

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ



ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

School of Economic Sciences

Master of Science in Economics

Statistical Methods Applied to Credit Scoring, Using German Credit Data

Maria Griva

Dissertation submitted to fulfill the necessary conditions for
obtaining the master's degree

Athens
(November 2018)



We approve Griva Maria's dissertation.

Sign

Name of Supervisor Professor:
Ekaterini Kyriazidou
Athens University of Economics and Business

Sign

Name of Examiner Professor:
Sofia Dimeli
Athens University of Economics and Business

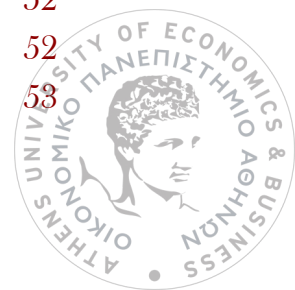
Sign

Name of Examiner Professor:
Theodoros Palivos
Athens University of Economics and Business



Contents

1.Introduction.....	4
2.Credit scoring.....	6
2.1 The Meaning of Data Basis.....	7
2.2 History of Credit Scoring	7
2.3 Benefits of Credit Scoring	9
2.4 Limitations of Credit Scoring.....	11
2.5 Literature Review.....	12
3.Explanation of Dataset and Regression Diagnostics.....	16
3.1 Explanation of Dataset	16
3.2 Regression Diagnostics.....	16
3.3 Testing Heteroscedasticity using Statistical test.....	21
4.Subset Selection Methods.....	23
4.1 Variable Selection	23
4.2 Forward Stepwise Selection	24
4.2.1 Forward Stepwise Selection Approach.....	25
4.3 Backward Stepwise Selection.....	26
4.3.1 Backward Stepwise Approach.....	27
4.4 Cross Validation and Best Subset Selection Method	28
4.4.1 Best Subset Selection Approach.....	28
4.5 Ridge Regression.....	29
4.5.1 Ridge Regression Approach.....	30
4.6 Lasso Method	31
4.6.1 Lasso Approach.....	32
4.7 Principal Component Regression.....	33
4.7.1 Principal Component Approach	33
4.8 Partial Least Squares	35
4.8.1 Partial Least Squares Approach.....	36
4.9 Conclusion.....	37
5.Discriminative Model	39
5.1 The logistic Regression (Linear Classifier).....	39
5.2 Why Logistic Regression.....	40
5.3 How the Logistic Regression works with our Dataset	40
6.Generative Model.....	45
6.1 Linear Discriminant Analysis.....	45
6.2 How the Linear Discriminant Analysis works with our Dataset.....	47
7.Quadratic Discriminant Analysis.....	50
7.1 How Quadratic Discriminant Analysis works with our Dataset.....	50
8. Classification Methods.....	52
8.1 K-Nearest Neighbors.....	52
8.1.1 How the K-NN works with our Dataset	53



8.2 Tree-Based Methods.....	55
8.2.1 Decision Trees.....	55
8.2.2 Classification trees.....	57
8.2.3 How the Decision trees work with our Dataset.....	58
8.3 Random Forest.....	60
8.3.1 How Random Forest works with our Dataset.....	61
9. Comparison of Credit Scoring Methods.....	65
10. Conclusion.....	66



1.Introduction

Credit Risk is defined as the risk of loss of principal or loss of a financial return resulting from a borrower's failure to repay a loan or to meet an agreed obligation. Credit risk arises and can be expected when a borrower does not meet their obligation in relation to future cash flows. Therefore, there is uncertainty over the borrower's financial performance in the future. As a result, in recent years, financiers seek tools and means to enable them to calculate the borrower's credit worthiness. A suitable credit limit can then be defined, risk-based pricing can be set, and subsequently adequate loan loss provisions can be kept safeguarding against the possible losses, in case the customers' obligations are not met.

Prediction of loan default has an obvious practical utility. The identification of default risk appears to be of paramount interest to banks. A lending major of a bank must evaluate tens or even hundreds of thousands of loan applications each year. These obviously cannot all be subjected to the scrutiny of a loan committee in the way that, say, a real estate loan might. Thus, statistical methods and automated procedures are essential. Banks typically use "credit scoring models". In principle, the credit score could incorporate any amount of relevant business information. In practice, credit scoring for loan applications appears to be focused narrowly on default risk. Basically, through credit scoring, lenders use scores to determine who qualifies for a loan, at what interest rate, and what credit limits. Lenders also use credit scores to determine which customers are likely to bring in the most revenue. The use of credit or identity scoring prior to authorizing access or granting credit is an implementation of a trusted system. Particularly, classification methods will provide results used in prediction or estimation.

The approaches for predicting qualitative responses is a process that is known as classification. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class. On the other hand, often the methods used for classification first predict the probability of each of the categories of a qualitative variable, as the basis for making the classification. In this sense they also behave like regression methods.



There are not only Credit Scoring Methods, but also Judgmental. Judgmental forecasting methods incorporate intuitive judgement, opinions and subjective probability estimates. Judgmental forecasting is used in cases where there is lack of historical data or during completely new and unique market conditions. Generally speaking, it seems that the only organizations which do not use credit scoring approaches are the smaller and/or more personal companies, and those concerned with corporate finance, where statistical methods have been slower to be adopted. However, although the financial community may have confidence in objective statistical credit scoring methods, there seems still to be some suspicion of them in the customer base. This stems in part from anxiety about the impersonal nature of the process and in part from concerns over the accuracy of the data relating to the individual applicant. (Hand & Henley, 1997), (GIETZEN, 2017)

However, we do not discuss about judgmental methods, but exclusively for credit scoring methods.

A German credit dataset of a bank with 20 independent variables is used in this thesis. Creditability is the dependent variable ($Y \rightarrow Cdblt$) and the dataset contains 1000 loan applicants. Different statistical-classification methods are performed and predict whether a loan applicant is creditworthy or not. In other words, “good” or bad”.

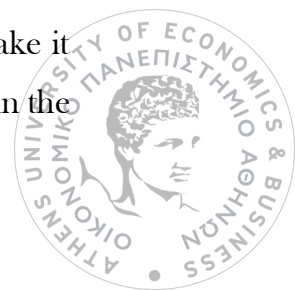
Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, K-Nearest Neighbors, Tree Based Methods: Classification trees (Decision Trees) and Random Forest are some statistical methods which are examined in this thesis.



2.Credit Scoring

Credit scoring is a statistical method used to predict the probability that a loan applicant or existing borrower will default or become delinquent. In other words, credit scoring is a method of evaluating the credit risk of loan applications. The method, introduced in the 1950s, is now widely used for consumer lending, especially credit cards, and is becoming more commonly used in mortgage lending. Credit scoring is already allowing large banks to expand into small-business lending, a market in which they have tended to be less active. Scoring is also an important step in making the securitization of small-business loans more feasible. The likely result would be increased availability of funding to small businesses, and at better terms, to the extent that securitization allows better diversification of risk.

Using historical data and statistical techniques, credit scoring tries to isolate the effects of various applicant characteristics on delinquencies and defaults. The method produces a “score” that a bank can use to rank its loan applicants or borrowers in terms of risk. To build a scoring model, developers analyze historical data on the performance of previously made loans to determine which borrower characteristics are useful in predicting whether the loan performed well. A well-designed model should give a higher percentage of high scores to borrowers whose loans will perform well and a higher percentage of low scores to borrowers whose loans won’t perform well. But no model is perfect, and some bad accounts will receive higher scores than some good accounts. Information on borrowers is obtained from their loan applications and from credit bureaus. Data such as the applicant’s monthly income, outstanding debt, financial assets, how long the applicant has been in the same job, whether the applicant has defaulted or was ever delinquent on a previous loan, whether the applicant owns or rents a home, and the type of bank account the applicant has are all potential factors related to loan performance. Regression analysis relates loan performance to these variables which are used to pick out which combination of factors best predicts delinquency or default, and how much weight should be given to each of the factors. Given the correlations between the factors, it is quite possible some of the factors the model developer begins with won’t make it into the final model, since they have little value added given the other variables in the



model. Even a good scoring system won't predict with certainty any individual loan's performance, but it should give a fairly accurate prediction of the likelihood that a loan applicant with certain characteristics will default.

To build a good scoring model, developers need historical data, which reflect loan performance in periods of both good and bad economic conditions. (Thomas, A survey of credit and behavioural scoring: forecasting financial, 2000) (Fernandes., 2016), (A. Abdou & Pointon, 2011)

2.1 The Meaning of Data Basis

Scoring lenders with small portfolios may never be able to use scoring. The data base must be computerized, and it ideally would include both approved and rejected applicants, although most lenders will have kept records only on approved applicants. The data base, as we mentioned above, should include a full range of characteristics of the borrower, the lender, and the loan, as well as data on the timing and length of each spell of arrears in each loan. These characteristics are all simple and inexpensive to collect, and most microfinance lenders already collect them when the loan officer visits a potential borrower. They could rate potential borrowers as very below average, below average, average, above average, or very above average on such qualities as reputation in the community, entrepreneurship, experience with debt, and informal support networks. The rigorous analysis of a data base of past loans may have vast power to improve management decisions. (Schreiner, 2000)

2.2 History of Credit Scoring

The history of credit scoring dates from the 1950s. Credit scoring is a way classifying borrowers into two groups- those who will default and those who will not- using the characteristics of the borrower and the loan. The first approach to solving this problem of identifying the groups in a population was introduced in statistics by Fisher. He sought to differentiate between two varieties of iris using measurements of the physical size of the plants and to differentiate the origins of skulls using their physical measurements. David Durand in 1941 was the first to recognize that one

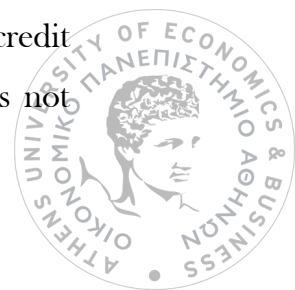


could use the same techniques to discriminate between good and bad loans. His was a research project for the U.S National Bureau of Economics Research and was not used for any predictive purpose.

During the 1930s some of the mail order companies had introduced numerical scoring systems to try and overcome the inconsistencies in credit decisions across credit analysis. With the start of World War II, all the finance houses and mail order firms began to experience difficulties with their credit management. The credit analysts were being drafted into military service, and there was a severe shortage of people who had this expertise. So, the firms got the analysts to write down the rules of the thumb they used to decide to whom to give loan. Some of these were the numerical scoring system already introduced, others were essentially setting of conditions that needed to be satisfied. These rules were then used by nonexperts to help make credit decisions- the first examples of experts' systems.

It did not take long after the war ended for some people to connect the automation of credit decisions and the classification techniques being developed in statistics and to see the benefit of statistically derived models in lending decisions. The first consultancy was formed in San Francisco by Bill Fair and Earl Isaac in the early 1950s, and their clients at the time were mainly finance houses, retailers, and mail order firms.

The arrival of credit cards in the late 1960s made the banks and credit card issuers realize the usefulness of credit scoring. The number of people applying for credit cards each day made it impossible both in economic and manpower terms to anything but automate the lending decision. The growth in computing power made it possible to undertake this automation. When these organizations used credit scoring, they found that it also was a much better predictor than any judgmental scheme, and default rates would drop by 50% or more. The only opposition came from those like Capon, who argued that “the brute force empiricism of credit scoring offends against the traditions of our society”. He felt that there should be more dependence on credit history and that it should be possible to explain why certain characteristics are needed in a scoring system and others are not. The event that ensured the complete acceptance of credit scoring was the passing of the Equal Credit Opportunity Acts in the U.S. in 1975 and 1976. These outlawed discriminating in the granting of credit unless the discrimination “was empirically derived and statistically valid”. It is not



often that lawmakers provide long-term employment for anyone but lawyers, but this ensured that credit scoring analysis was to be a growth profession for the next 30 years this is still the case, and this growth has spread from the U.S. across the world. This is partly because consumer credit has grown very quickly in countries with large populations, like China and India, and because of new types of consumer credit such as peer-to- peer lending.

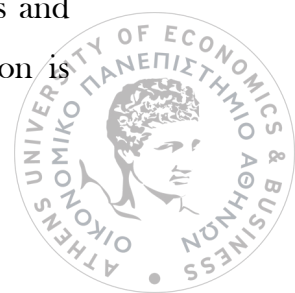
In the 1980s, the success of credit scoring in credit cards meant that banks started using scoring for their other products, such as mortgages and personal loans. Also, in the 1990s the growth in direct marketing has led to the use of scorecards to improve the response rates to advertising campaigns. In fact, this was one of the earliest uses in the 1950s when Sears used scoring to decide to whom to send is catalogues. Advances in computing allowed other techniques such as neural networks, support vector machines and random in the credit scoring context.

Lenders' objectives have changed from trying to minimize the chance a customer will default on one product to looking at how the lender can maximize the profit from that customer. Moreover, the original idea of estimating the risk of defaulting has been augmented by scorecards which estimate response, usage, retention, churn or early repayment, debt management and fraud scoring.

However, the greatest impact on credit scoring since 2000 is the advent of the Basel Accords. Since a credit score can be transformed into a probability of the borrower defaulting, credit scoring has become the mainstay of the models financial institutions have developed to meet these banking regulations. The financial institutions have developed to meet these banking regulations. The Basel Accords, Basel I (1988), Basel II (2005), and Basel III (2010), determine how much capital banks must set aside to meet the credit risk of their borrowers defaulting. (Greenspan, 2017), (Thomas, A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers, 2000)

2.3 Benefits of Credit Scoring

Credit reporting and credit scores can fuel economic growth, increase consumer access to essential resources and enable more efficient allocation of risk, costs and financial reserves. The reason for this is simple: where access to information is



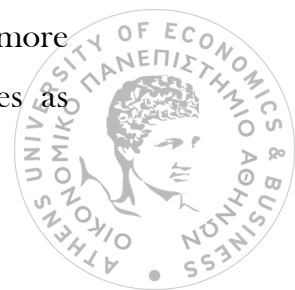
asymmetrical or unavailable, access to lending and credit resources becomes more difficult, expensive and inefficient. In contrast, the availability of credit information and the free flow of objective data are the cornerstones of a modern, successful market economy. Credit reporting and scoring enable consumers and private companies to freely transact with each other because the more objective information the business has, the more accurately it can meet consumer needs and preferences.

Credit histories are available for more than 200 million consumers, helping them achieve their financial and personal goals. Businesses use automated quantification of consumer credit histories-in the form of credit scores-to make more efficient, objective decisions about whether to extend credit and on what terms for such services as credit cards, consumer loans, mortgages and even insurance policies.

A credit score is the result of advanced analytical models that take a ‘snapshot’ of the consumer’s credit report and translate it into a three-digit number representing the amount of risk a consumer brings to a transaction, such as financial, insurance or even employment. Because of credit scoring, lenders can make faster, more objective decisions. Lenders retain complete control over their lending decisions and set their own score levels.

A wide range of industries take advantage of credit scores to improve fairness, effectiveness and efficiency. Financial companies use credit scores to predict the risk of delinquencies and losses, which enables them to better allocate costs. Insurance companies use specialized credit scores to make fairer underwriting decisions. Credit scores even provide benefits at the macroeconomic level by helping small enterprises attain the funds they need and by facilitating the securitization and sale of financial products in the secondary markets, substantially increasing the influx of capital into a country.

Credit scoring offers multiple benefits at every level of the economy. It has enabled lenders to extend into historically underserved market segments. In addition, decisions are now faster and more objective with most applicants receiving answers within minutes, rather than days. Finally, by using credit scores to predict risk more effectively, lenders have been able to reduce the cost of such vital services as



mortgages, personal loans and credit cards. Despite this expansion into traditionally underserved markets, moral hazard rates are lower with credit scores because lenders can more proactively monitor risk and maintain it at more appropriate levels.

Credit scoring plays a vital role in economic growth by helping expand access to credit markets, lowering the price of credit and reducing delinquencies and defaults. In the United States, credit scoring helps drive the American economy and makes credit affordable. For consumers, scoring is the key to homeownership and consumer credit. It increases competition among lenders, which drives down prices.

Decisions can be made faster and cheaper and more consumers can be approved. It helps spread risk more so vital resources, such as insurance and mortgages, are priced more fairly. For businesses, especially small and medium-sized enterprises, credit scoring increases access to financial resources, reduces costs and helps manage risk. For the national economy, credit scoring helps smooth consumption during cyclical periods of unemployment and reduces the swings of the business cycle. By enabling loans and credit products to be bundled according to risk and sold as securitized derivatives, credit scoring connects consumers to secondary capital markets and increases the amount of capital that is available to be extended or invested in economic growth. (A. Abdou & Pointon, 2011) (Kern, 2017)

2.4 Limitations of Credit Scoring

Accuracy is a very important consideration in using credit scoring. Even if the lender can lower its costs of evaluating loan applications by using scoring, if the models are not accurate, these cost savings would be eaten away by poorly performing loans. The accuracy of a credit scoring system will depend on the care with which it is developed. The data on which the system is based need to be a rich sample of both well-performing and poorly performing loans. The data should be up to date, and the models should be reestimated frequently to ensure that changes in the relationships between potential factors and loan performance are captured. If the bank using scoring increases its applicant pool by mass marketing, it must ensure that the new pool of applicants behaves similarly to the pool on which the model was built;



otherwise, the model may not accurately predict the behavior of these new applicants. The use of credit scoring itself may change a bank's applicant pool in unpredictable ways, since it changes the cost of lending to certain types of borrowers. Again, this change in applicant pool may hurt the accuracy of a model that was built using information from the past pool of applicants. Account should be taken not only of the characteristics of borrowers who were granted credit but also of those who were denied. Otherwise, a "selection bias" in the loan approval process could lead to bias in the estimated weights in the scoring model. A model's accuracy should be tested. A good model needs to make accurate predictions in good economic times and bad, so the data on which the model is based should cover both expansions and recessions. (A. Abdou & Pointon, 2011), (Yap, Ong, & Husain, Using data mining to improve assessment of credit worthiness via credit scoring models, 2011), (A. Abdou & Pointon, 2011).

