MEF UNIVERSITY

CREDIT RISK MODELS USING MACHINE LEARNING MODELS

Capstone Project

Özkan Akman

İSTANBUL, 2018

MEF UNIVERSITY

CREDIT RISK MODELS USING MACHINE LEARNING MODELS

Capstone Project

Özkan Akman

Advisor: Dr. Tuna Çakar

İSTANBUL, 2018

EXECUTIVE SUMMARY

CREDIT RISK MODELS USING MACHINE LEARNING MODELS

Özkan Akman

Advisor: Dr. Tuna Çakar

JANUARY, 2018, 29 pages

Credit scoring is an important subject in financial institutions, mainly in banks. I want to examine some machine learning techniques to find out a model that performs good in predicting or classifying the loaner person a good credit or a bad one by evaluating his/her demographic features as marital status, wealth, job seniority, monthly income and expenses. The purpose of the study is to build a classifier to predict the granting of retail credits with the help of a database of recent concessions of credits done by analysts.

First, I start with a general data analysis and understand the statistical figures in each variable such as max, min, mean and figure out if there any outlier or missing values I dataset. In other words: Explanatory Data Analysis (EDA) Feature Selection Methods, Missing Values Imputation are taken as initial steps in the project. R has already MICE library to handle missing value imputations. The data cleaning or data preprocessing is time-consuming but the most crucial step at the same time. However, if I want to come up with a working Machine Learning Model that performs for new dataset, it is quite important to make the data cleaning well done.

I have added two new variables in dataset: Loan-To-Value Ratio and Saving Capacity. Loan-to-Value has been calculated taking into account value of the total wealth of the person and the mortgage left to pay with the price of good that she/he would like to buy to express the ratio of money solicitated of an asset purchased. The higher the Loan-To-Value Ratio, the riskier the loan for a bank. The Saving Capacity is to express the power of borrower in saving money which is calculated through his/her income, expense, mortgage, and the amount of money solicitated. The lower the saving capacity, the riskier the loan is for a bank.

I proposed four different machine learning models in my project. Classification Tree, Kernel Support Vector Machine, Linear Discrimant Analysis, Logistic Regression Models. All these machine learning model algorithms try to estimate an applicant's credit situation so as to decide whether a bank should provide a loan to the applicant or not.

In order to compare different model performances, I prefer to use 10-fold cross validation method, each time the error is computed and stored in a vector for all four models with the average error. I have found that the least mean error is obtained when the decision tree classifier is used. I considered error as the measure for model selection. Therefore, I will choose "Classification tree" as the ML classifier model. The decision tree approach has

become a popular technique for developing credit scoring models because the resulting decision trees are easily interpretable and visualized.

Key Words: Credit Ranking, Credit Scoring Models, Machine Learning, Support Vector Machine, Decision Tree Model, Linear Discrimant Analysis, Loan-to-Value Ration, Saving Capacity, Logistic Regression Model

ÖZET

MAKİNE ÖĞRENMESİ UYGULAMALARI İLE KREDİ RİSK MODELLEME

Özkan Akman

Tez Danışmanı: Dr. Tuna Çakar

OCAK, 2018, 29 sayfa

Kredi skorlaması, başta bankalar olmak üzere finansal kuruluşlarda önemli bir konudur. Medeni durum, servet, iş kıdemi, aylık gelir-gider gibi demografik özelliklerini değerlendirerek, kredi sonucunu pozitif veya negatif sınıflandırmada iyi performans gösteren bir model bulmak için bazı makine öğrenme tekniklerini incelemek istedim. Projemin amacı, analistler tarafından yakın zamanda verilen kredi bilgilerinin durumunun yer aldığı bir veritabanı yardımıyla talep edilen perakende kredilerin sonucunu öngören bir sınıflandırıcı oluşturmaktır.

Öncelikle, genel bir veri analizi ile başlayıp, maks, min, ortalama gibi istatistiksel değerleri bulup, verisetide herhangi bir aykırı değer veya eksik değerlerin tespiti yapıldı. Başka bir deyişle: Açıklayıcı Veri Analizi, Özellik Seçme Yöntemleri, Eksik Değerleri Tamamlama projedeki atılan ilk adımlardı. R programı, mevcut veritabanındaki eksik değer muafiyetlerini işlemek için zaten "MICE" kütüphanesine sahiptir. Veri temizleme veya veri ön işleme, zaman alıcı ancak aynı zamanda en önemli adımdır. Bununla birlikte, eğer yeni bir veri kümesi için doğru sonularını veren bir Makine Öğrenme Modeli hazırlanmak istenirse, verilerin temizlenmesini sağlamak oldukça önemlidir.

