
Online_2_	1	0.1162377615	-0.203784379	-0.093205085	0.2131531576	-0.593027126
CreditCard_1_	0	-0.164963118	0.2759251544	-0.51226383	0.2034676197	-0.027574663
CreditCard_2_	1	0.1649631178	-0.275925154	0.5122638297	-0.20346762	0.0275746634
Age		-0.015191441	-0.113247855	0.18899542	0.5894180688	0.2398578654
Experience		-0.013623287	-0.11076212	0.1927843735	0.5846230549	0.2371731524
Income		0.300581279	0.3677559535	0.0672535995	-0.018505622	-0.041607271
ZIP_Code		0.008919303	-0.018128548	-0.016269196	-0.031747761	-0.034283945
CCAvg		0.2538127439	0.321427162	0.0470342244	-0.034685107	-0.02595682
Mortgage		0.112611671	0.1263300034	0.0223971163	-0.001213184	-0.00389237

In conclusion, the dimension reduction is not successful for this data. The method creates 11 PCs with an eigenvalue bigger than one, which altogether explains 87% of the variability. Considering that the data set itself has 13 explanatory variables, this is not a good result. However, interpreting the first two principal components show that both of them could be useful for deciding which characteristics loan-taker could have. PC1 describes individuals with a high income who also use different credit facilities of the bank as credit card and mortgage but also have stored money on a CD account. PC2 consists of young people with only a few years of experience but a high income. They use additional bank facilities as CD account and online banking, also different credit services as mortgage and credit card. The remaining nine PC can be discarded from the interpretation, as they do not inhabit further information.

4. Conclusion

The aim of the universal bank is to convert there liability customers into loan customers. They want to set up a new marketing campaign; hence, they need information about the connection between the variables given in the data and an enhancement of loan customers. The data is based on a campaign with the same purpose ran last year. Five data mining techniques are used in this study to determine a link between the 12 variables given in the data and the response variable personal loan. Those techniques are explanatory data analysis, classification trees, neural networks, cluster analysis and principal components analysis. Summing up the conclusions from the different data mining techniques, shows that three different groups of possible loan taker can be classified.

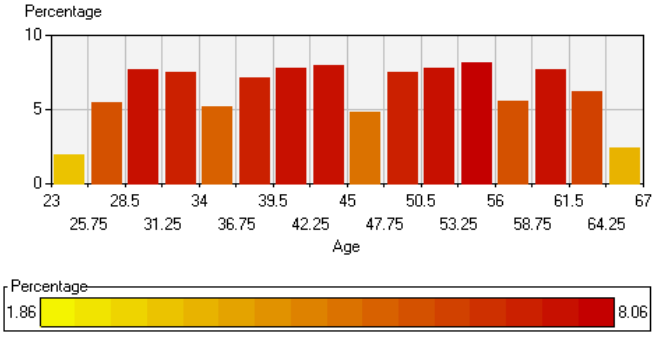
1) From the EDA, entropy tree, neural network and principal components analysis we follow that young people with a high income over $113500, many years of experience and low education level are likely to take a loan. This group might be described as “a new generation self-made man” as they do not have a high education level and still have a high income at a young age. A further characteristic of this group is a high average credit card spending per month of over $2800. Less important is that they probably have a CD or securities account and a family with several members. However, those two facts are not clear.

2) From the EDA, entropy tree, cluster analysis and PC analysis we find a second group of possible loan taker. This group consists of individuals with a high education people and with a high income of over $94500. They like to use extra services as online banking; thus, like to use new medium as internet. Moreover, these individuals use different credit possibilities of the bank, as they have a mortgage and exhibit a universal bank credit card. Some have a high average credit card spending over $2800. In addition, the use of investment possibilities as CD or securities account characterises this group. Overall, this is the group of “open minded people” due to there high education and openness to different facilities of the bank.

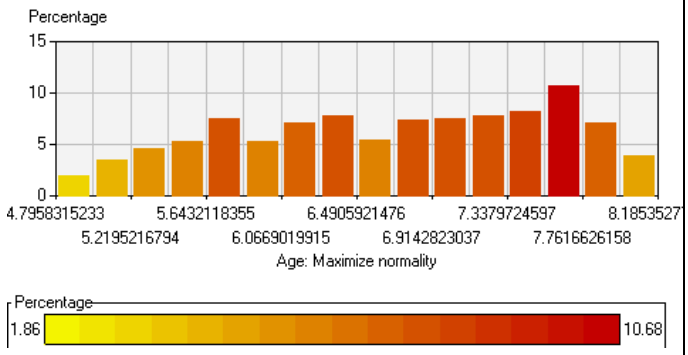
3) The cluster analysis reveals a different kind of group, which is composed of well-educated single people who are not interested in additional bank services as CD or securities account, online banking or credit card. This group can be named as “conservative” as the individuals are not open to new facilities that the bank has to offer.