2.5 Literature Review

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012) performed a review of methods for variable selection within one of the many modeling approaches for high-throughput data, Partial Least Squares Regression. The aim of their work was to collect and present the methods in such a way that the reader easily can get an understanding of the characteristics of the methods and to get a basis for selecting an appropriate method for own use.

The purpose Chong, I.-G., & Jun, C.-H. (2004) was to explore the nature of the variable importance in the projection method and to compare with other methods through computer simulation experiments. They designed 108 experiments where observations are generated from true models considering four factors—the proportion of the number of relevant predictors, the magnitude of correlations between predictors, the structure of regression coefficients, and the magnitude of signal to noise. Confusion matrix has adopted to evaluate the performance of PLS, the Lasso, and stepwise method. They have also discussed the proper cutoff value of the variable importance in the projection method to increase its performance.



Isabelle Guyon and Andr e Elisseeff (2003) addressed a common methodological flaw in the comparison of variable selection methods. Regarding the selection process they computed cross-validation performance estimates of the different variable subsets. Used with computationally intensive search algorithms, these estimates may overfit and yield biased predictions. Therefore, they cannot be used reliably to compare two selection methods, as is shown by the empirical results of this paper. They claimed that independent test sets should be used for determining the final performance.

Credit scoring has been regarded as a main tool of different companies or banks during the last few decades and has been widely investigated in different areas, such as finance and accounting. Different scoring techniques are being used in areas of classification and prediction, where statistical techniques have conventionally been used. A. Abdou, H., & Pointon, J. (2011) paper aims to research how important credit scoring have been and which are the key determinants in the construction of a scoring model through a widespread review of different statistical techniques and performance evaluation criteria. They have concluded that there is no overall best statistical technique used in building scoring models and the best technique for all circumstances does not yet exist. Also, the applications of the scoring methodologies have been widely extended to include different areas. For example, banks use these tools to predict their clients' behavior.

D. J. Hand and W. E. Henley (1997) mentioned that credit scoring is the term used to describe formal statistical methods used for classifying applicants for the credit into “good” and “bad” risk classes. Such methods have become important with a great increase in the consumption credit.

Hastie, T., Tibshirani, R., & Friedman, J. (May 2001) tried to bring together many of the important new ideas in learning and explain them in a statistical framework. They emphasized the methods and their conceptual underpinnings rather than their theoretical properties.

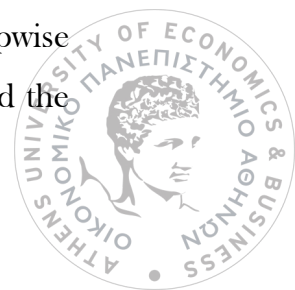


Financial institutions, due to the necessity for controlling and effectively managing credit risk, have improved the techniques designed for this purpose, resulting in the development of various quantitative models. Francisco Louzada, Anderson Ara and Guilherme B. Fernandes (2013) presented a systematic literature review relating theory and application of binary classification techniques for credit scoring financial analysis. The significance of the basic statistical techniques for credit scoring is illustrated from the included results.

Schreiner, M. (2000) claimed that scoring does have a place in microfinance. Although scoring is less powerful in poor countries than in rich countries, and although scoring will not replace the personal knowledge of character of loan officers or of loan groups, scoring can improve estimates of risk. Thus, scoring complements current microfinance technologies. Furthermore, the derivation of the scoring formula reveals how the characteristics of borrowers, loans, and lenders affect risk, and this knowledge is useful whether a lender uses predictions from scoring to inform daily decisions. It is claimed that in the next decade, many of the biggest microfinance lenders will likely make credit-scoring models one of their most important decision tools.

Lyn C. Thomas (2000) has mentioned the need to incorporate economic conditions into the scoring systems and the way the systems could change from estimating the probability of a consumer defaulting to estimating the profit a consumer will bring to the lending organization – two of the major developments being attempted in the area. It points out how successful has been this under-researched area of forecasting financial risk.

Murtaugh, P. A. (2009) evaluated the predictive ability of statistical models obtained by applying seven methods of variable selection to 12 ecological and environmental data sets. Cross validation, involving repeated splits of each data set into training and validation subsets, was used to obtain honest estimates of predictive ability that could be fairly compared among methods. There was surprisingly little difference in predictive ability among five methods based on multiple linear regression. Stepwise methods performed similarly to exhaustive algorithms for subset selection, and the



choice of criterion for comparing models (Akaike's information criterion, Schwarz's Bayesian information criterion or F statistics) had little effect on predictive ability. For most of the data sets, two methods based on regression trees yielded models with substantially lower predictive ability. Murtaugh, P. A. argued that there is no best method of variable selection and that any of the regression approaches discussed in this paper can yield useful predictive models.

Credit scoring model have been developed by banks and researchers to improve the process of assigning credit risk to either a "good risk" group that is likely to repay financial obligation or a "bad risk" group who has high possibility of defaulting on it. Data mining techniques are useful in this process. Using historical data on payments, demographic characteristics etc, credit scoring models can help providing a score for each bank applicant. Bee Wah Yap, Seng Huat Ong, Nor Huselina Mohamed Husain (2011) paper have illustrated how credit scoring models can predict the worthiness of customers. This study has applied the credit scoring techniques using data of payment history of members from a recreational club. The club has been facing a problem of rising number in defaulters in their monthly club subscription payments. The idea is that the management could have a model which can deploy to identify potential defaulters. The classification performance of logistic regression model and decision tree model were compared. The error rates for credit scorecard model, logistic regression and decision tree were 27.9%, 28.8% and 28.1%, respectively. Although no model outperforms the other, scores are relatively much easier to deploy in practical applications.

Statistical learning (James G., Witten D., Hastie T. and Tibshirani R.-2013) refers to a set of tools for modeling and understanding complex datasets. The field encompasses many methods such as the lasso and sparse regression, classification and regression trees, and boosting and support vector machines. With the explosion of "Big Data" problems, statistical learning has become a very hot field in many scientific areas as well as marketing, finance, and other business disciplines.



3.Explanation of Dataset and Regression Diagnostics

3.1 Explanation of Dataset

Our dataset has been imported in R and initially, some numeric and dummy variables has been converted to factor ones, which is more comfortable to see the number of characterized variables (see Appendix).

After converting the variables, using R the number of each category is calculated. We record the proportions of each category as well as the categories (see Appendix).

We perform with the help of R programming the distribution of the continuous variables:

Duration of Credit (DuCrd), Credit Amount (CrdAm) and Age.:

	Min	1st Qu	Median	Mean	3rd Qu	Max
DuCrd	4	12	18	20,9	24	72
CrdAm	250	1366	2320	3271	3972	18424
Age	19	27	33	35,54	42	75

At the above table we can see the minimum, 1st Qu., median, mean, 3rd Qu. and maximum. The Duration of Credit is $4 \leq \text{DuCrd} \leq 72$ per month. The Credit Amount fluctuates between 250 and 18424 DM, while its mean is 3271. The Age of loan applicants is between 19 and 75 as well as the mean age is 33.

The histograms and the boxplots of the continuous variables can be seen in order to have a quick view in Frequency of those (see Appendix).

3.2 Regression Diagnostics

Assume that we are fitting a multiple linear regression on the ger_cre data with $Y = \text{Crdblt}$ and $X_1, X_2, \dots, X_n = \text{AcBa}, \text{DuCrd}, \dots, \text{FrnWrkr}$.

$$\begin{aligned} \text{Crdblt} = & (-5.147\text{e-}02) + (9.879\text{e-}02) * \text{AcBa} - (4.396\text{e-}03) * \text{DuCrd} + (6.566\text{e-}02) * \text{PaStPrCrd} + (4.691\text{e-} \\ & 03) * \text{Purp} - (1.534\text{e-}05) * \text{CrdAm} + (3.424\text{e-}02) * \text{ValSavSto} + (2.482\text{e-}02) * \text{LenCurEmp} - (4.707\text{e-} \\ & 02) * \text{InPerCe} + (4.386\text{e-}02) * \text{SexMarSt} + (5.878\text{e-}02) * \text{Guara} - (2.859\text{e-}03) * \text{DurCurAddr} - (3.250\text{e-} \\ & 02) * \text{MoValAvAs} + (1.067\text{e-}03) * \text{Age} + (3.614\text{e-}02) * \text{ConcurCrd} + (4.988\text{e-}02) * \text{TypAp} - (4.226\text{e-} \\ & 02) * \text{NoCrdBa} + 4.991\text{e-}03 * \text{Occup} - (2.930\text{e-}02) * \text{NoDpnd} + (5.102\text{e-}02) * \text{Tlph} + (1.145\text{e-} \\ & 01) * \text{FrnWrkr} \end{aligned}$$



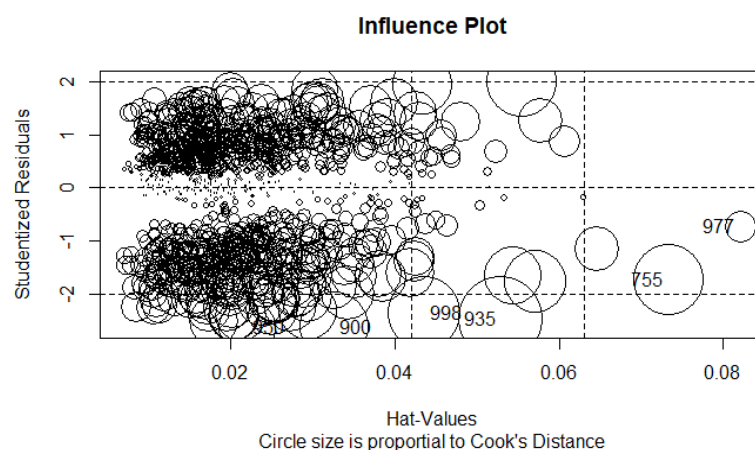
1. leverage plots

Outliers are one of those statistical issues that everyone knows about, but most people aren't sure how to deal with. Most parametric statistics, like means, standard deviations, and correlations, and every statistic based on these, are highly sensitive to outliers. And since the assumptions of common statistical procedures, like linear regression, are also based on these statistics, outliers can really mess up our analysis. Despite all this, as much as you'd like to, it is not acceptable to drop an observation just because it is an outlier. They can be legitimate observations and are sometimes the most interesting ones. It's important to investigate the nature of the outlier before deciding. (<https://www.investopedia.com/terms/a/adjusted-mean.asp>, n.d.)

From the leverage plots (see Appendix), we could basically say that only when $X=Occup$, $X=CrdAm$ and $X=DuCrd$, the outlier could be the 755 observation. Also, when $X=DuCrd$, it could be the 720 observation. We could check if there was a significant change of these slopes when we would omit the observation 755. However, in this case it is not necessary to drop these observations because as we can see they are not outlier for both axis X and Y , but only for axis Y .

Naturally, in diagnostic procedures, several transformations of the ordinary residuals have been suggested to overcome partially some of their shortcomings.

2. Influence Plot



The function in R creates a "bubble" plot of studentized residuals by hat values, with the areas of the circles representing the observations proportional to Cook's distances. Vertical reference lines are drawn at twice and three times the average hat value, horizontal reference lines at -2, 0, and 2 on the studentized-residual scale.



An influence plot shows the outlyingness, leverage, and influence of each case.

The plot shows the residual on the vertical axis, leverage on the horizontal axis, and the point size is the square root of Cook's D statistic, a measure of the influence of the point.

Outliers are cases that do not correspond to the model fitted to the bulk of the data. You can identify outliers as those cases with a large residual (usually greater than approximately ± 2), though not all cases with a large residual are outliers and not all outliers are bad. Some of the most interesting cases may be outliers.

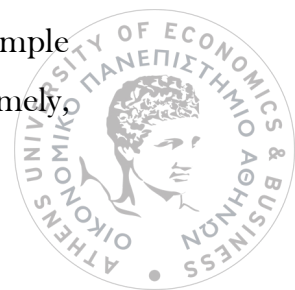
Leverage is the potential for a case to have an influence on the model. You can identify points with high leverage as those furthest to the right. A point with high leverage may not have much influence on the model if it fits the overall model without that case.

Influence combines the leverage and residual of a case to measure how the parameter estimates would change if that case were excluded. Points with a large residual and high leverage have the most influence. They can have an adverse effect on the model if they are changed or excluded, making the model less robust. Sometimes a small group of influential points can have an unduly large impact on the fit of the model.

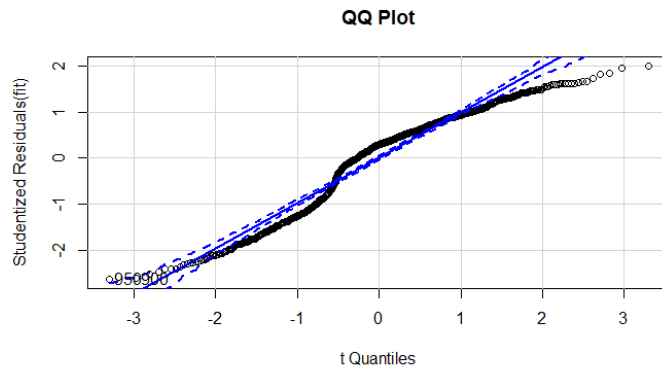
Regarding the above-mentioned Influence Plot, we can say that Standardized Residuals, which are above 2 and below -2, are outliers. Points with Hat-Values above 0.025 and standardized residuals between -2 and 2 are high leverage points. Data points with Hat-Values above 0.025 and standardized residuals above 2 and below -2 are influential points which significantly alter the overall trend. When we observe this plot, we deduce that we do not see outliers because although some Hat-Values are above 0.025 the correspondent standardized residuals are between -2 and 2. The same thing happens when the standardized residuals are above 2 and below -2.

3. Normality of Residuals

In this way, as part of regression diagnostics we can perform the Q-Q Normal Plot. We use a Q-Q plot to check for data Normality. In most cases, we don't want to compare two samples with each other, but compare a sample with a theoretical sample that comes from a certain distribution (for example, the normal distribution). Namely,



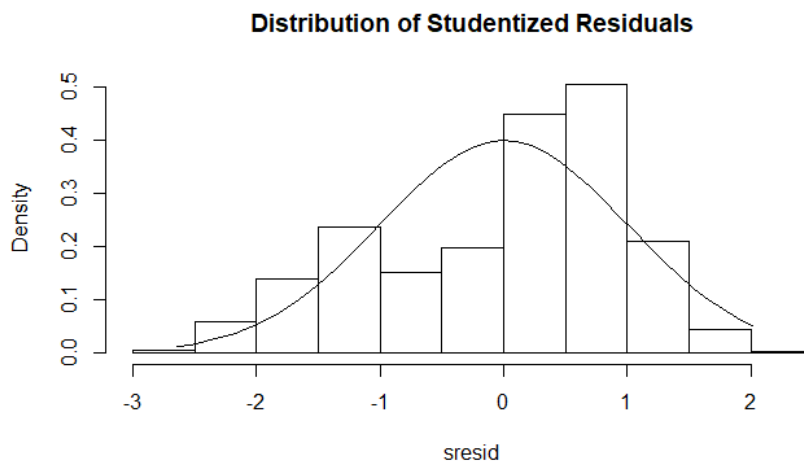
theoretical quantiles are the expected residuals/error which are distributed. To make a Q-Q plot this way, R has the special function. As the name implies, this function plots your sample against a normal distribution. We simply give the sample we want to plot as a first argument and add any graphical parameters you like. R then creates a sample with values coming from the standard normal distribution, or a normal distribution with a mean of zero and a standard deviation of one. With this second sample, R creates the Q-Q plot as explained before.



The closer all points lie to the line, the closer the distribution of your sample comes to the normal distribution.