Ayrıca, veri kümesinde iki yeni değişken ekledim: Kredi Değeri Oranı ve Tasarruf Kapasitesi. Kredi Değeri, kişinin toplam servetinin ve satın alınan bir varlığa ait talep edilen kredi tutarının oranını ifade etmek için satın almak istediği malın bedeli ile ödemek zorunda kalan ipotek hesaba katılarak hesaplanmıştır. Kredi Değeri Oranı ne kadar yüksekse, bir banka kredisi o kadar risklidir. Tasarruf Kapasitesi, gelir, gider, ipotek ve talep edilen para tutarıyla hesaplanan borçlunun ödeme gücünü ifade etmektir. Tasarruf kapasitesi ne kadar düşükse de kredi bir banka için o denli risklidir.

Projemde verisetini dört farklı makine öğrenme modeliyle çalıştırdım. Sınıflandırma Ağacı, Çekirdek Destek Vektör Makinesi, Doğrusal Ayırıcı Analiz, Lojistik Regresyon Modelleri. Tüm bu makine öğrenme modeli algoritmaları bir bankaya gelen başvuru sahibinin kredi talebinin sonucunu tahmin etmeye çalışır ve böylece bir bankanın başvurana kredi sağlayıp sağlamayacağına karar verir.

Farklı model performanslarını karşılaştırmak için, hata her hesaplandığında ortalama hata ile dört modelin tümü için bir vektöre depolandığında "10-kat çapraz doğrulama yöntemi" kullanmayı tercih ettim. Karar ağacı sınıflandırıcısı modeli kullanıldığında en küçük ortalama hata elde edildiğini gördüm. Hatayı, seçeceğim makine öğrenmesi model performansı için bir kriter olarak aldım. Bu nedenle makine öğrenmesi sınıflandırıcı modeli olarak "Karar Ağacı Modeli"ni seçtim. Karar Ağacı Modeli yaklaşımı, kredi puanlama

modelleri geliştirmede popüler bir teknik haline gelmiştir çünkü ortaya çıkan karar ağaçları kolayca yorumlanabilir ve görselleştirilebilir.

Anahtar Kelimeler: Kredi Sıralaması, Kredi Puanlama Modelleri, Makine Öğrenmesi, Destek Vektör Makinesi, Karar Ağacı Modeli, Lineer Ayrımcılık Analizi, Kredi-Değer Rasyon, Tasarruf Kapasitesi, Lojistik Regresyon Modeli

TABLE OF CONTENTS

1. INTRODUCTION

Credit evaluation is one of the most crucial processes in banks' credit management decisions. This process includes collecting, analyzing and classifying different credit elements and variables to assess the credit decisions. The quality of bank loans is the key determinant of competition, survival and profitability. One of the most important kits, to classify a bank's customers, as a part of the credit evaluation process to reduce the current and the expected risk of a customer being bad credit, is credit scoring. Hand & Jacka, (1998, p. 106) stated that 'the process (by financial institutions) of modelling creditworthiness is referred to as credit scoring'. It is also useful to provide further definitions of credit scoring

In "Credit Scoring", the process of evaluating the risk a customer poses of defaulting on a financial obligation (Hand & Henley, 1997) is to assign customers to one of two groups: "Good" and "Bad". Here I will concentrate on "Application Scoring" that's used at the time an application for credit is made and estimates an applicant's likelihood of default in a given time period. Credit scores which is a numerical value representing the creditworthiness of a person. This is evaluated by the analysis of person's portfolio which may include his/her revenues, information about mortgages, the property owned by person. The data used for model fitting for this task generally consists of financial and demographic information about a sample of previous applicants along with their good/bad status at some later date.

To define credit scoring, the term should be broken down into two components, credit and scoring. Firstly, simply the word "credit" means "buy now, pay later". It is derived from the Latin word "credo", which means "I believe" or "I trust in". Secondly, the word „scoring" refers to "the use of a numerical tool to rank order cases according to some real or perceived quality in order to discriminate between them, and ensure objective and consistent decisions". Therefore, scores might be presented as "numbers" to represent a single quality, or "grades" which may be presented as "letters" or "labels" to represent one or more qualities. Consequently, credit scoring can be simply defined as "the use of statistical models to transform relevant data into numerical measures that guide credit decisions. It is the industrialization of trust; a logical future development of the subjective credit ratings first provided by nineteenth century credit bureaux, that has been driven by a need for objective, fast and consistent decisions, and made possible by advances in technology. Credit scoring is the use of statistical models to determine the likelihood that a prospective borrower will

default on a loan. Credit scoring models are widely used to evaluate business, real estate, and consumer loans or the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit. These techniques decide who will get credit, how much credit they should get, and what operational strategies will enhance the profitability of the borrowers to the Lenders.