Appendix A : Histograms for interval and transformed interval variables

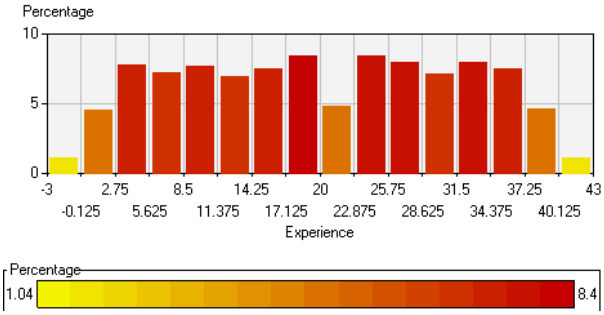
Age	
Skewness	-0.029
Kurtosis	-1.15
Mean	45.34
Std Dev	11.46



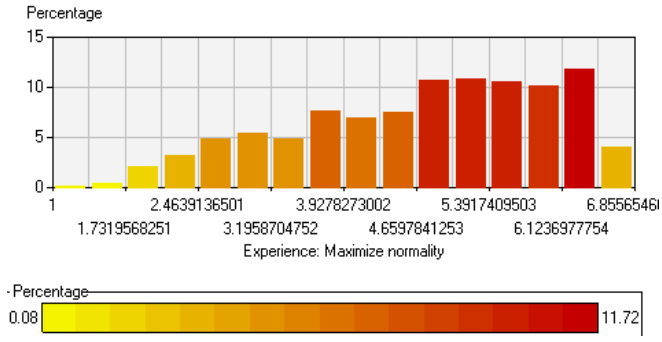
Sqrt (Age)	
Skewness	-0.2
Kurtosis	-1.09



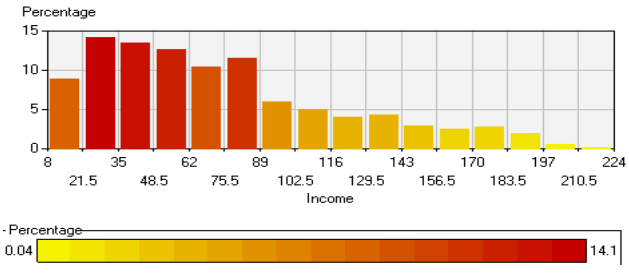
Experience	
Skewness	-0.03
Kurtosis	-1.12
Mean	20.1
Std Dev	11.47



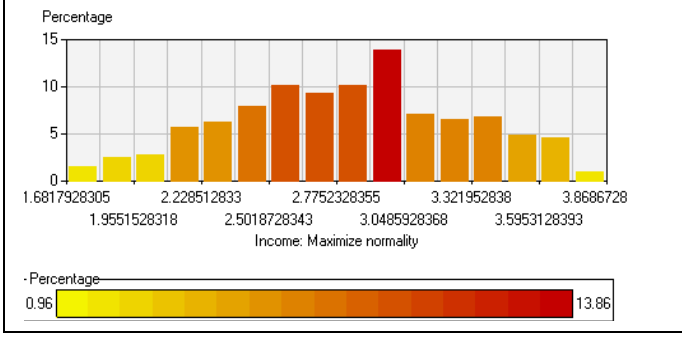
Sqrt (Experience + 4)	
Skewness	-0.44
Kurtosis	-0.76



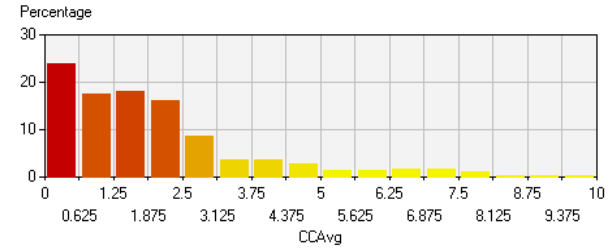
Income	
Skewness	0.84
Kurtosis	-0.04
Mean	73.77
Std Dev	46.03