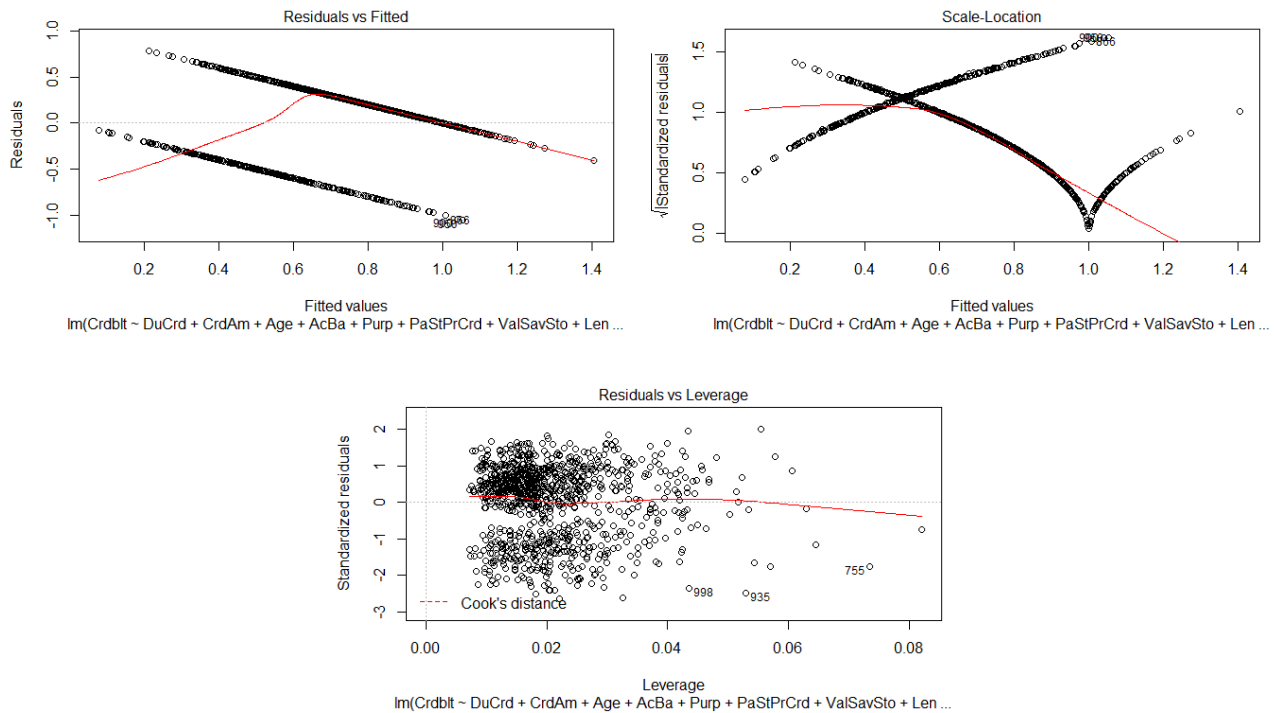
Here, we could say that the distribution is not normal. The points does not approximate the straight line. The points are deviations which are a bit larger right in the edge.

4. Distribution of Standardized Residuals



Here, in this graph it becomes clearer the distribution which is skewed to the left. So, we can say easier that the Standardized Residuals are not normally distributed.

5. Residuals vs Fitted/ Residuals vs Leverage

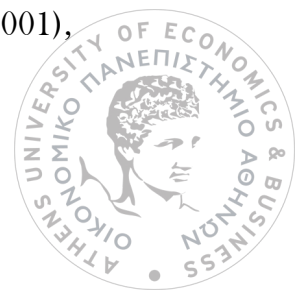


The red line in the first two plots has a strong curvy shape and it does not meet the Linearity Assumption. The red line of the residual vs leverage plot is close to linearity. However, the shape seems to be somehow conic, thus we would say that the model is heteroscedastic. So, the assumptions are not met for fitting a Linear Regression Model.

As part of regression diagnostics is also the evaluation of **Collinearity**. We use the function `vif()` via **R**. We intend to remove these variables which are more than 5. Threshold is from 5 to 10. We take these values via **R**:

DuCrd	CrdAm	Age	AcBa	Purp	PaStPrCrd	ValSavSto	LenCurEmp
1.955228	2.262017	1.308394	1.138452	1.072315	1.379162	1.09926	1.195133
InPerCe	SexMarSt	Guara	DurCurAddr	MoValAvAs	ConcurCrd	TypAp	NoCrdBa
1.310214	1.068769	1.074132	1.18242	1.391757	1.081218	1.313476	1.318022
Occup	NoDpnd	Tlph	FrgrWrkr				
1.33609	1.081393	1.276181	1.086762				

We see that these values **are less than 5**. So, we can safely conclude them. (Ching-Ti Liu, Jacqueline Milton, & Avery McIntosh, 2016), (Quick-R, n.d.), (University of Virginia Library, n.d.), (Hastie, Tibshirani, & Friedman, May, 2001), (Wooldridge, 2006).



3.3 Testing heteroscedasticity using Statistical test

The most accurate way to test for heteroscedasticity is through statistical tests. One of those is Breusch-Pagan test.

It tests whether the variance of the errors from a regression is dependent on the values of the independent variables. In that case, heteroskedasticity is presented. (see Appendix the results of the linear model).

lmtest package and the bptest function are used on our fitted model.

Let's do the Breusch-Pagan test:

Studentized Breusch-Pagan test
data: fit
BP =107.56
df = 20
p-value = 5.468e-14

While it doesn't give us the critical value to compare the test statistic, all you need to look at is the p-value to determine whether or not we should reject the null. If the p-value is less than the level of significance, then we reject the null hypothesis. Since $5.468e-14 < 0.05$, we can reject the null hypothesis. Hence, the model is heteroskedastic.

To get the correct standard errors we can use the `vconHC()` function from the `{sandwich}` package. Trough R, it may be particularly helpful to look just the coefficient matrix from the summary object. Then SEs are generated from the variance-covariance matrix for the coefficients. The variance estimates for the coefficients are on the diagonal. To convert these to SEs, we simply take the squared root. Now that we know where the regular SEs are coming from, let's get the heteroskedasticity-consistent SEs for this model from sandwich. The SEs come from the `vcovHC` function and the resulting object is the variance-covariance matrix for the coefficients. This is, again, a variance-covariance matrix for the coefficients. So, to get SES, we take the square root of the diagonal. (see Appendix all related results).



The summary output now reflects the correct SEs:

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-5.147e-02	1.538e-01	-0.346	0.729203
DuCrd	-4.396e-03	1.661e-03	-2.962	0.003126 **
CrdAm	-1.534e-05	8.077e-06	-2.251	0.024634 *
Age	1.067e-03	1.345e-03	0.827	0.408229
AcBa	9.879e-02	1.108e-02	9.100	< 2e-16 ***
Purp	4.691e-03	5.121e-03	0.972	0.331492
PaStPrCrd	6.566e-02	1.467e-02	4.733	2.54e-06 ***
ValSavSto	3.424e-02	8.313e-03	4.032	5.95e-05 ***
LenCurEmp	2.482e-02	1.221e-02	2.144	0.032297 *
InPerCe	-4.707e-02	1.376e-02	-3.595	0.000341 ***
SexMarSt	4.386e-02	1.973e-02	2.348	0.019091 *
Guara	5.878e-02	2.909e-02	2.117	0.034487 *
DurCurAddr	-2.859e-03	1.290e-02	-0.227	0.820664
MoValAvAs	-3.250e-02	1.431e-02	-2.261	0.024003 *
ConcurCrd	3.614e-02	2.066e-02	1.916	0.055607 .
TypAp	4.988e-02	2.917e-02	1.803	0.071679 .
NoCrdBa	-4.226e-02	2.707e-02	-1.662	0.096907 .
Occup	4.991e-03	2.377e-02	0.221	0.825502
NoDpnd	-2.930e-02	3.737e-02	-0.797	0.425529
Tlph	5.102e-02	2.876e-02	1.733	0.083470 .
FrgnWrkr	1.145e-01	5.699e-02	1.621	0.105444
Residual standard error:	0.4045 on 979 degrees of freedom			
Multiple R-squared:	0.2374			
Adjusted R-squared:	0.2218			
F-statistic:	15.24 on 20 and 979 DF			
p-value:	< 2.2e-16			



4. Subset Selection methods

4.1 Variable Selection

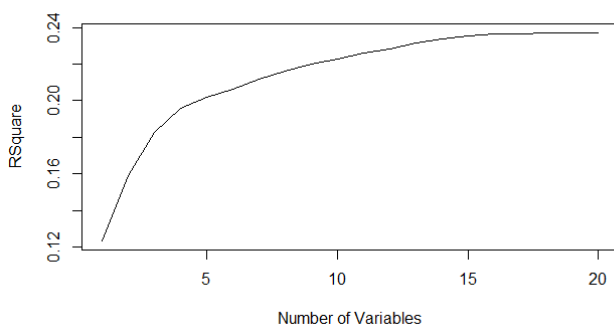
In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

Feature selection techniques are used for four reasons:

- Simplification of models to make them easier to interpret by researchers/users
- Shorter training times
- To avoid the curse of dimensionality
- Enhanced generalization by reducing overfitting (formally, reduction of variance) (Wikipedia, n.d.)

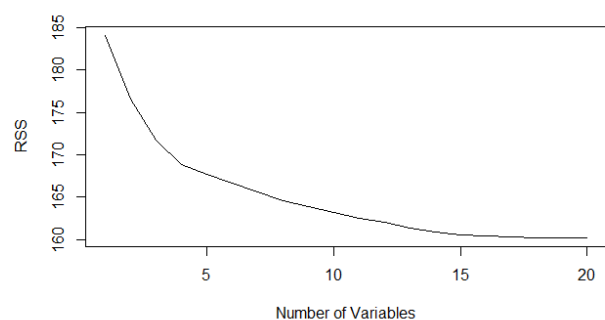
Using the appropriate code in **R** we have some measurers to select the best variable that we want to fit our data set. We are going to use 'rsq', 'rss', 'adjr2', 'cp' and 'bic'. (James, Witten, Hastie, & Tibshirani, 2013)

RSquared



When we include more than 8 variables, we reach almost the highest RSquared we can get for 23%. So, we will probably use 16 Variables for this measure.

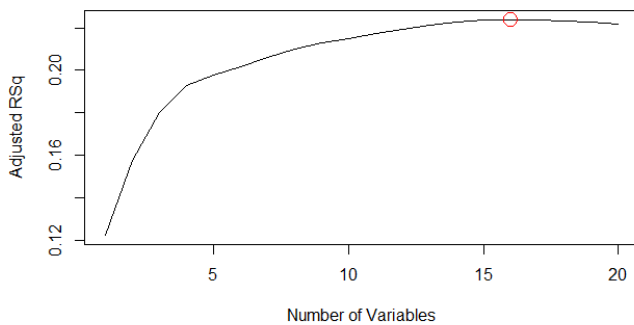
RSS



It is intended to minimize RSS. We want the lowest variable to include in model, hence we would say 16 or 4 because there is a steep slope.

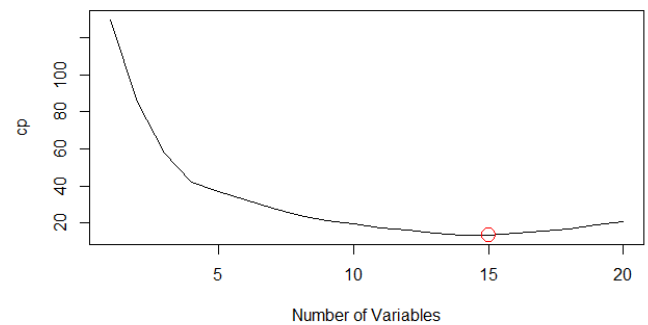


Adj r^2



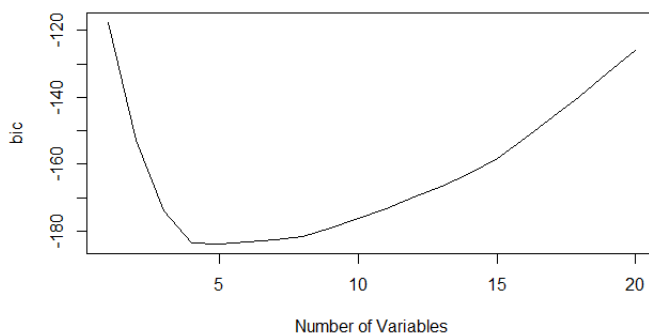
The highest Adj r^2 has 16 variables.

C p



The lowest Cp has 15 variables. So, for this measure when we include 15 variables, have the best model.

Bic

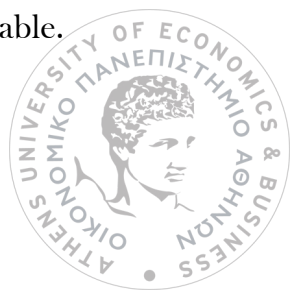


For this measure, when we include 5 variables we have the best model.
We have the lowest Bic when include 5 variables.

(Hastie, Tibshirani, & Friedman, May, 2001), (Wooldridge, 2006)

4.2 Forward Stepwise Selection

Stepwise selection is a method that allows moves in either direction, dropping or adding variables at the various steps. Stepwise regression can be achieved either by trying out one independent variable at a time and including it in the regression model if it is statistically significant, or by including all potential independent variables in the model and eliminating those that are not statistically significant, or by a combination of both methods. Tests for significance are conducted via F-tests, t-tests, adjusted R squared, and a few other less common methods. The goal is to find a set of independent variables which significantly influence the dependent variable. Conducting these tests automatically can potentially save time for the individual.



Forward stepwise selection is also a possibility, though not as common. In the forward approach, variables once entered may be dropped if they are no longer significant as other variables are added.

Forward selection is a very attractive approach, because it's both tractable and it gives a good sequence of models.

- Start with a null model. The null model has no predictors, just one intercept (The mean over Y).
- Fit p simple linear regression models, each with one of the variables in and the intercept. So basically, you just search through all the single-variable models the best one (the one that results in the lowest residual sum of squares). You pick and fix this one in the model.
- Now search through the remaining p minus 1 variables and find out which variable should be added to the current model to best improve the residual sum of squares.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold. (James, Witten, Hastie, & Tibshirani, 2013)

4.2.1 Forward Stepwise Selection Approach

We run the appropriate R code with our dataset and through the algorithm we take the “best” variables for our model. Because of using the C_p measure and it gave us that our model should include 15 variables, we keep the 15 strongest variables come from Forward Stepwise Selection. Naturally we could use another measure such as ‘rsq’, ‘rss’, ‘adjr2’ and ‘bic’.

Here, we ¹split the dataset into train and test set (60% training and 40% test).

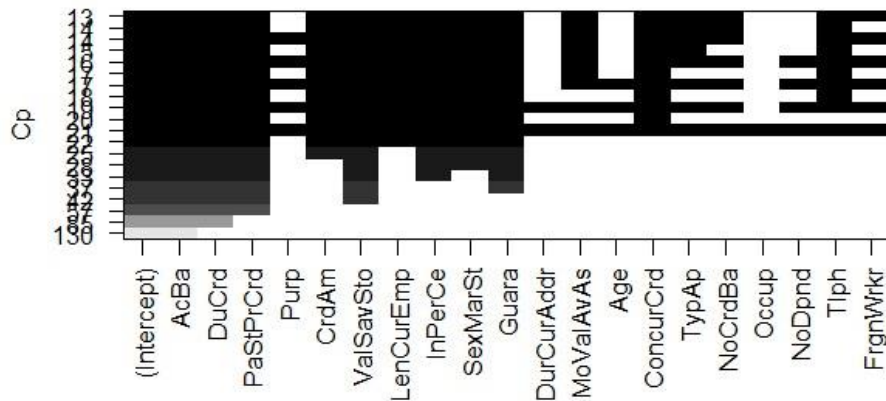
Hence, this method gives the following Variables:

AcBa, DuCrd, PaStPrCrd, CrdAm, ValSavSto, InPerCe, SexMarSt, Guara,
LenCurEmp, ConcurCrd, Tlph, MoValAvAs, TypAp, NoCrdBa, FrgnWrkr.

¹ In all sections and methods of the thesis, we split the dataset in the same way (60% training and 40% test)



The following plot can give a picture of the selected variables:



Plot shows the selected variables for the best model through the Forward Stepwise Method

To calculate the **train** and **test** means squared error we can use the linear regression model putting the 15 selected variables. We could also use another model for example a non-linear. Running R, we see that the **Test MSE= 0.2221826** and **Train MSE= 0.1510432**.

4.3 Backward Stepwise Selection

Backward stepwise selection involves starting off in a backward approach and then potentially adding back variables if they later appear to be significant. The process is one of alternation between choosing the least significant variable to drop and then re-considering all dropped variables (except the most recently dropped) for re-introduction into the model. This means that two separate significance levels must be chosen for deletion from the model and for adding to the model. The second significance must be more stringent than the first.

Namely, forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time. To be able to perform backward selection, we need to be in a situation where we have more observations than variables because we can do least squares regression when n is greater than p . If p is greater than n , we cannot fit a least squares model. It's not even defined.



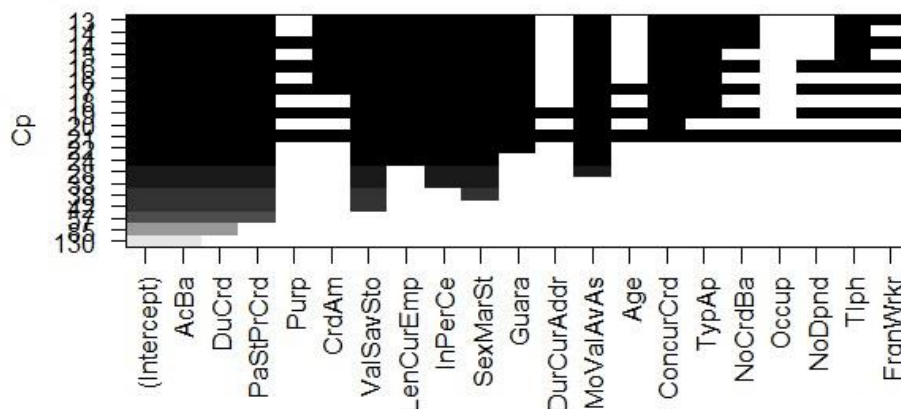
- Start with all variables in the model.
- Remove the variable with the largest p-value that is, the variable that is the least statistically significant.
- The new (p - 1)-variable model is t , and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold. (James, Witten, Hastie, & Tibshirani, 2013)

4.3.1 Backward Stepwise Approach

Here, we also run the R code and use the same measure Cp. For this reason, we continue to have 15 variables for the best model. We also split the dataset into train and test set keeping the same test proportion as Forward Stepwise Selection.

The backward stepwise method gives the following variables:

AcBa, DuCrd, PaStPrCrd, ValSavSto, InPerCe, SexMarSt, LenCurEmp, MoValAvAs, Guara, ConcurCrd, TypAp, CrdAm, Tlph, NoCrdBa, LenCurEm



Plot shows the selected variables for the best model through the Backward Stepwise Method

We also try the linear model with the above-mentioned variables, hence the **Test MSE= 0.2221996** and **Train MSE= 0.1643574**.