1.1. Data Description

The credit data set consist of 4.455 consumers' credit from a bank. The response variable of interest is 'status', which is given in categorical form (y = 1 for positive, y = 2 for negative). In addition, 13 variables are included in dataset as they are assumed to influence creditability. Total number of variables: 14

- Y ("Opinion"-response variable): a factor with levels '1- positive', '2-negative'
- Number.of.employment.years: a continous variable showing the years the person worked
- House.type: a factor variable for the type of house where the person is living; 1-rented, 2-ownerwithdeed, 3-private, 4-ignore, 5-parents, 6-others
- term.in.months: a continous variable showing the duration of loan in months
- Age: a continous variable indicating the age of person
- Marital.status: a factor variable for the marital status of the person; 1-single, 2-married, 3-widower, 4-separated, 5-divorced
- Registers :a factor variable indicating whether the person registers a house or not; 1- no, 2-yes
- Type.of.employment: a factor variable indicating the type of employment the person have; 1- Permanent, 2-Temporaray, 3-Self, 4-Others
- Expenses: a continous variable indicating the expenses of the person
- Income: a continous variable indicating the total income of the person.
- Total.wealth: a continous variable indicating the total wealth a person have
- Mortgage.left: a continous variable indicating the mortgage left to pay by the person
- Amount.of.money.solicitated: a continous variable indicating the total wealth of the person taking into account the mortgage left to pay
- Price.of.good.to.buy: a continous variable indicating the price of good the person wants to buy.

1.2. Project Definition

The purpose of the study is to build a classifier to predict the granting of retail credits with the help of a database of recent concessions of credits done by analysts. The goal of the classifier is to train our data and find out machine learning models that will perform in acceptable levels for the decision of granting loans to the customer to buy an asset in the future. In this Project, a number of specific challenges were encountered during the EDA(Exploratory Data Analysis) and modeling stages in the above framework.

However, the historical loan repayment performance of the credit applicant has not been provided in the dataset, so I can not measure behavioural scoring that allows lenders to regularly monitor customers and help coordinate customer-level decision making. To be profitable financial institution must accurately predict customers' likelihood of default over different time horizons (1 month, 3 months, 6 months, etc.). Banks needs to continue monitoring on changes in the debtors characteristics such as, past and current delinquency levels, time on books, amounts delinquent, whether the account exceeds the limit and by how much, among others.

1.3. Methodology

The credit scoring problem in this project is building a classifier to predict whether a customer should be granted for credit or not. The methodology to do such task is divided into following steps:

- Exploratory Data Analysis (EDA)
- Feature Extraction
- Building a classifier
- Cross validation for selecting best model
- Applying ML models on Test Data

I have used in my project, the following R libraries: FactoMineR, MICE, kernlab, MASS, class, tree, e1071 and rpart.

- FactoMineR: Multivariate Exploratory Data Analysis and Data Mining
- mice: Multivariate Imputation by Chained Equations
- kernlab: Kernel-Based Machine Learning Lab
- MASS: Support Functions and Datasets for Modern Applied Statistics
- class: Functions for Classification

- tree: Classification and Regression Trees
- e1071: Misc Functions of the Statistics, Probability Theory
- rpart: Recursive Partitioning and Regression Trees

The data is divided into two parts as: train data (2/3) and test data(1/3) Both train and test dataset were cleaned and preprocessed first with imputational actions for the missing values and outlier detection before implementing any models .

2. EXPLORATORY DATA ANALYSIS (EDA)

When I explore the summary of the dataset, I can see that there are some missing or incorrect values in the original dataset.

To handle the missing values, I prefer to replace them as NA. The numerical variables which appear as "99999999" were stated in the problem statement that the missing values are denoted by "99999999". The categorical variables which appear as "0" were assumed to be a missing value due to some mistake in the data.

Then, I factor the variables which should be the categorical variable and label the modalities: House.type, Marital.status, Registers, Registers and Opinion

The missing values are filled using the process of multiple imputations via using MICE library in R. MICE stands for Multivariate Imputation via Chained Equations. This package is well described in van Buuren & Groothuis-Oudshoorn (2011). Under the assumption that missing data are Missing at Random (MAR), the missing values are predicted by regression on observed values. Continues missing values are by default predicted by linear regression. The mice package in R, helps us imputing missing values with plausible data values. These plausible values are drawn from a distribution specifically designed for each missing data point. Each variable has its own imputation model. Built-in imputation models are provided for continuous data (predictive mean matching, normal), binary data (logistic regression), unordered categorical data (polytomous logistic regression) and ordered categorical data (proportional odds). MICE can also impute continuous two-level data (normal model, pan, second-level variables). Passive imputation can be used to maintain consistency between variables. Various diagnostic plots are available to inspect the quality of the imputations.