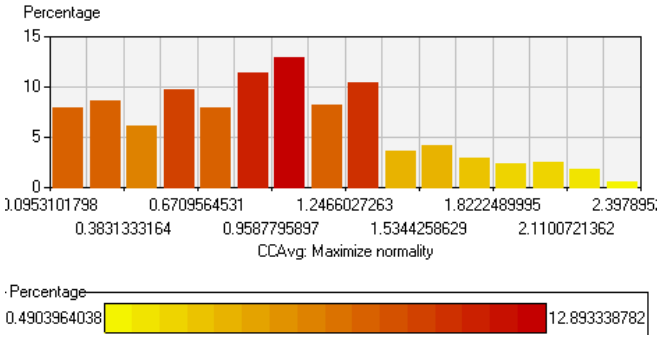
(Income)** 0.25	
Skewness	-0.08
Kurtosis	-0.62



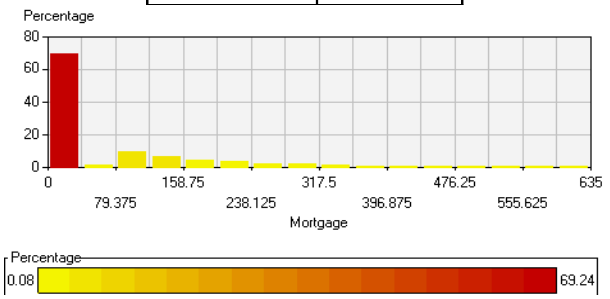
CC Avg	
Skewness	1.6
Kurtosis	2.65
Mean	1.94
Std Dev	1.75



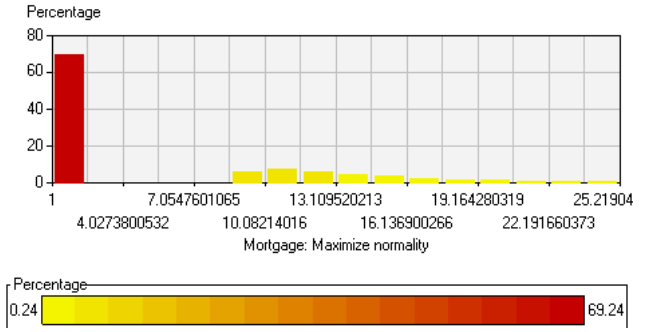
Log (CC AVG + 1)	
Skewness	0.36
Kurtosis	-0.44



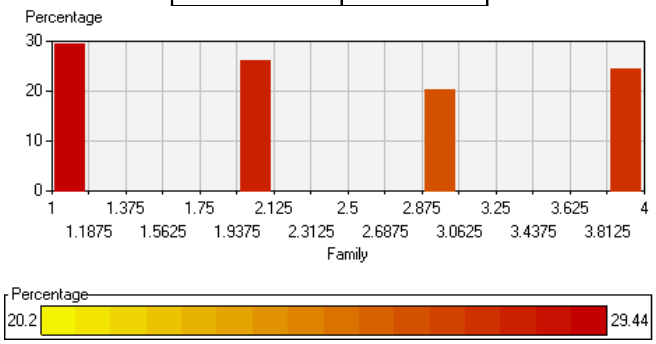
Mortgage	
Skewness	2.1
Kurtosis	4.76
Mean	56.5
Std Dev	101.7



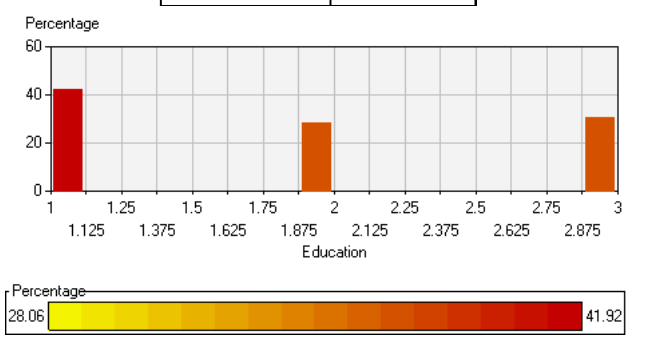
Sqrt (MORTGAGE + 1)	
Skewness	1.2
Kurtosis	0.03



Family	
Skewness	0.16
Kurtosis	-1.04
Mean	2.4
Std Dev	1.15



Education	
Skewness	0.23
Kurtosis	-1.55
Mean	1.88
Std Dev	0.84



5. References

Dr. Jordan, Claire: Practical Data Mining Study Guide. 2007