4.4 Cross Validation and Best Subset Selection Method

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples.

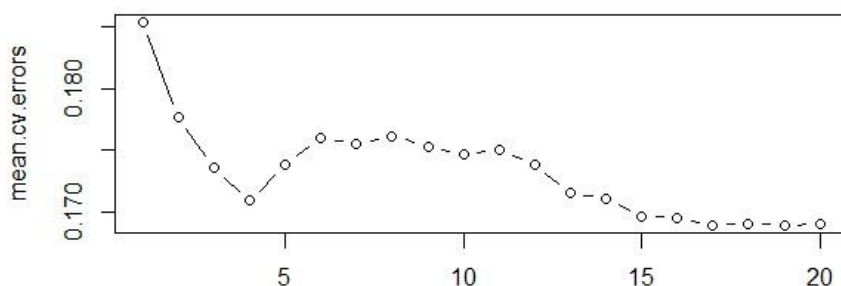
In this part, the sample is randomly partitioned into 10 equal size subsamples ($k=10$). Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 10-1 subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

Generally, cross-validation procedure is repeated n times, yielding n random partitions of the original sample.

The n results are again averaged (or otherwise combined) to produce a single estimation. (James, Witten, Hastie, & Tibshirani, 2013), (Hastie, Tibshirani, & Friedman, May, 2001)

4.4.1 Best Subset Selection Approach

Having used cross validation ($k=10$), the Best Subset Selection method gives the following interesting plot:



As we see that the minimum mean.cv.error is 4. From 4 up to 8 the graph illustrates an augmentation and then a decrease to 15, which is basically the minimum.

We will try both options. First, we will run a linear model with the 3 first strongest variables and then run one more time a linear model with 15 Variables. The dataset is splitted in the same way as Backward and Forward methods.

In the first option we use the 3 strongest Variables, AcBa, DuCrd, PaStPrCrd and run a liner regression. In this case, we have **Test MSE= 0.2246371 and **Train MSE= 0.1734628**. (We do it only to see the results because of the existence of this decrease in point 4. Actually, we do not use only three variables)**

In the second option we use the 15 strongest Variables, AcBa, DuCrd, PaStPrCrd, CrdAm, ValSavSto, LenCurEmp, InPerCe, SexMarSt, Guara, MoValAvAs, ConcurCrd, TypAp, NoCrdBa, Tlph, FrgnWrkr. In this case, we have **Test MSE= 0.2286232 and **Train MSE= 0.1611355**.**

4.5 Ridge Regression

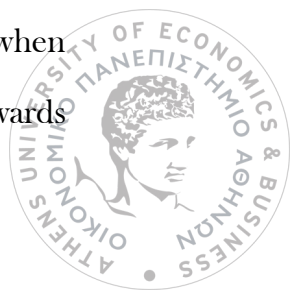
The least squares fitting procedure estimates $\beta_0, \beta_1, \dots, \beta_p$ using the values that minimize

$$RSS = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2$$

Ridge regression is very similar to least squares, except that the coefficients ridge are estimated by minimizing a slightly different quantity. In particular, the ridge regression coefficient estimates $\hat{\beta}_0$ R are the values that minimize

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = RSS + \lambda \sum_{j=1}^p \beta_j^2$$

Where $\lambda \geq 0$ is a tuning parameter, to be determined separately. As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small. However, the second term, $\lambda \sum \beta_i^2$, called a shrinkage penalty, is small when β_1, \dots, β_p are close to zero, and so it has the effect of shrinking the estimates of β_i towards



zero. The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates. When $\lambda=0$, the penalty term has no effect, and ridge regression will produce the least squares estimates. However, as $\lambda \rightarrow \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero. Unlike least squared estimates, which generates only one set of coefficient estimates, ridge regression will produce a different set of coefficient estimates, $\hat{\beta}_\lambda^R$, for each value of λ . Selecting a good value of λ is critical.

The shrinkage penalty is applied to β_1, \dots, β_p , but not to the intercept β_0 . We want to shrink the estimated association of each variable with the response; however, we do not want to shrink the intercept, which is simply a measure of the mean value of the response when $x_{i1} = x_{i2} = \dots = x_{ip} = 0$. If we assume that the variables—that is, the columns of the data matrix X —have been centered to have mean zero before ridge regression is performed, then the estimated intercept will take the form

$$\hat{\beta}_0 = \bar{y} = \sum_{i=1}^n y_i / n. \text{ (James, Witten, Hastie, \& Tibshirani, 2013)}$$

4.5.1 Ridge Regression Approach

Ridge Regression will give us the best λ via R. So, the best $\lambda=0.114304$ and we will use it to get the prediction on the test set. Namely, it is predicted Y using the test set. We see that the **MSE=0.1488859**

Ridge Coefficients:

Intercept)	AcBa	DuCrd	PaStPrCrd	Purp	CrdAm
6.144675e-02	8.207573e-02	-3.993675e-03	5.326140e-02	2.765153e-03	-1.196121e-05
ValSavSto	LenCurEmp	InPerCe	SexMarSt	Guara	DurCurAddr
2.957114e-02	2.171938e-02	-3.430186e-02	3.600842e-02	4.166183e-02	-3.512266e-03
MoValAvAs	Age	ConcurCrd	TypAp	NoCrdBa	Occup
-2.854295e-02	1.192410e-03	3.333577e-02	3.414358e-02	-2.291479e-02	1.658908e-03
NoDpnd	Tlph	FrgnWrkr			
-2.038924e-02	3.928789e-02	1.002559e-01			



4.6 Lasso Method

Lasso was introduced to improve the prediction accuracy and interpretability of regression models by altering the model fitting process to select only a subset of the provided covariates for use in the final model rather than using all of them. It was developed independently in geophysics, based on prior work that used the ℓ penalty for both fitting and penalization of the coefficients, and by the statistician, Robert Tibshirani based on Breiman's nonnegative garrote.

Prior to lasso, the most widely used method for choosing which covariates to include was stepwise selection, which only improves prediction accuracy in certain cases, such as when only a few covariates have a strong relationship with the outcome. However, in other cases, it can make prediction error worse. Also, at the time, ridge regression was the most popular technique for improving prediction accuracy. Ridge regression improves prediction error by shrinking large regression coefficients in order to reduce overfitting, but it does not perform covariate selection and therefore does not help to make the model more interpretable.

Lasso can achieve both of these goals by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value, which forces certain coefficients to be set to zero, effectively choosing a simpler model that does not include those coefficients. This idea is similar to ridge regression, in which the sum of the squares of the coefficients is forced to be less than a fixed value, though in the case of ridge regression, this only shrinks the size of the coefficients, it does not set any of them to zero.

General Form

Lasso regularization can be extended to a wide variety of objective functions such as those for generalized linear models, generalized estimating equations, proportional hazards models, and ²M-estimators in general, in the obvious way.

² Note: In statistics, M-estimators are a broad class of extremum estimators for which the objective function is a sample average. Both non-linear least squares and maximum likelihood estimation are special cases of M-estimators. The definition of M-estimators was motivated by robust statistics, which contributed new types of M-estimators. The statistical procedure of evaluating an M-estimator on a data set is called M-estimation.



Given the objective function:

$$\frac{1}{N} \sum_{i=1}^N f(x_i, y_i, \alpha, \beta)$$

the lasso regularized version of the estimator will be the solution to

$$\min_{\alpha, \beta} \frac{1}{N} \sum_{i=1}^N f(x_i, y_i, \alpha, \beta) \text{ subject to } \|\beta\|_1 \leq t$$

where only β penalized while α is free to take any allowed value. (James, Witten, Hastie, & Tibshirani, 2013), (Hastie, Tibshirani, & Friedman, May, 2001), (Wikipedia, n.d.)

4.6.1 Lasso Approach

We apply the same procedure on the Lasso. The best $\lambda=0.003572817$ and we use it to predict Y on the test set. The **MSE=0.1562664**.

Lasso Coefficients:

(Intercept)	AcBa	DuCrd	PaStPrCrd	Purp	CrdAm
1.208788e-01	9.618908e-02	-4.428161e-03	5.443169e-02	0.000000e+00	-8.877161e-06
ValSavSto	LenCurEmp	InPerCe	SexMarSt	Guara	DurCurAddr
2.849844e-02	1.776306e-02	-2.944218e-02	3.001855e-02	3.622153e-02	0.000000e+00
MoValAvAs	Age	ConcurCrd	TypAp	NoCrdBa	Occup
-2.441292e-02	5.272074e-04	2.477608e-02	2.179406e-02	-1.259050e-02	0.000000e+00
NoDpnd	Tlph	FrgnWrkr			
0.000000e+00	2.495584e-02	7.036923e-02			

We compare the Ridge MSE to Lasso MSE and it seems that Ridge Regression is doing better job because **0.1488859<0.1562664**.

One thing to point out is that on the Ridge regression some of insignificant coefficients can be very small. However, in the Lasso regression some coefficients are zero. As some coefficients shrink to zero, we can eliminate those variables. So, these are the following variables which are going to use to predict the Y variable:

(Intercept)	AcBa	DuCrd	PaStPrCrd	CrdAm	ValSavSto
1.208788e-01	9.618908e-02	-4.428161e-03	5.443169e-02	-8.877161e-06	2.849844e-02
LenCurEmp	InPerCe	SexMarSt	Guara	MoValAvAs	Age
1.776306e-02	-2.944218e-02	3.001855e-02	3.622153e-02	-2.441292e-02	5.272074e-04
ConcurCrd	TypAp	NoCrdBa	Tlph	FrgnWrkr	
2.477608e-02	2.179406e-02	-1.259050e-02	2.495584e-02	7.036923e-02	

(Hastie, Tibshirani, & Friedman, May, 2001)



4.7 Principal Component Regression

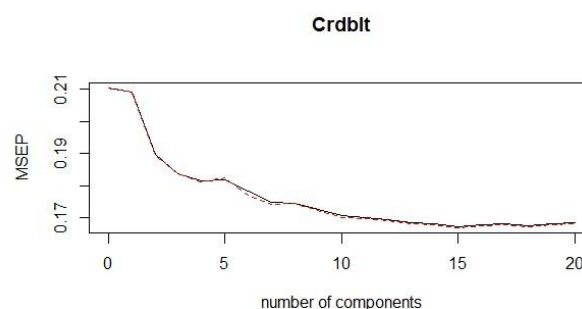
In statistics, principal component regression is a regression analysis technique that is based on principal component analysis. Typically, it considers regressing the outcome on a set of covariates based on a standard linear regression model.

In PCR, instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables are used as regressors. One typically uses only a subset of all the principal components for regression, thus making PCR some kind of a regularized procedure. Often the principal components with higher variances are selected as regressors. However, for the purpose of predicting the outcome, the principal components with low variances may also be important, in some cases even more important.

One major use of PCR lies in overcoming the multicollinearity problem which arises when two or more of the explanatory variables are close to being collinear. PCR can aptly deal with such situations by excluding some of the low-variance principal components in the regression step. In addition, by usually regressing on only a subset of all the principal components, PCR can result in dimension reduction through substantially lowering the effective number of parameters characterizing the underlying model. This can be particularly useful in settings with high-dimensional covariates. Also, through appropriate selection of the principal components to be used for regression, PCR can lead to efficient prediction of the outcome based on the assumed model. (James, Witten, Hastie, & Tibshirani, 2013)

4.7.1 Principal Component Approach

We use cross-validation:



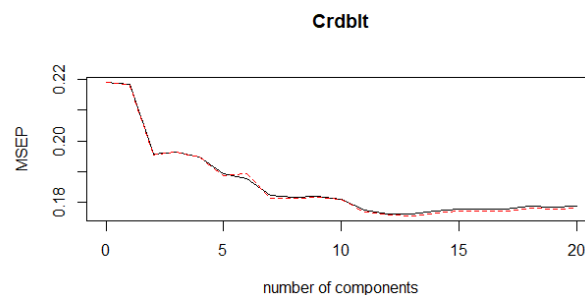
VALIDATION: RMSEP

Cross-validated using 10 random segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps
CV	0.4679	0.4671	0.4424	0.4431	0.4412	0.4353	0.4332	0.4272	0.4262
adjCV	0.4679	0.4670	0.4421	0.4431	0.4414	0.4343	0.4351	0.4259	0.4258
	9 comps	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps	16 comps	
CV	0.4266	0.4254	0.4213	0.4200	0.4199	0.4210	0.4218	0.4216	
adjCV	0.4263	0.4255	0.4206	0.4196	0.4192	0.4204	0.4211	0.4210	
	17 comps	18 comps	19 comps	20 comps					
CV	0.4218	0.4230	0.4225	0.4230					
adjCV	0.4212	0.4223	0.4218	0.4222					

From the above plot it is seen that the 15th component has the lowest MSE. It can also be seen from the table above, where the 15th component is 0.4218 and is the lowest.

Now we use cross-validation on the train set and have the following plot:



VALIDATION: RMSEP

Cross-validated using 10 random segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps
CV	0.4679	0.4671	0.4424	0.4431	0.4412	0.4353	0.4332	0.4272	0.4262
adjCV	0.4679	0.4670	0.4421	0.4431	0.4414	0.4343	0.4351	0.4259	0.4258
	9 comps	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps	16 comps	
CV	0.4266	0.4254	0.4213	0.4200	0.4199	0.4210	0.4218	0.4216	
adjCV	0.4263	0.4255	0.4206	0.4196	0.4192	0.4204	0.4211	0.4210	
	17 comps	18 comps	19 comps	20 comps					
CV	0.4218	0.4230	0.4225	0.4230					
adjCV	0.4212	0.4223	0.4218	0.4222					

In the table as in the plot above, it seems that the 13th component has the lowest MSE.

Hence, using 13 components, R gives that **MSE= 0.1566044**.

All in all, we resulted in **0.1566044**, because we took the lowest MSE into account.

Of course, we should change the number of the ³ components depending on the dimension and we could choose less than 13 or more. For example, if we are interested in almost 85% of X explained, then probably we will use 14 components.

³ Note: Each component is NOT a Variable, is some variance of 20 Variables.



See the following table:

TRAINING: % variance explained										
	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps	
X	13.376	23.10	30.45	37.15	43.56	49.60	55.38	60.61	65.43	
Crdblt	1.094	10.26	10.26	10.29	11.18	16.36	17.58	17.92	18.33	
	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps	16 comps	17 comps	18 comps	
X	70.09	74.46	78.33	81.98	85.43	88.69	91.71	94.34	96.74	
Crdblt	19.47	20.95	20.99	21.37	22.05	22.05	22.55	22.57	22.57	
	19 comps	20 comps								
X	98.85	100.00								
Crdblt	23.14	23.34								

Then, R gives **MSE=0.157955**

4.8 Partial Least Squares

Partial Least Squares Regression is an extension of the multiple linear regression model (see, e.g., Multiple Regression or General Stepwise Regression). In its simplest form, a linear model specifies the (linear) relationship between a dependent (response) variable Y , and a set of predictor variables, the X 's, so that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p.$$

In this equation β_0 is the regression coefficient for the intercept and the β_i values are the regression coefficients (for variables 1 through p) computed from the data.

For example, you could estimate (predict) a person's weight as a function of the person's height and gender. You could use linear regression to estimate the respective regression coefficients from a sample of data, measuring height, weight, and observing the subjects' gender. For many data analysis problems, estimates of the linear relationships between variables are adequate to describe the observed data, and to make reasonable predictions for new observations.

The multiple linear regression model has been extended in several ways to address more sophisticated data analysis problems. The multiple linear regression model serves as the basis for a number of multivariate methods such as discriminant analysis (i.e., the prediction of group membership from the levels of continuous predictor variables), principal components regression (i.e., the prediction of responses on the dependent variables from factors underlying the levels of the predictor variables), and canonical correlation (i.e., the prediction of factors underlying responses on the dependent variables from factors underlying the levels of the predictor variables). These multivariate methods all have two important properties in common. These methods impose restrictions such that factors



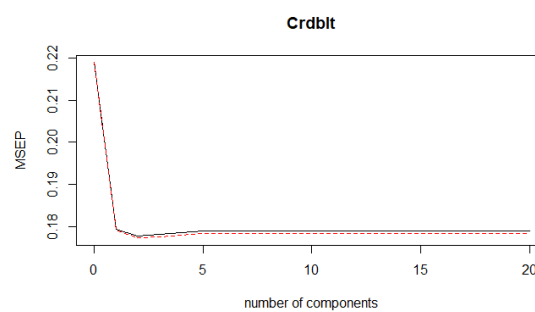
underlying the Y and X variables are extracted from the $Y'Y$ and $X'X$ matrices, respectively, and never from cross-product matrices involving both the Y and X variables, and the number of prediction functions can never exceed the minimum of the number of Y variables and X variables.

Partial Least Squares Regression extends multiple linear regression without imposing the restrictions employed by discriminant analysis, principal components regression, and canonical correlation. In partial least squares regression, prediction functions are represented by factors extracted from the $Y'XX'Y$ matrix. The number of such prediction functions that can be extracted typically will exceed the maximum of the number of Y and X variables.

In short, partial least squares regression is probably the least restrictive of the various multivariate extensions of the multiple linear regression model. This flexibility allows it to be used in situations where the use of traditional multivariate methods is severely limited, such as when there are fewer observations than predictor variables. Furthermore, partial least squares regression can be used as an exploratory analysis tool to select suitable predictor variables and to identify outliers before classical linear regression. (James, Witten, Hastie, & Tibshirani, 2013).