2.1. Chi-Square Test

To evaluate the link between each category of variables, a chi-square test is performed (alpha=0.05). The more significant the test is, the more the considered category and categorical variable are linked. In this dataset, chi-square results indicating that p.value of all categorical variables are less than 0.05 (at 95% confident). Thus the categorical variables here are significant. In order to see the link between one category of opinion (Positive and Negative) and another category of another categorical variable of the data set, the function compares two proportions:

- The proportion of individuals who possess the 2. category among those who possess 1.
- The global percentage of individuals who possess 2. category

From the result it can be seen that Type of employment = Permanent is over represented (as v-test>0) while Type.of.employment=Temporary is under represented (as v-test<0) among individual who has positive response for credit approval.

For each category of "opinion" and each continuous variable, I use quantitative variables to see the global description of the variable by the quantitative variables with the square correlation coefficient and the p-value of the test in a one-way ANOVA.

3. FEATURE EXTRACTION

Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction. Here, some variables are converted to categorical variables via decision trees and for each explanatory variable, a classification decision tree is constructed where predicted values are plotted against the explanatory variable.

Decision trees (TREES). Classification and Regression Trees (Breiman et al., 1984) is a classification method which uses historical data to construct so-called decision rules organized into tree-like architectures. In general, the purpose of this method is to determine a set of if-then logical conditions that permit prediction or classification of cases. There are three usual tree's algorithms: chi-square automatic interaction detector (CHAID),

classification and regression tree (CART) and C5, which differ by the criterion of tree construction, CART uses gini as the splitting criterion, C5 uses entropy, while CHAID uses the chi-square test to exhibit a rule based model implementation in a stock selection. Bijak and Thomas (2012) used CHAID and CART to verify the segmentation value in the performance capability. He proposes a combination of a Bayesian behavior scoring model and a CART-based credit scoring model. Other possible and particular methods of decision trees are C4.5 decision trees algorithm and J4.8 decision trees algorithm. Here I prefer to use CART, gini as the splitting criterion.

This feature extraction will help in reducing the number of features so that the model doesn't overfit the data leading to bad results. Overfitting is the production of an analysis that corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably

To better evaluate the data: two new variables are computed and added to the dataset: Loan-To-Value Ratio and Saving Capacity.

Loan-To-Value is calculated taking into account value of the total wealth of the person and the mortgage left to pay with the price of good that she/he would like to buy to express the ratio of money solicitated of an asset purchased. The higher the Loan-To-Value Ratio, the riskier the loan for a bank.

Loan-To-Value = Amount.of.money.solicitated / price.of.good.to.buy * 100

Saving Capacity is to express the power of borrower in saving money which is calculated through his/her income, expense, mortgage, and the amount of money solicitated. The lower the saving capacity, the riskier the loan is for a bank.

SavingCapacity=[Income- Expense- (mortgage. left/100)]/[Amount. of.money. solicitaded/ Terms.in.monhts]*100

In summary, we found that Age, Income, Amount.Of.Money.Solicitated, Mortgage, Price.Of.Good.To.Buy can be taken as linear, however Total.Wealth, Expenses, Number.Of.Employment.Years, Term.In.Months, ltv.Ratio And Saving.Capacity need to be transformed to Categorical or Binary as they have discrete values.

Table 1: Feature Extraction Table (Tranformation to Categorical Variables)

Variable Name	Description
Number.of.employment.years	[Number.of.employment.years<1] <- 1 [Number.of.employment.years>=1 & [Number.of.employment.years< 5] <- 2 [Number.of.employment.years>=5] <- 3
term.in.months	[term.in.months==12] <- 1 [term.in.months==24] <- 2 [term.in.months==36] <- 3 [term.in.months==48] <- 4 [term.in.months==60] <- 5
Expenses	[Expenses< 44.5] <- 1 [Expenses>=44.5 & Expenses< 85.5] <- 2 [Expenses>= 85.5 & Expenses < 145] <- 3 [Expenses>= 145] <- 4
total.wealth	total.wealth [p[,1]>0.4] <- 1 total.wealth [p[,1]<=0.4] <- 0
ltv..ratio	[ltv.ratio< 70] <- 1 [ltv.ratio>=70 & ltv.ratio< 98] <- 2 [ltv.ratio>=99] <- 3
Saving.capacity	[saving.capacity< 1] <- 1 [saving.capacity>=1] <- 2