4.8.1 Partial Least Squares Approach

Here we also use cross-validation on the train set and have the following plot:



The 2th component has the lowest MSE. It can also be seen from the table below as well as from the above plot that the 2th component is 0.426 and is the lowest.

Hence, using 13 components, R gives that **MSE= 0.1561894**



VALIDATION: RMSEP

Cross-validated using 10 random segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps
CV		0.4236	0.4216	0.4220	0.4226	0.4229	0.4230	0.4230	0.4230
adjCV	0.4679	0.4232	0.4210	0.4213	0.4218	0.4222	0.4222	0.4222	0.4222
	9 comps	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps	16 comps	
CV	0.4230	0.4230	0.4230	0.4230	0.4230	0.4230	0.4230	0.4230	
adjCV	0.4222	0.4222	0.4222	0.4222	0.4222	0.4222	0.4222	0.4222	
	17 comps	18 comps	19 comps	20 comps					
CV	0.4230	0.4230	0.4230	0.4230					
adjCV	0.4222	0.4222	0.4222	0.4222					

However, we could also change the numbers of components components depending on the dimension and we could choose less than 2 or more. We will use 15 components because we can look at the percentage of variance explained. The first 15 components explain almost 80% (79.61%) of X variable and as a result R gives **MSE= 0.1550208**.

See the following table:

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps
X	9.315	17.15	27.30	32.73	36.50	41.08	44.41	48.42	53.00
Crdbl1t	20.931	23.16	23.41	23.50	23.51	23.51	23.51	23.51	23.51
	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps	16 comps	17 comps	
X	57.64	63.16	67.27	70.89	75.42	79.61	84.48	88.76	
Crdbl1t	23.51	23.51	23.51	23.51	23.51	23.51	23.51	23.51	
	18 comps	19 comps	20 comps						
X	93.29	96.29	100.00						
Crdbl1t	23.51	23.51	23.51						

(Chong & Jun, 2004), (Mehmood, Liland, Snipen, & Sæbø, 2012), (Guyon & Elisseeff, 2003), (Murtaugh, 2009)

4.9 Conclusion

Linear model:

As we saw, the dataset consists of 21 variables which characterize 1000 loan applicants. 30% of those is not credit worthy (“bad”) and 70% is credit worthy (“good”) applicants. We started assuming that the model is linear and examined if it meets linearity assumptions. Through regression diagnostics graphs, it is illustrated that the model is heteroscedastic and non-normal. Using the Breusch-Pagan test, we also showed that the model is heteroscedastic. Next, the SEs have been corrected since heteroscedasticity affects only them.

Non-normality of the model is illustrated in the Q-Q plot as well as in the distribution of standardized residuals graph, which is skewed to the left.



Subset selection methods:

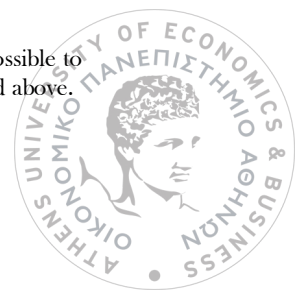
Since all above mentioned subset selection methods have been explained, we have to decide which variables should be included in the model. Actually, there is not a specific answer. Each editor evaluates all results and decides according his estimation which variables will keep. Our goal is to get out the variables which we think as unnecessary. Initially, the first thought is to prefer the Subset Selection method which decreases the test MSE and as a result train MSE. However, in this case, the possibility of overfitting rises. Overfitting the model generally takes the form of making an overly complex model to explain idiosyncrasies in the data under study. In reality, the data often studied has some degree of error or random noise within it. Thus, we attempt to make the model conforms too closely to slightly inaccurate data which can infect the model with substantial errors and reduce its predictive power.

So, we do not select ⁴Ridge Regression and Lasso because MSEs are low enough and we are skeptical about that. Additionally, Ridge Regression takes all variables into account and as a results MSE is very low. Regarding LASSO we could say that it is a method which rejects the insignificant variables and it keeps 16 variables. Although the MSE of this method is low enough, if we will try the 16 Variables using linear regression, it gives test MSE=0.235116 and train MSE=0.1691912 which are higher enough than BEST. PCR and PLS have almost the same train MSE, but a bit higher than Forward, Backward and Best. Forward has the same test MSE as Backward and they are slightly lower than BEST. In this way, we can choose any of the last tree methods because they give the lowest test MSEs and the same variables as the most appropriate ones.

Consequently, we select the 15 following Variables for our model:

AcBa, DuCrd, PaStPrCrd, CrdAm, ValSavSto, LenCurEmp, InPerCe, SexMarSt, Guara, MoValAvAs, ConcurCrd, TypAp, NoCrdBa, Tlph, FrgnWrkr.

⁴ Note: Although we chose those variables which Forward, Backward and Best selection give, it would be also possible to used only Ridge and Lasso, that is 16 Variables, instead of examining the rest Subset Selection Methods mentioned above.



5. Discriminative Models

Discriminative Model also called conditional models, are a class of models used in machine learning for modelling the dependence of unobserved variables y on observed variables x . Within a probabilistic framework, this is done by modeling the conditional probability distribution $P(y | x)$, which can be used for predicting y from x . Discriminative models, as opposed to generative models, do not allow one to generate samples from the joint distribution of observed and target variables. However, for tasks such as classification and regression that do not require the joint distribution, discriminative models can yield superior performance. On the other hand, generative models are typically more flexible than discriminative models in expressing dependencies in complex learning tasks. In addition, most discriminative models are inherently supervised and cannot easily support unsupervised learning. Application-specific details ultimately dictate the suitability of selecting a discriminative versus generative model. Such model is the logistic regression.

5.1 The logistic Regression (Linear Classifier)

In statistics, the logistic model (or logit model) is a statistical model that is usually taken to apply to a binary dependent variable. In regression analysis, logistic regression or logit regression is estimating the parameters of a logistic model. More formally, a logistic model is one where the log-odds of the probability of an event is a linear combination of independent or predictor variables. The two possible dependent variable values are often labelled as "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. The binary logistic regression model can be generalized to more than two levels of the dependent variable: categorical outputs with more than two values are modelled by multinomial logistic regression, and if the multiple categories are ordered, by ordinal logistic regression, for example the proportional odds ordinal logistic model.



5.2 Why Logistic Regression

Equation of straight line: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \rightarrow$ Range of Y is from $-(\infty)$ to ∞ .

Let's try to reduce the Logistic Equation from this equation.

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \rightarrow$ In Logistic Regression Y cannot be between 0 and.

Now to get the range of Y between 0 and infinity, let's transform Y \rightarrow

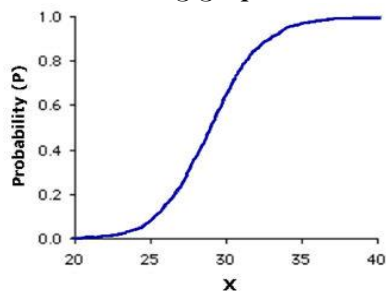
$\frac{Y}{1-Y} \left\{ \begin{array}{l} Y=0 \mid 0 \\ Y=1 \mid \infty \end{array} \right.$ Now we have the range between 0 and infinity.

Let us transform it further, to get the range between $-(\infty)$ and ∞ \rightarrow

$\log(Y/1-Y) \rightarrow \log(Y/1-Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$

Logistic or Logit Regression Model is a regression model where the dependent variable (Crdbl) is categorical, namely variables can be only fixed values such as A, B or C or Yes or No.

The following graph shows a Logit graphic:



(James, Witten, Hastie, & Tibshirani, 2013), (Hastie, Tibshirani, & Friedman, May, 2001)

5.3 How the Logistic Regression works with our Dataset

Let's take our sample dataset in R, which is called ger_cre in which we have kept the 15 Variables. Our aim is to predict if a loan applicant will default or become delinquent. So, we have the response **Creditability** (Crdbl) which can be 'worthy'=1 or 'not worthy'=0.



The formula to predict a logit transformation of the probability of presence of the characteristic of Creditability is: $\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$, where p is the probability of presence of the characteristic of Creditability.

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

Estimating the regression coefficients:

The coefficients β_0 and β_1 are unknown, and must be estimated based on the available training data. In this case maximum likelihood is preferred, since it has better statistical properties. The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows: we seek estimates for β_0 and β_1 such that the predicted probability $p(x_i)$ to be each applicant ‘good’ or ‘bad’, corresponds as closely as possible to the individual’s observed creditability status. In other words, we try to find β_0 and β_1 such that plugging these estimates into the model for $P(X)$, yields a number close to 1 for all applicants who are ‘good’ and close to 0 for those who are not. This intuition can be formalized using a mathematical equation called a likelihood function.

Logistic regression and other models can be easily fit using a statistical software package such as R and so we do not need to concern ourselves with the details of the maximum likelihood fitting procedure. (James, Witten, Hastie, & Tibshirani, 2013), (Hastie, Tibshirani, & Friedman, May, 2001)

Before creating a model, we divide our dataset into training and testing.

Running `summary(model)` via R, we take the value of our coefficients:

Deviance Residuals:				
Min	1Q	Median	3Q	Max
-2.6516	-0.7794	0.4512	0.7459	2.0994



Coefficients:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.919e+00	9.543e-01	4.106	4.02e-05 ***
AcBa	5.858e-01	6.946e-02	8.434	< 2e-16 ***
DuCrd	-2.423e-02	8.592e-03	-2.820	0.004799 **
PaStPrCrd	3.772e-01	8.630e-02	4.371	1.23e-05 ***
CrdAm	-9.144e-05	3.967e-05	-2.305	0.021156 *
ValSavSto	2.362e-01	5.780e-02	4.086	4.39e-05 ***
LenCurEmp	1.615e-01	6.767e-02	2.386	0.017021 *
NoCrdBa	-2.253e-01	1.575e-01	-1.431	0.152503
InPerCe	-2.848e-01	8.145e-02	-3.496	0.000471 ***
SexMarSt	2.384e-01	1.141e-01	2.090	0.036615 *
Guara	3.459e-01	1.778e-01	1.945	0.051748.
MoValAvAs	-1.879e-01	8.824e-02	-2.130	0.033201 *
ConcurCrd	2.323e-01	1.100e-01	2.113	0.034605 *
TypAp	3.358e-01	1.587e-01	2.116	0.034309 *
Tlph	3.391e-01	1.765e-01	1.921	0.054709 .
FrgnWrkr	1.137e+00	6.110e-01	1.861	0.062759 .

Null deviance:	1221.73 on 999 degrees of freedom
Residual deviance:	959.32 on 984 degrees of freedom
AIC:	991.32

From the above table, firstly, we can see the significant codes. These significant codes specify how much significant our independent variables are. Consequently, it seems how much significant our dataset is. Basically, our dataset is significant as most Variables have significance level over 90%. (*** 99.9%, ** 99%, * 95%, . 90%), except from NoCrdBa which is insignificant. So, we do not need optimize our model that is to remove NoCrdBa variable.

Null Deviance shows how well the response Variable is predicted by a model that includes only the intercept. Residual Deviance shows how well the response variable is predicted with inclusion of independent variables.

To do a prediction as to whether the loan applicant will be “good or “bad”, we must convert these predicted probabilities into class labels, “1” or “0”.

CrdBlt		
logpred	0	1
0	141	73
1	159	627

correct prediction: $(141+627)/1000=0.768$
error rate: 0.232



As we can see our model correctly predicted that 627 loan applicants will be “good” and “bad” 141, for a total of $141+627=768$ correct predictions.

Hence, logistic regression predicted 76.8% correctly which is relatively high percentage. However, we should examine out dataset by splitting it, because we trained and tested the model on the same set of 1000 observations.

To better assess the accuracy of the logistic regression model in this setting, we can fit the model using part of the data, and then examine how well it predicts the held-out data. This will yield a more realistic error rate, in the sense that in practice we will be interested in our model’s performance not on the data that we used to fit the model, but rather on these that will happen in the future and are unknown.

Trying the train data set

	Crdblttrain	
logpred	0	1
0	79	41
1	96	395

correct prediction: $(79+395)/611=0.776$

error rate: 0.224

Now trying the test data set

	Crdblttest	
logpred1	0	1
0	63	31
1	64	249

The results are rather good: the test error rate is 23.3%.

Hence, Logistic Model is 76.7% accurate.

(Louzada, Ara, & Fernandes, 2013), (A. Abdou & Pointon, 2011), (James, Witten, Hastie, & Tibshirani, 2013)

We can also examine the same method taking different thresholds into account:

1.Threshold 0.3:

Train data:

	Crdblttrain	
logpred5	0	1
0	40	9
1	143	424

correct prediction: 0.753

error rate: 0.247

Test data:

	Crdblttest	
logpred1	0	1
0	37	10
1	87	273

correct prediction: 0.762

error rate: 0.238



2.Threshold 0.7:

Train data:

	Crdbltrain		
logpred5	0	1	
	0	123	127
	1	53	303

correct prediction: 0.703

error rate: 0.297

Test data:

	Crdbltest		
logpred1	0	1	
	0	81	81
	1	37	206

correct prediction: 0.709

error rate: 0.291



6. Generative Model

Generative modeling is the use of artificial intelligence, statistics and probability in applications to produce a representation or abstraction of observed phenomena or target variables that can be calculated from observations. Generative modeling is used in unsupervised machine learning as a means to describe phenomena in data, enabling computers to understand the real world. This artificial intelligence understanding can be used to predict all manner of probabilities on a subject from modeled data. An example of such model is the discriminant analysis model.

6.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA), normal discriminant analysis or discriminant function analysis is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

LDA is closely related to analysis of variance (ANOVA) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements. However, ANOVA uses categorical independent variables and a continuous dependent variable, whereas discriminant analysis has continuous independent variables and a categorical dependent variable. Logistic regression and probit regression are more similar to LDA than ANOVA is, as they also explain a categorical variable by the values of continuous independent variables. These other methods are preferable in applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA method.

LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor



analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made.

LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis.

Discriminant analysis is used when groups are known a priori (unlike in cluster analysis). Each case must have a score on one or more quantitative predictor measures, and a score on a group measure. In simple terms, discriminant function analysis is classification - the act of distributing things into groups, classes or categories of the same type.

Summarizing the LDA approach in 5 steps

Listed below are the 5 general steps for performing a linear discriminant analysis:

1. Compute the d-dimensional mean vectors for the different classes from the dataset.
2. Compute the scatter matrices (in-between-class and within-class scatter matrix).
3. Compute the eigenvectors (e_1, e_2, \dots, e_d) and corresponding eigenvalues ($\lambda_1, \lambda_2, \dots, \lambda_d$) for the scatter matrices.
4. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a $d \times k$ dimensional matrix W (where every column represents an eigenvector).

Use this $d \times k$ eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: $Y = X \times W$ (where X is a $n \times d$ -dimensional matrix representing the n samples, and y are the transformed $n \times k$ -dimensional samples in the new subspace). (James, Witten, Hastie, & Tibshirani, 2013), (Louzada, Ara, & Fernandes, 2013), (A. Abdou & Pointon, 2011), (Hastie, Tibshirani, & Friedman, May, 2001)



6.1.2 How the Linear Discriminant Analysis works with our Dataset

Firstly, we use the linear discriminant function and through R we take the following train results:

Prior probabilities of groups:	
0	0.2906977
1	0.7093023

Group means:						
	AcBa	DuCrd	PaStPrCrd	CrdAm	ValSavSto	
0	1.925714	24.5657	2.217143	4006.314	1.697143	
1	2.93911	19.1663	2.702576	2938.721	2.229508	
	LenCurEmp	InPerCe	SexMarSt	Guara	MoValAvAs	
0	3.194286	3.01714	2.525714	1.091429	2.588571	
1	3.440281	2.87822	2.737705	1.133489	2.285714	
	ConcurCrd	TypAp	NoCrdBa	Tlph	FrgnWrkr	
0	2.571429	1.92	1.36	1.371429	1.005714	
1	2.711944	1.95082	1.416862	1.405152	1.053864	

Coefficients of linear discriminants:	
LD1	
AcBa	0.564937303
DuCrd	-0.012710697
PaStPrCrd	0.268137465
CrdAm	-0.000133522
ValSavSto	0.174240854
LenCurEmp	0.085968087
InPerCe	-0.217122673
SexMarSt	0.320839621
Guara	0.423803912
MoValAvAs	-0.128512999
ConcurCrd	0.16447292
TypAp	0.286250291
NoCrdBa	-0.143801835
Tlph	0.271674527
FrgnWrkr	0.790907047

Here, we see the proportion of each category (“0”: 29% and “1”: 71%).

The Group means shows the mean of each variable in each group.

First discriminant function (LD1) is a linear combination of the 15 predictor variables:

$0.5649373028 * \text{AcBa} + (-0.0127106973 * \text{DuCrd}) + \dots + 0.7909070470 * \text{FrgnWrkr}$.



Next let's evaluate the prediction accuracy of our model. Firstly, we'll run the model against the **training set** used to verify the model fits the data properly by using the command predict. The table output below is a confusion matrix with the actual values at the row labels and the predicted values at the column labels. We use threshold 0.5.

	Actual	
Predicted	0	1
0	81	47
1	94	380

$(89+384)/602 = 0.7657807$: correct prediction

The total number of correctly predicted observations is the sum of the diagonal. So, this model fit the training data correctly for almost every observation. Verifying the training set doesn't prove accuracy, but a poor fit to the **training data** could be a sign that the model isn't a good one.

Now let's run our **test set** against this model to determine its accuracy.

	Actual	
Predicted	0	1
0	63	28
1	62	245

$(63+245)/398 = 0.7738693$: correct prediction

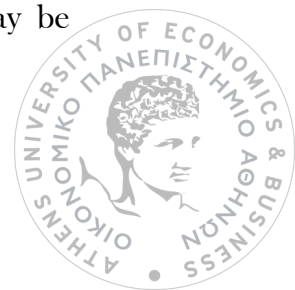
$(62+28)/398 = 0.22613$: error rate

Let's examine the accuracy of our test set.

In practice, a binary classifier such as this one can make two types of category, or it can incorrectly assign an applicant who is "bad" to "good" category. It is often of interest to determine which of these two types of errors are being made.

A confusion matrix, shown for the Creditability data, is a convenient way to display this information. The matrix table reveals that LDA predicted that a total of 307 people would be "good". Of these people, 245 actually are "good" and 63 are not. Hence, 63 out of 125 (or 50.4%) of the individuals who are "bad" are incorrectly labelled. However, of the 273 individuals who are "good", 28 (or 10.25%) are missed by LDA. The overall error rate is 23%. So, although the overall error rate is not so high, the error rate among individuals who are "bad" is 50.4%.

From the perspective of a bank that is trying to identify high-risk individuals, an error rate of $63/125 = 50.4\%$ among individuals who are "bad" loan applicant may be unacceptable.



However, if we are concerned about incorrectly predicting the creditability status for individuals who are “good”, then we can consider lowering this threshold. For instance, we might label any applicant with a ⁵posterior (*) probability of being “good” above 30% to the default class:

	Actual	
	0	1
FALSE	31	202
TRUE	94	71

* FALSE=1, TRUE=0

Here, LDA predicts that 165 individuals will be “bad”. Of the 273 individuals who are “good”, LDA correctly predicts all but 71, or 26%. This is not an improvement toward the error rate of 10.25% that resulted from using the threshold of 50%.

	Actual	
	0	1
FALSE	55	240
TRUE	70	33

→ when $P(\text{good}=1 \mid X=x) \geq 0.47$

So, after trying via R we conclude that the $P(\text{good}=1 \mid X=x) \geq 0.47$ or 47%, because we have the lower total error rate.

Hence, the LDA prediction is 77.889% accurate.

⁵ Note: Posterior probability is the revised probability of an event occurring after taking into consideration new information. Posterior probability is calculated by updating the prior probability by using Bayes' theorem. In statistical terms, the posterior probability is the probability of event A occurring given that event B has occurred.)



7. Quadratic Discriminant Analysis

Quadratic discriminant analysis (QDA) is closely related to linear discriminant analysis (LDA), where it is assumed that the measurements from each class are normally distributed. Unlike LDA however, in QDA there is no assumption that the covariance of each of the classes is identical. When the normality assumption is true, the best possible test for the hypothesis that a given measurement is from a given class is the likelihood ratio test. (James, Witten, Hastie, & Tibshirani, 2013), (Yap, Ong, & Husain, Using data mining to improve assessment of credit worthiness via credit scoring models, 2011), (VOJTEK & KOČENDA*, 2005)

7.1 How Quadratic Discriminant Analysis works with our Dataset

Group means:						
		AcBa	DuCrd	PaStPrCrd	CrdAm	ValSavSto
	0	1.903333	24.86	2.166667	3938.127	1.673333
	1	2.865714	19.20714	2.707143	2985.443	2.29
		LenCurEmp	InPerCe	SexMarSt	Guara	MoValAvAs
	0	3.17	3.096667	2.586667	1.126667	2.586667
	1	3.475714	2.92	2.722857	1.152857	2.26
		ConcurCrd	TypAp	NoCrdBa	Tlph	FrgnWrkr
	0	2.556667	1.913333	1.366667	1.376667	1.013333
	1	2.725714	1.934286	1.424286	1.415714	1.047143

The output contains the group means. But it does not contain the coefficients of the linear discriminants, because the QDA classifier involves a quadratic, rather than a linear, function of the predictors.

Taking all observations into account, the table with the correct prediction is shown:

	CrdBlt	
qda.class	0	1
0	203	129
1	97	571

Here, the correct prediction is 0.774.

Confusion matrix for the **train data**:

	Actual	
	0	1
FALSE	93	379
TRUE	101	60

correct prediction is 0.758.



Yet, if we want to compare QDL with LDA we should show the table with **test data** which have the lowest error rate:

	Actual	
	0	1
FALSE	45	263
TRUE	62	41

error rate=0.21

correct prediction=0.79

So, the QDA prediction is 79% accurate.



8. Classification Methods

Classification is a data mining task of predicting the value of a categorical variable by building a model based on one or more numerical and/or categorical variables (predictors or attributes). Data mining is a critical step in knowledge discovery involving theories, methodologies and tools for revealing patterns in data. It developed in fields other than statistics, e.g., machine learning and signal processing, are also introduced. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class. Often the methods used for classification first predict the probability of each of the categories of a qualitative variable, as the basis for making the classification. In this sense they also behave like regression methods.

We will examine the following classification methods: k-nearest neighbors, tree-based methods: decision trees, random forest.

8.1 K-Nearest Neighbors

The KNN algorithm is a robust and versatile classifier that is often used as a benchmark for more complex classifiers such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM). Despite its simplicity, KNN can outperform more powerful classifiers and is used in a variety of applications such as economic forecasting, data compression and genetics.

As we mentioned k-nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. This algorithm segregates unlabeled data points into well-defined groups. Choosing the number of nearest neighbors i.e. determining the value of k plays a significant role in determining the efficacy of the model. Thus, selection of k will determine how well the data can be utilized to generalize the results of the k-NN algorithm. A large k value has benefits which include reducing the variance due to the noisy data; the side effect being developing a bias due to which the learner tends to ignore the smaller patterns which may have useful insights.



KNN has been around for a long time and has been very well studied. As such, different disciplines have different names for it, for example:

- **Instance-Based Learning:** The raw training instances are used to make predictions. As such KNN is often referred to as instance-based learning or a case-based learning (where each training instance is a case from the problem domain).
- **Lazy Learning:** No learning of the model is required, and all of the work happens at the time a prediction is requested. As such, KNN is often referred to as a lazy learning algorithm.
- **Non-Parametric:** KNN makes no assumptions about the functional form of the problem being solved. As such KNN is referred to as a non-parametric machine learning algorithm.

K-NN Pros and Cons

Pros: The algorithm is highly unbiased in nature and makes no prior assumption of the underlying data. Being simple and effective in nature, it is easy to implement and has gained good popularity.

Cons: k-NN algorithm has drawn a lot of flake for being extremely simple! If we take a deeper look, this doesn't create a model since there's no abstraction process involved. Yes, the training process is fast as the data is stored verbatim (hence lazy learner) but the prediction time is pretty high with useful insights missing at times. Therefore, building this algorithm requires time to be invested in data preparation (especially treating the missing data and categorical features) to obtain a robust model. (James, Witten, Hastie, & Tibshirani, 2013)

8.1.1 How the K-NN works with our Dataset

The first thing we need to do is the loading of our dataset in CSV format. Via R we construct the train and test data set. Then, we predict on a test set of 400 observations and the rest (600) is used as train set.

Let's use an odd number near the square root of the observations size (1000) of our data set. The result is 31.



It is examined the KNN for the **test data set** with the following tables:

We will first try for **k=31** and see how it works in our model.

test.crd		
knn. 31	0	1
0	7	113
1	5	275

$$282/400=0.705$$

Using **k = 31**, 70.5% of the observations is correctly predicted.

Let's use k values as 1, 5 and 20 to see how the perform in terms of correct proportion of classification a success rate.

For k=1: 211

For k=5: 224

For k=20: 257

We can also perform the table for **k=1**:

test.crd		
knn. 1	0	1
0	6	183
1	6	205

$$211/400= 0.5275$$

The results using k = 1 are not very good, since only 52.75% of the observations are correctly predicted.

k=1 overfits and this can only be seen in test error rate which is 47.25%.

Below, we repeat the analysis using **k=5**:

test.crd		
knn. 5	0	1
0	7	171
1	5	217

$$224/400=0.56$$

We see that the results have improved slightly. Now 56% of the observations are correctly predicted.

Finally, we perform the table using **k=20**:

test.crd		
knn. 20	0	1
0	6	137
1	6	251

$$257/400=0.6425$$

Here, the results have improved more, and the prediction is 64.25%.

However, if we try until k=55 the prediction is 75.75%:

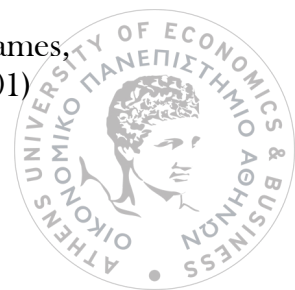
test.crd		
knn. 55	0	1
0	2	86
1	10	302

And if we continue with k=57 we see that the prediction is not improved but decreased to 74.25%.

Hence, we could say that the most appropriate prediction is 75.75% whose k=55.

(See Appendix the train data)

(VOJTEK & KOČENDA *, 2005), (Louzada, Ara, & Fernandes, 2013), (James, Witten, Hastie, & Tibshirani, 2013), (Hastie, Tibshirani, & Friedman, May, 2001)



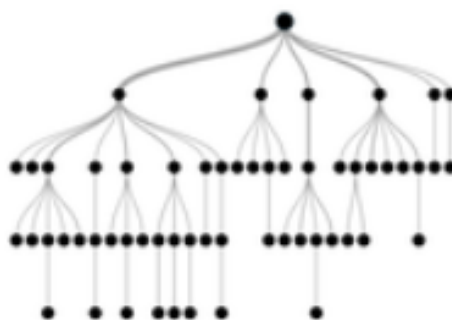
8.2 Tree-Based Methods

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression).

8.2.1 Decision Trees

Decision trees are powerful non-linear classifiers, which utilize a tree structure to model the relationships among the features and the potential outcomes. A decision tree classifier uses a structure of branching decisions, which channel examples into a final predicted class value.

Decision trees is being popularly used in all kinds of data science problems. Hence, for every analyst, it's important to learn these algorithms and use them for modeling. This is a type of algorithm which is mostly used in classification problems. It works for both categorical and continuous input and output variables. (Louzada, Ara, & Fernandes, 2013), (A. Abdou & Pointon, 2011)



This machine-learning approach is used to classify data into classes and to represent the results in a flowchart, such as a tree structure. This model classifies data in a dataset by flowing through a query structure from the root until it reaches the leaf, which represents one class. The root represents the attribute that plays a main role in

classification, and the leaf represents the class. The decision tree model follows the steps outlined below in classifying data:

1. It puts all training examples to a root.
2. It divides training examples based on selected attributes.
3. It selects attributes by using some statistical measures.
4. Recursive partitioning continues until no training example remains, or until no attribute remains, or the remaining training examples belong to the same class.

Types of decision tree is based on the type of target variable we have. It can be of two types:

- **Classification Trees:** where the target variable is categorical, and the tree is used to identify the class within which a target variable would likely fall into. For example, the target variable has two value YES or NO.
- **Regression Trees:** where the target variable is continuous, and tree is used to predict its value.

Advantages of classification trees:

1. **Easy to Understand:** Decision tree output is very easy to understand even for people from non-analytical background. It does not require any statistical knowledge to read and interpret them. Its graphical representation is very intuitive, and users can easily relate their hypothesis.
2. **Useful in Data exploration:** Decision tree is one of the fastest ways to identify most significant variables and relation between two or more variables. With the help of decision trees, we can create new variables that have better power to predict. It can also be used in data exploration stage. For example, we are working on a problem where we have information available in hundreds of variables, there decision tree will help to identify most significant variable.
3. **Less data cleaning required:** It requires less data cleaning compared to some other modeling techniques. It is not influenced by outliers and missing values to a fair degree.
4. **Data type is not a constraint:** It can handle both numerical and categorical variables.



5. **Non-Parametric Method:** Decision tree is considered to be a non-parametric method. This means that decision trees have no assumptions about the space distribution and the classifier structure.

Disadvantages:

The modification of a single variable may change the whole tree if this variable is located near the top of the tree. This results in a lack of robustness. You can overcome it by resampling, in which one can construct the trees on many successive samples and can aggregate by a vote or a mean. But this will lead to losing the simplicity and readability of the model, which are the advantages of decision trees. For example, an individual item has all categories of a group A except the value of a variable that splits the tree. In this case, a variable is misclassified in another group as the tree has tested this variable. (James, Witten, Hastie, & Tibshirani, 2013)

8.2.2 Classification trees

A classification tree is very similar to a regression tree, except that it is classification used to predict a qualitative response rather than a quantitative one. For a regression tree, the predicted response for an observation is given by the mean response of the training observations that belong to the same terminal node. In contrast, for a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs. In interpreting the results of a classification tree, we are often interested not only in the class prediction corresponding to a terminal node region, but also in the class proportions among the training observations that fall into that region.

Classification Trees vs Regression Trees

Both the trees work almost similar to each other, let's look at the primary differences & similarity between classification and regression trees:

1. Regression trees are used when dependent variable is continuous. Classification trees are used when dependent variable is categorical.



2. In case of regression tree, the value obtained by terminal nodes in the training data is the mean response of observation falling in that region. Thus, if an unseen data observation falls in that region, we'll make its prediction with mean value.
3. In case of classification tree, the value (class) obtained by terminal node in the training data is the mode of observations falling in that region. Thus, if an unseen data observation falls in that region, we'll make its prediction with mode value.
4. Both the trees divide the predictor space (independent variables) into distinct and non-overlapping regions. For the sake of simplicity, you can think of these regions as high dimensional boxes or boxes.
5. Both the trees follow a top-down greedy approach known as recursive binary splitting. We call it as 'top-down' because it begins from the top of tree when all the observations are available in a single region and successively splits the predictor space into two new branches down the tree. It is known as 'greedy' because, the algorithm cares (looks for best variable available) about only the current split, and not about future splits which will lead to a better tree.
6. This splitting process is continued until a user defined stopping criterion is reached.
7. In both the cases, the splitting process results in fully grown trees until the stopping criteria is reached. But the fully-grown tree is likely to overfit data, leading to poor accuracy on unseen data. This bring 'pruning'. Pruning is one of the techniques used tackle overfitting.

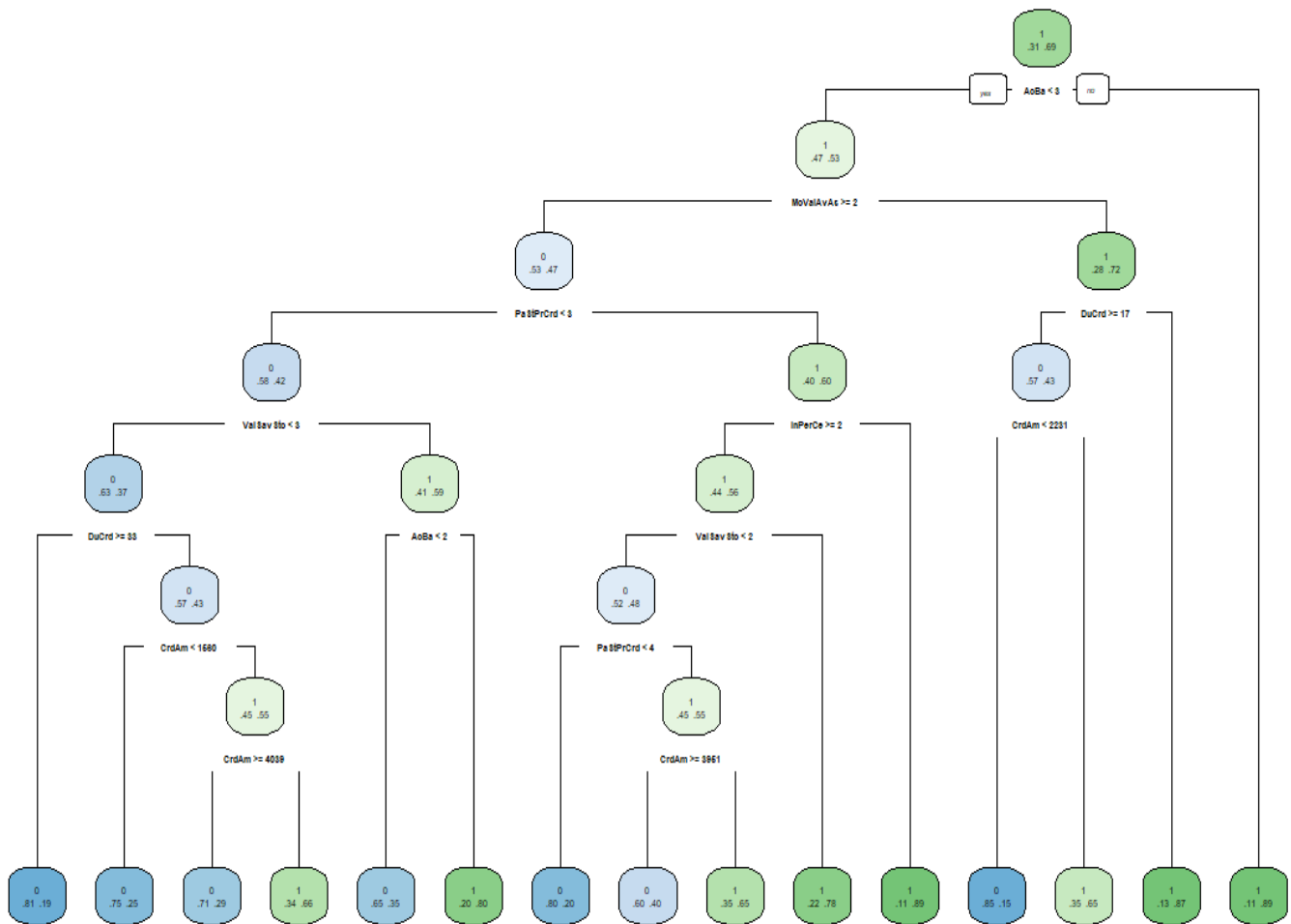
8.2.3 How the Decision trees work with our Dataset

Decision trees

First, we introduce the dataset which contains 16 variables and 100 observations. Data are splitted to train and test. We use library(rpart) for running the tree model (see Appendix).