4. IMPLEMENTING DIFFERENT MACHINE LEARNING MODELS

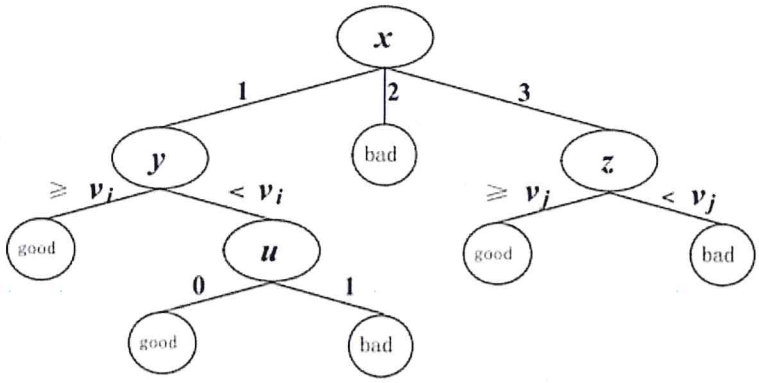
4.1. Classification Tree or Decision Tree Model:

A decision tree is a mapping from observations about an item to conclusion about its target value as a predictive model in data mining and machine learning. Generally, for such tree models, other descriptive names are used as classification tree (discrete target) or regression tree (continuous target). In these tree structures, the leaf nodes represent classifications, the inner nodes represent the current predictive attributes and branches represent conjunctions of attributions that lead to the final classifications. The popular decision trees algorithms include ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993) which is an extension of ID3 algorithm and CART. The C4.5 will be simply introduced as follows.

C4.5 builds decision trees from a set of training data with every sample of that classified, using the concept of information entropy. The training data is a set S = (s1, s2, ..., sn) and each sample si = (x1, x2, ..., xm, ci) is a vector, where x1, x2, ..., xm represents predictive attributes or features of the sample and ci represents the class of the sample si as a target attribute. At each inner node of the tree, C4.5 chooses one attribute that most effectively splits its set of samples into subsets. Its criterion is the normalized

information gain that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. Then the C4.5 algorithm is recursively executed for the smaller subsets. The C4.5 algorithm then recurses on the smaller subsets. When the tree is built, it would have some base cases as follows (Quinlan, 1993). 1. All the samples in the subsets belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class. 2. None of the attributes provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class. 3. Instance of previously unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value. The Fig. 1 is an illustration of the structure of decision tree built by some credit database with the C4.5, where x, y, z, u in inner nodes of the tree are predictive attributes and ''good" and ''bad" are the classifications of target attribute in the credit database.

Figure 1: A Structure of Decision Tree



I use here decision tree which is implemented using "rpart" function in R package. The maximum depth of the tree is taken to be 4 and cp to be 0.001

4.2. Kernel Support Vector Machine Model

The support vector machines (SVM) approach was first proposed by Cortes and Vapnik(1995). The main idea of SVM is to minimize the upper bound of the generalization error. SVM usually maps the input variables into a high-dimensional feature space through

some nonlinear mapping. In that space, an optimal separating hyper plane, which is one that separates the data with the maximal margin, is constructed by solving a constrained quadratic optimization problem.

Here, SVM is implemented in "ksvm" function of the R package. I take cost parameter as 5. My model has given a training error value of 0.185841

4.3. Linear Discriminant Analysis (LDA) Model

Linear Discriminant analysis (LDA) is introduced by Fisher (1986), based on the construction of one or more linear functions involving the explanatory variables. Consequently, the general model is given by $Z = \alpha + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$, where Z represents the discrimination score, α the intercept, βi represents the coeffi- cient responsible for the linear contribution of the i th explanatory variable Xi , where
$i = 1, 2, \ldots, p$.

This technique has the following assumptions:

(1) the covariance matrices of each classification subset are equal.

(2) Each classification group follows a multivariate normal distribution.

The purpose of discriminant analysis is to classify objects into one of two or more groups based on a set of features that describe the objects. A basic principal is to maximize the difference between two groups, while the differences among particular members of the same group are minimized. Within credit risk models, one group consists of good borrowers (non-defaulted – group A), while the other includes bad ones (already defaulted – group B). The differences are measured by means of the discriminant variable – score z. For a given borrower i, we calculate the score as follows:

$$Z_i = \sum_{j=1}^{n} \gamma_j x_{j,i} ,$$

where x denotes a given feature (usually financial indicator, e.g. obtained from financial statements), γ is its coefficient within the estimated model and n is a number of indicators. Linear discriminant analysis can be used to produce a direct estimate of the probability of default. It can by shown that the company's probability of default is given by:

$$PD = p(B|x_i) = \frac{1}{1 + \frac{1-\pi_B}{\pi_B} e^{z_i - \alpha}},$$

where zi is quantity defined in the previous formula above, πB represents the prior probability of default;

$$\alpha = 0.5 y'(x_A - x_B),$$

where xA and xB are the mean values of the n independent variables for the group of good financial positioned and bad financial positioned.

Frequently, the linear discriminant analysis is compared with other credit scoring techniques.

This classifier uses linear discriminant analysis which is implemented in "LDA" function of the R package.

I selected from the variables, age, income, amount.of.money.solicitated, price.of.good to buy and coefficients of linear discriminants are given as below from the LDA model:

Age	-0.026545312
Income	-0.009817864
Amount.of.money.solic.	0.002093565
Price.of.good.to.buy	-0.000960713

4.4. Logistic Regression Model:

Logistic regression builds a linear model based on a transformed target variable that can attain only values in the interval between zero and one. Thus the transformed target variable can be interpreted as a probability of belonging to the specific class. For binary classification problem, the model is especially simple, comprising just single linear function.

$\text{Log}\,[\,P(Y = +1 \mid X = x)\,/P(Y = -1 \mid X = x)] = w \cdot x + b$

The maximum likelihood estimation (MLE) method is usually used to find the parameters w and b, using the conditional likelihood of Y given the X. The left-hand side of formula is called the log-odds ratio. Note, that in the situation when both probabilities, the one in the numerator as well as the one in the denominator, will be equal to 0.5, the log-odds ratio is zero. The equation will then simplify to. This is the equation describing the

decision boundary in the logistic regression model; the set of points where the logistic regression is indecisive

Logistic regression model is implemented via "glm" function in R package.

5. RESULTS

I prefer to use 10-fold cross validation method where the dataset is formed by replicating the initial training data replicated 5 times and merged together. So dividing the initial data into 10 parts and assigning the same fold number to each row in 5 sets of initial data set. Each time the error is computed and stored in a vector for all four models and at the end the average error is computed for all models.

I have found that the least mean error is obtained when the decision tree classifier is used.

I considered error as the measure for model selection. Therefore, I will choose "Classification tree" as the ML classifier model. In particular, the decision tree approach has become a popular technique for developing credit scoring models because the resulting decision trees are easily interpretable and visualized. The parameter I used for building tree model is given below. I consider to build tree such that it compromises following 2 aspects: Low training error and a tree that is not too large

Table 4: Cross Validation Performance for Classifier Models

MODEL	MEAN ERROR
Classification tree	0.1995916
Kernel Support Vector Machines	0.2054186
Linear Discriminant Analysis	0.2754255
Generalized Linear Models	0.2006807

I added also to my test dataset two additional ratios calculated in train data: Loan-To-Value Ratio and Saving Capacity. Then I have implemented the same methodology done for preprocessing the dataset, outlier detection, missing values imputation. Then "decision tree classifier" is applied on the obtained test data set. The error on this test data set is found out to be equal to 0.201123.

6. DISCUSSION

Credit scoring models helps financial organizations to decide whether or not to grant the credit. In this project, I have identified the most common methods used in the process of credit scoring of applicants for loans and concentrate on the most relevant methods, which correspond to their use in the practice of banks. Although my dataset is only a small percentage of the Bank's total customer base, I am quite optimistic about the developments in Machine Learning Techniques. The aggregation of machine-learning forecasts of individuals could contribute a lot to the management of enterprise and systemic risk and machine-learning forecasts are considerably more adaptive to the absolute levels of default rates.

This project is done by taking into account the historical data and extracting the features in order to access the creditworthiness of the applicants. The model is derived using classification tree since it has least error comparing to another algorithms. However, there is a case where it is required analyst to make the final decision on credit approval and to minimize the resources and time the cost is introduced to assist the proper threshold.

Answering the question of which method to choose is not straightforward and depends mainly on the bank's preferences, data availability, its characteristics, etc. As follows from our short survey, the various methods are often very comparable in results. This fact can be partly explained by the mathematical relationships between these models. Often, there is no superior method for diverse data sets, however, the classification tree method has given the best performance for my dataset here. The rules that are constructed on the basis of some of these methods can be hard to explain to a manager as well as to a client, however.

6.1. Social and Ethical Aspect

When we think about social and ethical aspect for the credit scoring models, there are mainly 2 basic concerns. First of them is that the models' prediction results are constructed on the basis of some algorithms that can be hard to explain to a manager as well as to a client. We should definitely have use the advantages of these machine learning models but as we still have significant error rates in these model performance, the final decision should be made by a professional expert such as branch manager etc. With his/her expertise, he/she can better evaluate the ML model results and decide whether to give a loan or not. (Bank Personnel's Accountability for His/Her Decision to grant credit). Second one is the data protection

regulations, which can influence the utilization of the variables apart from statistical performance analyses. Additionally, legislation can affect the use of specific covariates. (Client Data Protection Laws)