Tree plot:



Generally, the tree has the root at the top and leaves at the bottom. The most important variable to the prediction model is AcBa which is at the top of the tree. Basically, it is the most significant variable in helping to classify the observation. We start from the root. If $AcBa < 3$ or if “yes” then we go to MoValAvAs. In this knot, the probability of the loan applicant to be “bad” is 0.47 and “good” is 0.53 and we follow the same procedure until the last tree leaf. If “no”, then the loan applicant is more likely to be “good” with probability 0.89 and 0.11 “bad”.

Now, we will use the tree model to calculate misclassification error as well as the correct prediction for **train data**. So, what we are doing here is that we are creating the following table:

	0	1
0	90	64
1	94	346

In this table the line indicates the reality or the actual classification. The column indicates the prediction from the tree model. So, the number in the diagonal (90, 346) show the correct classification and those in the other diagonal (94,64) show the misclassification.

So, **correct prediction** = 0.734 or 73.4% and **misclassification error** = 0.265 or 26.5%.

Then, we will use the tree model to calculate misclassification error as well as the correct prediction for **test data**. So, what we are doing here is that we are creating the following table:

testPred	0	1
0	55	54
1	61	236

Here, the **correct prediction**=0.7167 or 71.67% and **misclassification error** =0.2832 or 28.32%.

So, this prediction is 71.61% accurate.

8.3 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

Random Forest is a supervised learning algorithm. As we can see from its name, it creates a forest and makes it somehow random. The “forest” it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. In other words, random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. (James, Witten, Hastie, & Tibshirani, 2013)



Random Forest:

- Develop by aggregating trees.
- Can be used for classification or regression.
- Avoid overfitting.
- Can deal with large numbers of features.
- Helps with feature selection based on importance
- User friendly: only two parameters.

1) Trees- ntree

2) Variables randomly sampled as candidates at each split

8.3.1 How Random Forest works with our Dataset

First, we split the dataset to train and test and then will find the appropriate mtry through R to run the random forest model and find the less error rate. Actually, it is intended to find the less OOB of error rate.

For Random Forest we calculate OOB (out of bag error). That we do is for each bootstrap iteration and related tree, we calculate prediction error using data not in bootstrap sample. When we are doing classification the OOB is the Accuracy.

Since we find that it should be used mtry=3 or 3 number of variables available for splitting at each tree node, we will firstly run the model using the **train data**.

Random Forest	
Number of trees:	500
No. of variables tried at each split:	3
OOB estimate of error rate:	22%

Confusion matrix:			
	0	1	Class. Error
0	90	96	0.516129
1	38	383	0.902613

As we see the OOB estimate error rate is 22.08%. We have used 500 trees in the model with mtry equals 3. Classification errors for “0” or “bad” is 51.6% and for “1” or “good” is 9.02%, much better.



Let's look at the **training error**:

Confusion Matrix and Statistics			
	Reference		
Prediction	0	1	
0	185	0	
1	1	421	

Accuracy:	0.9984
95% CI:	(0.9909, 1)
No Information Rate:	0.6936
P-Value [Acc > NIR]:	<2e-16
Kappa:	0.9961
Mcnemar's Test P-Value:	1
Sensitivity:	0.9946
Specificity:	1
Pos Pred Value:	1
Neg Pred Value:	0.9976
Prevalence:	0.3064
Detection Rate:	0.3048
Detection Prevalence	0.3048
Balanced Accuracy	0.9973
Positive' Class:	0

For the train data Accuracy level is 99.84%.

However, the real test is going to be best on **test data**. So, let's get the test data confusion matrix:

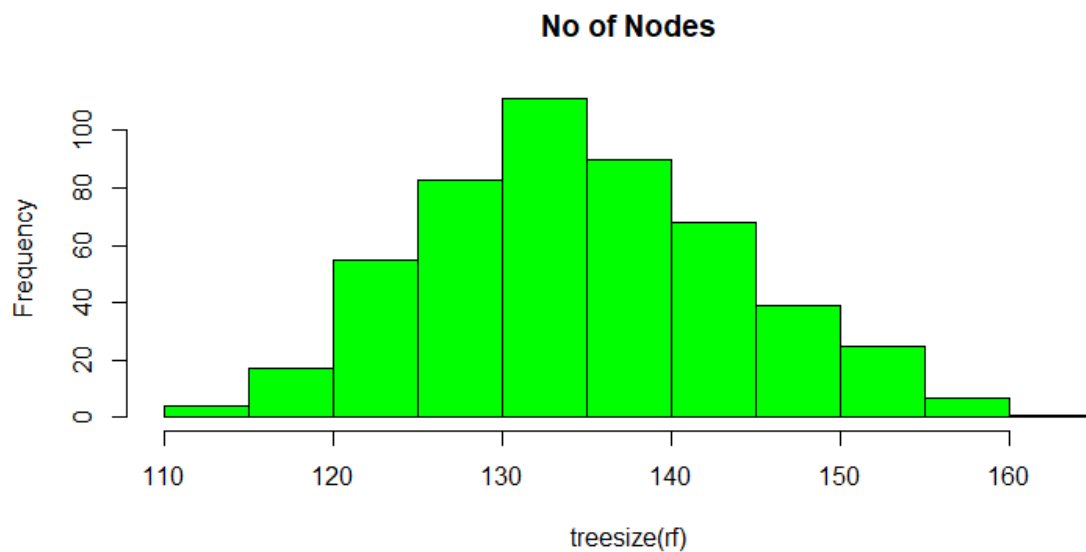
Confusion Matrix and Statistics			
	Reference		
Prediction	0	1	
0	47	25	
1	67	254	

Accuracy:	0.7659
95% CI:	(0.7209, 0.8069)
No Information Rate:	0.7099
P-Value [Acc > NIR]:	0.007522
Kappa:	0.3621
Mcnemar's Test P-Value:	1.92E-05
Sensitivity:	0.4123
Specificity:	0.9104
Pos Pred Value:	0.6528
Neg Pred Value:	0.7913
Prevalence:	0.2901
Detection Rate:	0.1196
Detection Prevalence	0.1832
Balanced Accuracy	0.6613
Positive' Class:	0

Hence, it gives Accuracy level 76.56% and as a result the prediction is 76.5% accurate.

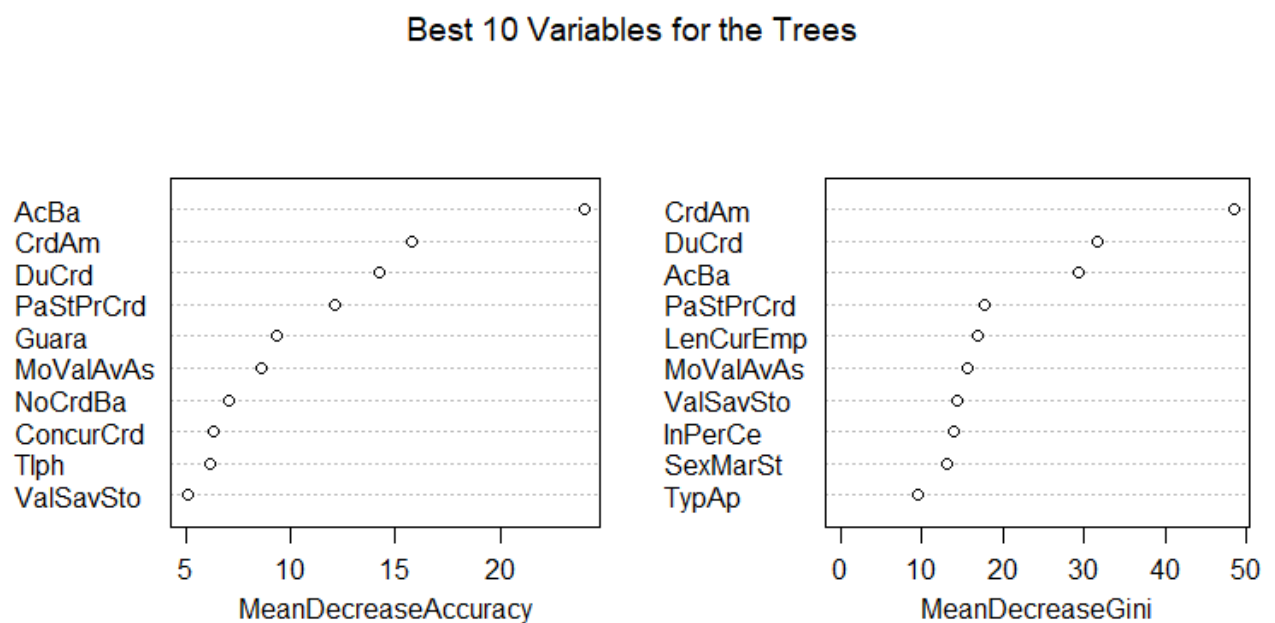


Numbers of Nodes:



The histogram shows the distribution of number of nodes in each of those 500 trees. The biggest Var is close to 115. There exist 115 trees which have 130-135 nodes. The smallest Var is at the left side and there are almost 5 trees with 110-115 nodes.

Which variables play an important role in the model:



The left graph indicates how worse the model performs without each variable. The first variable AcBa is on the top and CrdAm, DuCrd and PaStPrCrd have important contribution. The contribution of ValSavSto is the least and it equals 5.

The right graph measures how pure the nodes are at the end of the tree without each variable. (Nasa & Suman, 2012), (Yap, Ong, & Husain, Using data mining to improve assessment of credit worthiness via credit scoring models, 2011), (James, Witten, Hastie, & Tibshirani, 2013).



9. Comparison of Credit Scoring Methods

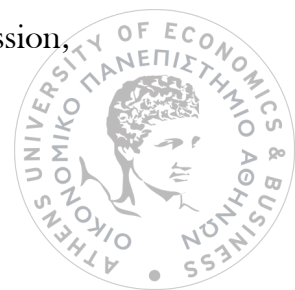
To compare the classification model, we put in the following table the basic indicators; that is, sensitivity and specificity. The sensitivity rate is the true positive rate (the percentage of defaulters predicted correctly as defaulters), while specificity is the true negative rate (percentage of the non-defaulters predicted correctly as non-defaulters). These values of indicators are based on the training and the validation subsets respectively. To compare predictive models, we need to look at the Type I error (a good credit customer being misclassified as bad credit customer) and Type II error (a bad credit customer being misclassified as a good credit customer) of the models. We also include Type I error and Type II error at the table below.

Table (Sensitivity, Specificity, Type I error and Type II error)

Model	Sample	Sensitivity	Specificity	Type I error	Type II error
Logistic Regression	Training	0.658	0.804	0.196	0.342
	Validation	0.67	0.796	0.204	0.33
Linear Discriminant Analysis	Training	0.63	0.8	0.2	0.37
	Validation	0.68	0.81	0.19	0.32
Quadratic Discriminant Analysis	Training	0.63	0.72	0.28	0.37
	Validation	0.6	0.85	0.15	0.4
K-Nearest Neighbor	Training	0.57	0.56	0.44	0.43
	Validation	0.02	0.98	0.03	0.98
Decision tree	Training	0.58	0.79	0.21	0.42
	Validation	0.5	0.79	0.21	0.5
Random Forest	Training	-	0.998	0.002	0
	Validation	0.65	0.79	0.21	0.34

Generally, Type II errors are higher than Type I errors, except from the Random Forest. It is found that the LDA model has the highest sensitivity and the lowest Type II error (a defaulter misclassified as non-defaulter).

To examine the results, we look at the **validation subset** of the models. It is obvious that the K-nearest neighbor is the worst model as it has the highest Type II error and the lowest sensitivity. Although Random forest has low Type II error 0.34 and Sensitivity 0.65, we see that there is a great difference between test and train data. Moreover, we could say that the decision tree is not an appropriate model, because the Type II error equals Sensitivity which equals 0.5. Hence, Logistic Regression, LDA and QLDA are the “best” models.



10. Conclusion

Statistical techniques consist credit decision-makers' tools which are used by the banks and some companies to assess if the customers or loan applicants are capable to repay their obligations. In other words, if the customers are creditworthy or not ("good" or "bad"). The institutions collect data related to the customers characteristics such as level of income, assets that the customer possesses job stability, profession, amount lent in relation to monthly income, case history of loans and payments, marital status and children as well as age. Furthermore, these performance evaluation criteria can also help them to choose the best model based on their aims and objectives. Classification methods, especially for loan applications, have become the most important forecasting techniques or tools which by using algorithms can provide results with significant accuracy. Nowadays, it is important to have a database with different types of characteristics in order to make predictions using statistical techniques. Of course, the possibility of finding data interests not only the managers in banking or finance, but also researchers in academic field. In reality, we cannot say for sure that there is a "best" model. It should be emphasized that there is no ideal credit scoring modelling procedure. It depends on the data structure, data quality and the objective of the classification.

Here, we used a German credit dataset and with the help of R programming we tried to examine six different statistical methods in order to make the best prediction. In our case, LDA and QLDA illustrates the most appropriate results.

We conclude to the choice of Discriminant Analysis as the appropriate technique, since it has the best results. We found that 22% of the applicants are predicted false and 78% of them correct, which is relatively high. That is to say, if we were credit officers, we would conclude that the model at hand, predicts approximately 8 out of 10 the true status of each loan candidate, or else, if had a "good" candidate model would identify him 4 out of 5 times.



APPENDIX

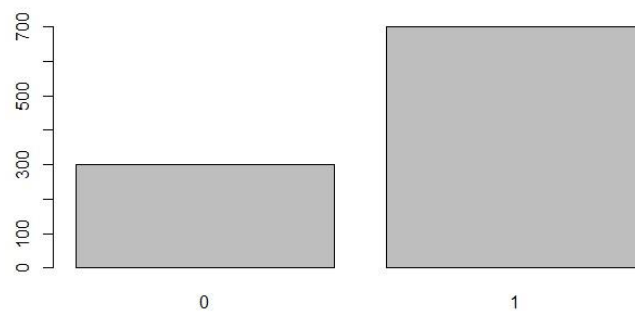
Page 16:

Names of Variables	
Creditability	Crdblt
Account Balance	AcBa
Duration of Credit (month)	DuCrd
Payment Status of Previous Credit	PaStPrCrd
Purpose	Purp
Credit Amount	CrdAm
Value Savings/Stocks	ValSavSto
Length of current employment	LenCurEmp
Instalment per cent	InPerCe
Sex & Marital Status	SexMarSt
Guarantors	Guara
Duration in Current address	DurCurAddr
Most valuable available asset	MoValAvAs
Age (years)	Age
Concurrent Credits	ConcurCrd
Type of apartment	TypAp
No of Credits at this Bank	NoCrdBa
Occupation	Occup
No of dependents	NoDpnd
Telephone	Tlph
Foreign Worker	FrgnWrkr

Dataset: <https://onlinecourses.science.psu.edu/stat857/node/222>

Page 16:

Creditability			
Not credit worthy	0	300	30%
Credit worthy	1	700	70%



Account Balance			
No running account	1	274	27.4%
No balance or debit	2	269	26.9%
0 <= ... < 200 DM	3	63	6.3%
... >= 200 DM or checking account for at least 1 year	4	394	39.4%

Payment Status of Previous Credit			
Hesitant payment of previous credits	0	40	4%
Problematic running account / credits running at other banks	1	49	4.9%
No previous credits / paid back all previous credits	2	530	53%
No problems with current credits at this bank	3	88	8.8%
Paid back previous credits at this bank	4	293	29.3%

Purpose			
Other	0	234	23.4%
New car	1	103	10.3%
Used car	2	181	18.1%
Items of furniture	3	280	28%
Radio / television	4	12	1.2%
Household appliances	5	22	2.2%
repair	6	50	5%
education	7	0	0%
vacation	8	9	0.9%
retraining	9	97	9.70%
business	10	12	1.20%

Value Saving/Stocks			
No available/no savings	1	603	60.3%
< 100,- DM	2	103	10.3%
100,-<=...<500,DM	3	63	6.3%
500,<=...<1000,DM	4	48	4.8%
>= 1000,- DM	5	183	18.3%

Length of Current Employment			
unemployed	1	62	6.2%
<= 1 year	2	172	17.2%
1<=...<4years	3	339	33.9%
4<=...<7years	4	174	17.4%
>= 7 years	5	253	25.3%

Instalment per cent of available income			
>= 35	1	136	13.6%
25<=...<35	2	231	23.1%
20<=...<25	3	157	15.7%
< 20	4	476	47.6%



Sex & Marital Status			
male:divorced/living apart	1	50	5%
male: single	2	310	31%
male:married/widowed	3	548	54.8%
female	4	92	9.2%

Guarantors			
none	1	907	90.7%
Co-Applicant	2	41	4.1%
Guarantor	3	52	5.2%

Duration in Current Address			
< 1 year	1	130	13%
1 <= ... < 4 years	2	308	30.8%
4 <= ... < 7 years	3	149	14.9%
>= 7 years	4	413	41.3%

Most Valuable Available Asset			
not available / no assets	1	282	28.2%
Car / Other	2	232	23.2%
Savings contract with a building society / Life insurance	3	332	33.2%
Ownership of house or land	4	154	15.4%