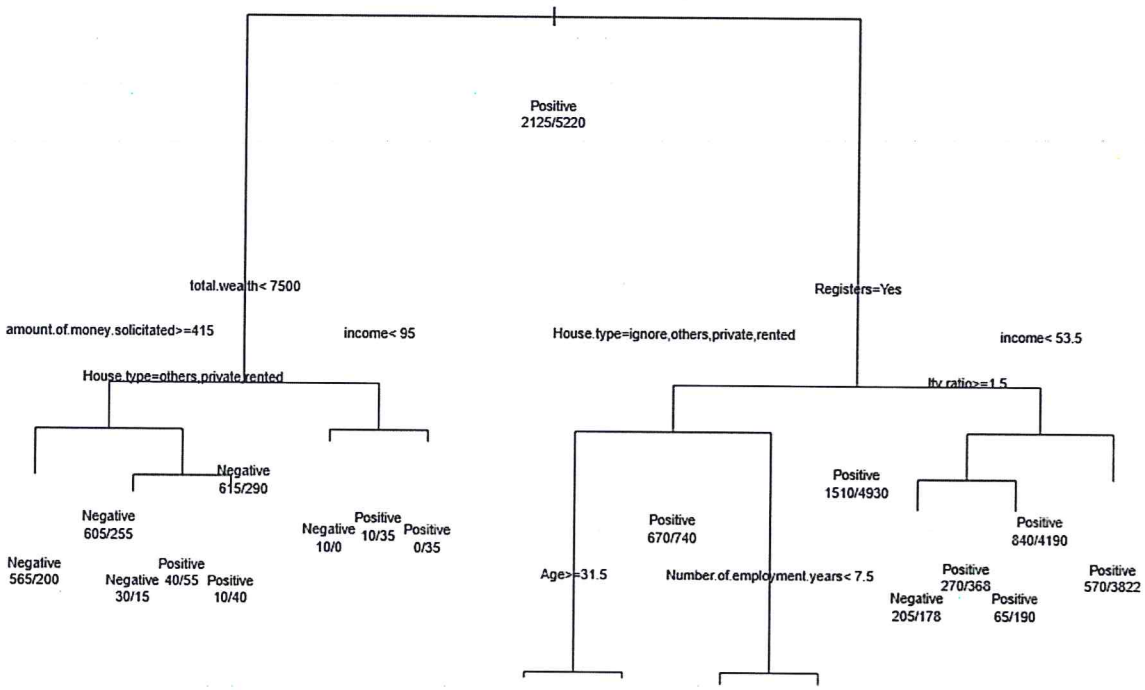
6.2. Value Delivered (Contribution)

Every year, there are millions of credit applications and financial institutions are expected to evaluate and decide on granting credits with their limited personnel and time. These financial institutes use classical statistical models to evaluate credit appraisals, and in the granting and supervision of credit. However, these classical models are not flexible or adaptive when it comes to large amounts of data input; as a result, some of the assumptions in the classical probabilistic analysis may fail. What if the analysis fail, that means the financial institution increase their risk of losing money.

Here we have the question of how they can have better accuracy of prediction and of model generalization. Using historical data and learning from them what factors matter most and how they affect the loans regular payment or delay, machine learning models start to make generalizations and provide an ML models that has given the financial innstution for new applications a single value known in a fast and efficient way as a credit score representing the lending risk. This value can help guide the decision process. The higher the credit score, the more confident a lender can be of the customer's creditworthiness. Credit scoring models here can also be defined as a form of predictive modeling that assesses the likelihood of a customer defaulting on a credit obligation. The predictive model "learns" by utilizing a customer's historical data together with similar loan data to predict the probability of that customer displaying a defined behavior in future.