Concurrent Credits			
at other banks	1	139	13.9%
at department store or mail order house	2	47	4.7%
no further running credits	3	814	81.4%

Type of Apartment			
free apartment	1	179	17.9%
rented flat	2	714	71.4%
owner-occupied flat	3	107	10.7%

No of Credits at this Bank			
one	1	633	63.3%
two or three	2	333	33.3%
four or five	3	28	2.8%
six or more	4	6	0.6%



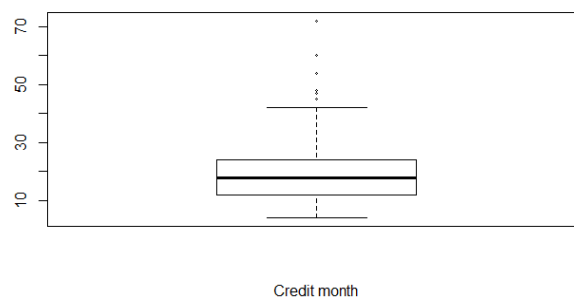
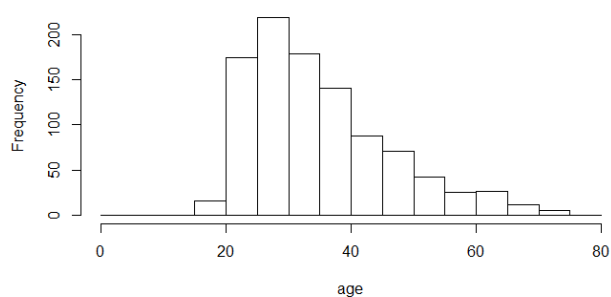
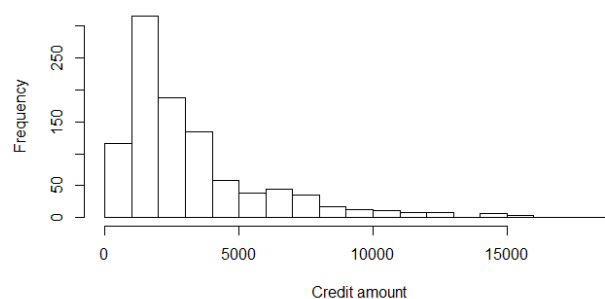
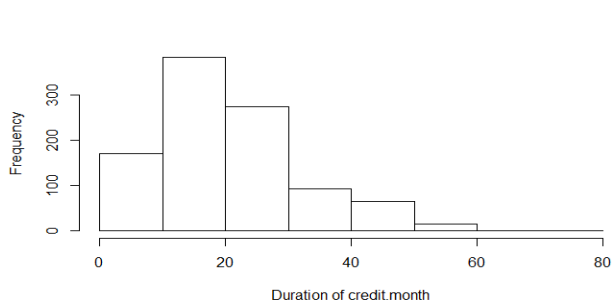
Occupation			
unemployed / unskilled with no permanent residence	1	22	2.2%
unskilled with permanent residence	2	200	20%
skilled worker / skilled employee / minor civil servant	3	630	63%
executive / self-employed / higher civil servant	4	148	14.8%

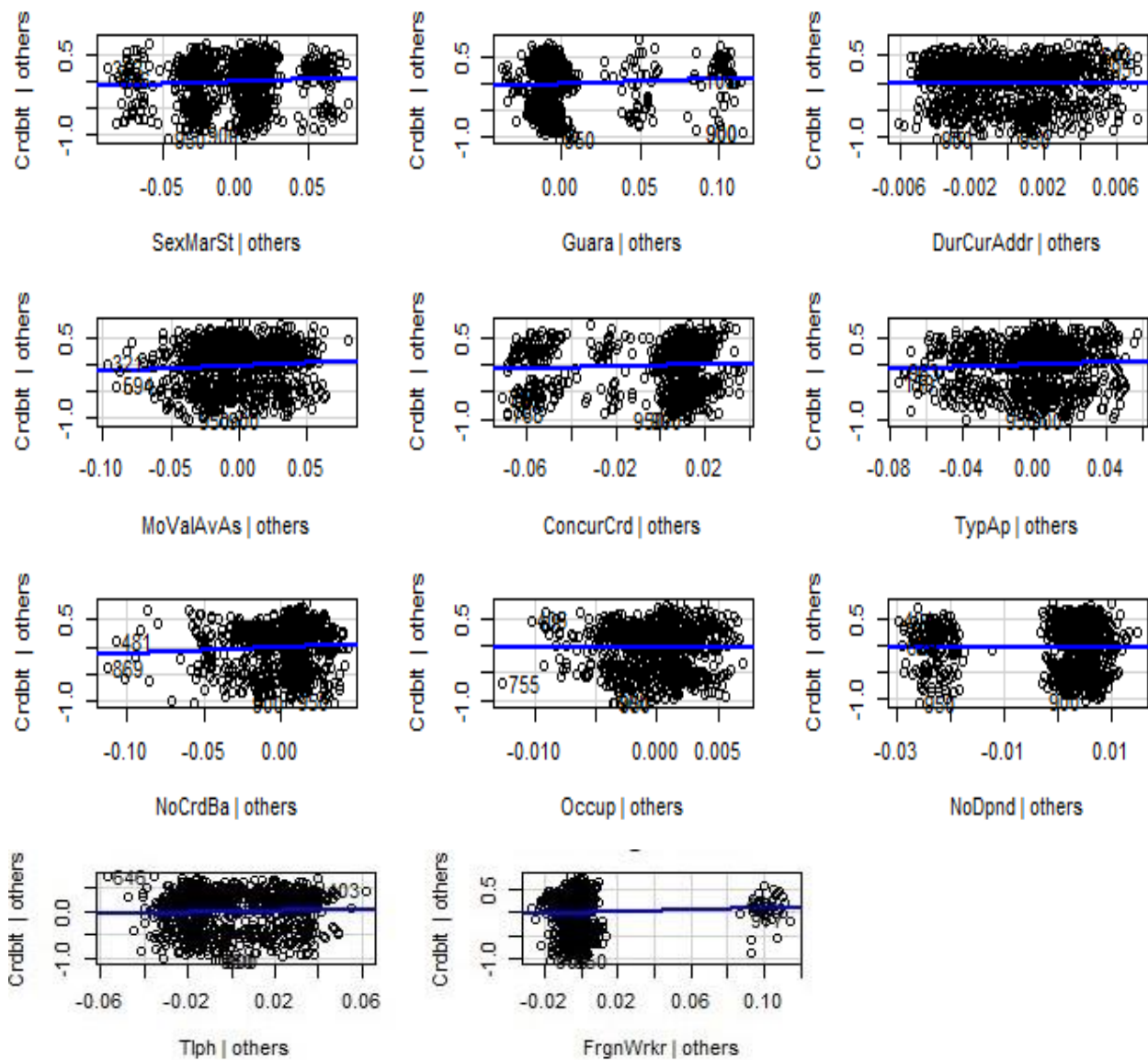
No of Dependents			
3 and more	1	845	84.5%
0 to 2	2	155	15.5%

Telephone			
No	1	596	59.6%
Yes	2	404	40.4%

Foreign Worker			
No	1	936	93.6%
Yes	2	37	3.7%

Page 16:





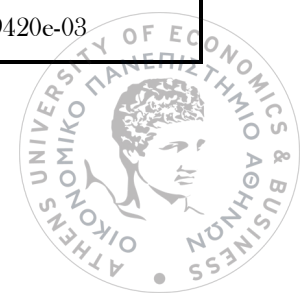
Residuals:				
Min	1Q	Median	3Q	Max
-1.0534	-0.3479	0.1186	0.2997	0.7878

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-5.147e-02	1.486e-01	-0.346	0.729203
DuCrd	-4.396e-03	1.484e-03	-2.962	0.003126 **
CrdAm	-1.534e-05	6.818e-06	-2.251	0.024634 *
Age	1.067e-03	1.289e-03	0.827	0.408229
AcBa	9.879e-02	1.086e-02	9.100	< 2e-16 ***
Purp	4.691e-03	4.828e-03	0.972	0.331492
PaStPrCrd	6.566e-02	1.387e-02	4.733	2.54e-06 ***
ValSavSto	3.424e-02	8.491e-03	4.032	5.95e-05 ***
LenCurEmp	2.482e-02	1.158e-02	2.144	0.032297 *
InPerCe	-4.707e-02	1.309e-02	-3.595	0.000341 ***
SexMarSt	4.386e-02	1.868e-02	2.348	0.019091 *
Guara	5.878e-02	2.776e-02	2.117	0.034487 *
DurCurAddr	-2.859e-03	1.261e-02	-0.227	0.820664
MoValAvAs	-3.250e-02	1.437e-02	-2.261	0.024003 *
ConcurCrd	3.614e-02	1.886e-02	1.916	0.055607 .
TypAp	4.988e-02	2.766e-02	1.803	0.071679 .
NoCrdBa	-4.226e-02	2.543e-02	-1.662	0.096907 .
Occup	4.991e-03	2.263e-02	0.221	0.825502
NoDpnd	-2.930e-02	3.675e-02	-0.797	0.425529
Tlph	5.102e-02	2.945e-02	1.733	0.083470 .
FrngWrkr	1.145e-01	7.064e-02	1.621	0.105444

Residual standard error:	0.4045 on 979 degrees of freedom
Multiple R-squared:	0.2374
Adjusted R-squared:	0.2218
F-statistic:	15.24 on 20 and 979 DF
p-value:	< 2.2e-16

1.The variance estimates for the coefficients

(Intercept)	DuCrd	CrdAm	Age	AcBa	Purp	PaStPrCrd
2.209470e-02	2.201742e-06	4.648671e-11	1.662344e-06	1.178642e-04	2.331275e-05	1.925044e-04
ValSavSto	LenCurEmp	InPerCe	SexMarSt	Guara	DurCurAddr	MoValAvAs
7.210277e-05	1.340419e-04	1.714281e-04	3.490579e-04	7.707512e-04	1.589403e-04	2.066286e-04
ConcurCrd	TypAp	NoCrdBa	Occup	NoDpnd	Tlph	FrngWrkr
3.556093e-04	7.651458e-04	6.467920e-04	5.121195e-04	1.350638e-03	8.670188e-04	4.989420e-03



2. Take the squared root/ SEs estimate from the standard formula

(Intercept)	DuCrđ	CrđAm	Age	AcBa	Purp	PaStPrCrđ
1.486429e-01	1.483827e-03	6.818116e-06	1.289319e-03	1.085653e-02	4.828328e-03	1.387460e-02
ValSavSto	LenCurEmp	InPerCe	SexMarSt	Guara	DurCurAddr	MoValAvAs
8.491335e-03	1.157765e-02	1.309305e-02	1.868309e-02	2.776241e-02	1.260715e-02	1.437458e-02
ConcurCrđ	TypAp	NoCrđBa	Occup	NoDpnd	Tlph	FrđnWrkr
1.885761e-02	2.766127e-02	2.543211e-02	2.263006e-02	3.675103e-02	2.944518e-02	7.063583e-02

3. To get SEs, we take the square root of the diagonal of a variance-covariance matrix for the coefficients, like we did above (heteroskedasticity-consistent formula).

(Intercept)	DuCrđ	CrđAm	Age	AcBa	Purp	PaStPrCrđ
1.537522e-01	1.660503e-03	8.076843e-06	1.345379e-03	1.107567e-02	5.120733e-03	1.467496e-02
ValSavSto	LenCurEmp	InPerCe	SexMarSt	Guara	DurCurAddr	MoValAvAs
8.313477e-03	1.220727e-02	1.375592e-02	1.973157e-02	2.909345e-02	1.290001e-02	1.431056e-02
ConcurCrđ	TypAp	NoCrđBa	Occup	NoDpnd	Tlph	FrđnWrkr
2.066148e-02	2.917151e-02	2.706832e-02	2.377242e-02	3.736566e-02	2.876303e-02	5.698500e-02

Page 54:

train.crd		
knn..55	0	1
	0	107 81
	1	181 231

The correct prediction of the train data, when k=55, is 56.33%

Page 58:

n= 594

node), split, n, loss, yval, (yprob)
* denotes terminal node

```

1) root 594 184 1 (0.3097643 0.6902357)
 2) AcBa< 2.5 331 154 1 (0.4652568 0.5347432)
   4) MoValAvAs>=1.5 246 116 0 (0.5284553 0.4715447)
     8) PaStPrCrđ< 2.5 171 71 0 (0.5847953 0.4152047)
       16) ValSavSto< 2.5 134 49 0 (0.6343284 0.3656716)
         32) DuCrđ>=33 36 7 0 (0.8055556 0.1944444) *
         33) DuCrđ< 33 98 42 0 (0.5714286 0.4285714)
           66) CrđAm< 1560 40 10 0 (0.7500000 0.2500000) *
           67) CrđAm>=1560 58 26 1 (0.4482759 0.5517241)
             134) CrđAm>=4038.5 17 5 0 (0.7058824 0.2941176) *
             135) CrđAm< 4038.5 41 14 1 (0.3414634 0.6585366) *
       17) ValSavSto>=2.5 37 15 1 (0.4054054 0.5945946)
         34) AcBa< 1.5 17 6 0 (0.6470588 0.3529412) *
         35) AcBa>=1.5 20 4 1 (0.2000000 0.8000000) *
       9) PaStPrCrđ>=2.5 75 30 1 (0.4000000 0.6000000)
       18) InPerCe>=1.5 66 29 1 (0.4393939 0.5606061)
         36) ValSavSto< 1.5 48 23 0 (0.5208333 0.4791667)
           72) PaStPrCrđ< 3.5 10 2 0 (0.8000000 0.2000000) *
           73) PaStPrCrđ>=3.5 38 17 1 (0.4473684 0.5526316)
             146) CrđAm>=3950.5 15 6 0 (0.6000000 0.4000000) *
             147) CrđAm< 3950.5 23 8 1 (0.3478261 0.6521739) *
           37) ValSavSto>=1.5 18 4 1 (0.2222222 0.7777778) *
         19) InPerCe< 1.5 9 1 1 (0.1111111 0.8888889) *
       5) MoValAvAs< 1.5 85 24 1 (0.2823529 0.7176471)
         10) DuCrđ>=16.5 30 13 0 (0.5666667 0.4333333)
           20) CrđAm< 2230.5 13 2 0 (0.8461538 0.1538462) *
           21) CrđAm>=2230.5 17 6 1 (0.3529412 0.6470588) *
           11) DuCrđ< 16.5 55 7 1 (0.1272727 0.8727273) *
       3) AcBa>=2.5 263 30 1 (0.1140684 0.8859316) *

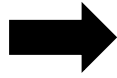
```

Page 49:



Confusion Matrix

	0	1
0	α	β
1	γ	δ



Sensitivity = $\alpha / (\alpha + \beta)$

Specificity = $\delta / (\gamma + \delta)$

Type I error = $\gamma / (\gamma + \delta)$

Type II error = $\beta / (\alpha + \beta)$

Bibliography

- A. Abdou, H., & Pointon, J. (2011). Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature. *Intelligent Systems in Accounting, Finance & Management*.
- Ching-Ti Liu, P. A., Jacqueline Milton, P. C., & Avery McIntosh, d. c. (2016). *Correlation and Regression with R*. Retrieved from <https://www.statmethods.net/stats/rdiagnostics.html>
- Chong, I.-G., & Jun, C.-H. (2004). Performance of some variable selection methods when multicollinearity. *Elsevier*.
- Fernandes., F. L. (2016). Classification methods applied to credit scoring: A systematic review and overall comparison.
- GIETZEN, T. (2017, June). Credit Scoring vs. Expert Judgment, A Randomized Controlled Trial. St. Gallen.
- Greenspan, E. (2017). Credit Scoring and its Applications. Philadelphia, USA: David Marshall.
- Guyon, I., & Elisseeff, A. (2003). Overfitting in Making Comparisons. *Journal of Machine Learning Research*.
- Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society*.
- Hastie, T., Tibshirani, R., & Friedman, J. (May, 2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Stanford, California: Springer.
- <https://www.investopedia.com/terms/a/adjusted-mean.asp>. (n.d.). Retrieved from Investopedia.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Application in R*. New York: Springer.
- Kern, A. M. (2017). CREDIT SCORE ANALYSIS.
- Louzada, F., Ara, A., & Fernandes, G. B. (2013). Classification methods applied to credit scoring: A systematic review and overall comparison. *Elsevier*.
- Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Elsevier*.
- Murtaugh, P. A. (2009). Performance of several variable-selection methods applied to real ecological data. 1061-1068. Department of Statistics, Oregon State University, Corvallis, OR 97331, USA.
- Nasa, C., & Suman. (2012). Evaluation of Different Classification Techniques for WEB Data. *International Journal of Computer Applications*.
- Quick-R*. (n.d.). Retrieved from <https://www.statmethods.net/stats/rdiagnostics.html>
- Schreiner, M. (2000). Credit Scoring for Microfinance: Can It Work? *Journal of Microfinance*.



Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers.

International Journal of Forecasting.

University of Virginia Library. (n.d.). Retrieved from Research Data Services+Sciences:

<https://data.library.virginia.edu/diagnostic-plots/>

VOJTEK, M., & KOČENDA*, E. (2005). Credit Scoring Methods. Charles University and Academy of Sciences, Prague.

Wikipedia. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Feature_selection

Wooldridge, J. M. (2006). *Introductory Econometrics. A Modern Approach*. United States of America, Michigan State University.

Yap, B. W., Ong, S. H., & Husain, N. H. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Elsevier*.