APPENDIX A

Decision Tree Model



APPENDIX B

General Linear Model Output

```
Call:
glm(formula = Opinion ~ Number.of.employment.years + House.type +
    term.in.months + Age + Marital.status + Registers + Type.of.employme
nt +
    Expenses + income + total.wealth + mortgage.left + amount.of.money.s
olicitated +
    price.of.good.to.buy + saving.capacity + ltv.ratio, family = binomia
l(link = "logit"),
    data = cmp3)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.468  -0.671  -0.385   0.548   3.122

Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  4.336e+00  5.072e-01   8.548  < 2e-16 ***
Number.of.employment.years  -9.832e-01  5.029e-02 -19.550  < 2e-16 ***
House.typeownerwithdeed     -8.370e-01  5.875e-02 -14.247  < 2e-16 ***
House.typeprivate           -7.727e-02  1.020e-01  -0.758 0.448748
House.typeignore             9.780e-01  3.266e-01   2.995 0.002748 **
House.typeparents           -7.014e-01  7.670e-02  -9.144  < 2e-16 ***
House.typeothers             2.535e-01  8.756e-02   2.895 0.003786 **
term.in.months              -1.202e-02  4.189e-03  -2.870 0.004102 **
Age                         -3.813e-03  2.621e-03  -1.455 0.145791
Marital.statusmarried       -5.364e-02  8.134e-02  -0.659 0.509599
Marital.statuswidower       -5.882e-01  2.218e-01  -2.652 0.007993 **
Marital.statusseparated      9.349e-01  1.297e-01   7.205 5.79e-13 ***
Marital.statusdivorced       7.912e-01  2.218e-01   3.568 0.000360 ***
RegistersYes                 1.872e+00  5.899e-02  31.741  < 2e-16 ***
Type.of.employmentTemporaray 1.239e+00  7.117e-02  17.404  < 2e-16 ***
Type.of.employmentSelf       4.720e-01  5.490e-02   8.598  < 2e-16 ***
Type.of.employmentOthers     8.505e-01  1.107e-01   7.682 1.56e-14 ***
Expensesless miser          -2.071e-01  7.728e-02  -2.679 0.007379 **
Expensesless spenthrift      5.952e-01  1.022e-01   5.823 5.78e-09 ***
Expensesspenthrift           2.025e+00  5.745e-01   3.524 0.000424 ***
income                      -7.690e-03  3.277e-04 -23.467  < 2e-16 ***
total.wealth                -2.612e+00  4.827e-01  -5.412 6.25e-08 ***
mortgage.left                1.503e-04  2.044e-05   7.352 1.95e-13 ***
amount.of.money.solicitated  2.117e-03  8.834e-05  23.968  < 2e-16 ***
price.of.good.to.buy        -1.064e-03  7.061e-05 -15.061  < 2e-16 ***
saving.capacity                     NA         NA      NA       NA
ltv.ratio                    1.533e-03  1.919e-03   0.799 0.424375
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 17638  on 14929  degrees of freedom
Residual deviance: 12872  on 14904  degrees of freedom
AIC: 12924
Number of Fisher Scoring iterations: 5
```

APPENDIX C

Support Vector Machine (SVM)

```
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  5.80132466935897e-06

Number of Support Vectors : 7807

Objective Function Value : -33972.75
Training error : 0.204354
```

APPENDIX D

Linear Discriminant Analysis (LDA)

```
Call:
lda(cmp3$Opinion ~ cmp3$Age + cmp3$income + cmp3$amount.of.money.solicit
ated +
    cmp3$price.of.good.to.buy, data = cmp3, CV = FALSE)

Prior probabilities of groups:
 Positive  Negative
0.7223711 0.2776289

Group means:
         cmp3$Age cmp3$income cmp3$amount.of.money.solicitated cmp3$pric
e.of.good.to.buy
Positive 37.75522    146.6305                          982.1813
1454.160
Negative 35.69119    104.5578                         1150.4318
1466.698

Coefficients of linear discriminants:
                                        LD1
cmp3$Age                        -0.0193210356
cmp3$income                     -0.0078075807
cmp3$amount.of.money.solicitated  0.0023067302
cmp3$price.of.good.to.buy       -0.0009713067
```

REFERENCES

Hand, D.J. and Henley, W.E. (1997) Statistical Classification Methods in Consumer Credit Scoring: A Review. Journal of Royal Statistical Society, 160, 523-541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Bijak, Katarzyna and Thomas, Lyn C., Modelling LGD for Unsecured Retail Loans Using Bayesian Methods (February 2015). Journal of the Operational Research Society, Vol. 66, Issue 2, pp. 342-352, 2015. Available at
SSRN: https://ssrn.com/abstract=2545857 or http://dx.doi.org/10.1057/jors.2014.9

Quinlan, J. Mach Learn (1986) 1: 81. https://doi.org/10.1023/A:1022643204877

Abdou, H. & Pointon, J. (2011) 'Credit scoring, statistical techniques and evaluation criteria: a review of the literature ', Intelligent Systems in Accounting, Finance & Management, 18 (2-3), pp. 59-88.

Kennedy, K., Mac Namee, B., Delany, S. J., O'Sullivan, M., & Watson, N. (2013). A window of opportunity: Assessing behavioural scoring. Expert Systems with Applications: An International Journal, 40(4), 1372-1380. doi:10.1016/j.eswa.2012.08.052

Kennedy, K. (2013). Credit scoring using machine learning. Doctoral thesis. Dublin Institute of Technology. doi:10.21427/D7NC7J.

van Buuren, S., & Groothuis-Oudshoorn, C. G. M. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of statistical software, 45(3).

Cortes, C. & Vapnik, V. Machine Learning (1995) 20: 273.
https://doi.org/10.1023/A:1022627411411

Louzada, F., Ara, A., Fernandes, G. B.(2016) Classification methods applied to credit scoring: A systematic review and overall comparison. Surveys in Operations Research and Management Science, 21(2), pp.117-134

Zhang D., Zhou X., Leung, S.C.H., Zheng J. (2010) Expert Systems with Applications. 37 pp. 7838–7843

Gurný, P., Gurný, M. (2013) Comparison of credit scoring models on probability of default estimation for US banks. Prague Economic Papers. doi: 10.18267/j.pep.446