UNIVERSITÉ DE LORRAINE - FACULTÉ DE DROIT, SCIENCES ÉCONOMIQUES ET GESTION

MASTER 2
ECONOMIE APPLIQUEE
INGENIERIE FINANCIERE DE MARCHE

EXAM OF

MULTIVARIATE ANALYSIS AND DATA MINING FOR FINANCIAL TIME SERIES

LOANS DEFAULT PREDICTION - GIVE ME CREDIT

TASKS REPARTITION	
Karima BOUSSAA **50% of the work**	Literature Review, Statistical learning method(s) description, Data description and analysis
Kossi Olivier ADANOU **50% of the work**	Coding (full RMarkdown script) Statistical and empirical programming and analysis Results (outputs/graphics) presentation & interpretation Conclusion.

Submitted by:

Kossi Olivier ADANOU
Karima BOUSSAA

Course Coordinator:

Sophie BEREAU

2022/2023

Table des matières

Introduction

Humans or companies apply for loans for various reasons. People may apply for loans for a purchase or personal investment. Organizations and companies might take a loan to grow their businesses. Banks play crucial role in providing these entities with funds so that the market and society should function properly. Unfortunately, not every loan application can be approved. In order to make a decision on a loan application, the financial establishments (banks) looks at the credit history of these entities. One of the core functions of a bank is to give out loans to consumers and companies. For each loan, the bank is at risk of not receiving back the entire principal. The amount of risk usually has an influence on interest the bank will receive. For a lender it is valuable to be able to estimate the risk associated with each client, it can help a bank in two ways. First, it can be used in the loan origination process. This is the process consisting of all the steps a borrower and lender go through to process the application of a new loan. The second situation where risk estimation is crucial, is in monitoring the already accepted loans. If a bank is capable of estimating which loans are likely to default, those loans can be handled with more attention. This can be done with the aim of increasing the recovered amount after a default, or preventing the loan from going into default at all. There are number of parameters to determine if a an entity is eligible for a loan. It's can be used to predict whether an entity will be suitable for giving a loan or not. Predicting whether a borrower would default on his or her loan is of vital importance for bankers, as default prediction accuracy will have great impact on their profitability. Previous efforts have been made in this domain using machine learning based on different attributes. Credit scoring algorithms, which make a guess at the probability of default, are method banks use to determine whether or not a loan should be granted. Recent surveys show that credit institutions are increasingly adopting Machine Learning (ML) tools in several areas of credit risk management, like regulatory capital calculation, optimizing provisions, credit-scoring or monitoring outstanding loans (IIF 2019, BoE 2019, Fernández

2019). So we shall use here ML to provide a prediction of default status. In this exercise we requires to improve on the state of the art in credit scoring, by predicting the probability that somebody will experience financial distress in the next two years. The goal of this study is to build a model that borrowers can use to help make the best financial decisions.

The rest of this exam work is organized as follows. Background and related work are covered in Section 2. Data collection and exploration is covered in Section 3. The methodology of machine learning analysis is outlined in Section 4. This is followed by a detailed discussion of our analysis, culminating with the final results being presented in Sections 7,8 and 9.

Problem description

As mentioned before, estimating the risk of a loan is an important task within banking. This makes it interesting to research the possible improvements that can be reached by applying statistical learning methods to provide a prediction of default status and document the best performing method(s) by means of a cross-validation exercise. Several scientific papers have been written about the expected benefits of using machine learning in default prediction (Abellán and Mantas, 2014; Harris, 2015; Huang, Chen, and Wang, 2007). These show that machine learning can lead to a higher in accuracy default prediction, compared to conventional methods.

Study objective

Now that the problem has been described, an objective for this exam project is defined. The objective is to develop knowledge about the predictive performance of different machine learning algorithms when used to predict defaults in loans. This is done by implementing different machine learning algorithms (statistical methods) which can be used to classify samples. These different algorithms will be used to predict the defaults on loans. The loans come from two data sets with loans of which it is known whether or not they went into default

Learning Methods

To predict borrowers default, we use for this project Logistic Regression, Support Vector Machine (SVM), Bagging, Random Forest, Boosting, Naive Bayes, Neural Networks, Decision tree on test dataset, XGBOOST, cross validation and K-Nearest Neighbor (KNN) models.

1. Methodology

Now that the required theoretical background has been discussed, the next step is to describe in more detail how the objectives of this project will be achieved. This experiment design consists of a combination of topics discussed in the previous chapter. The design is split into two stages, the first stage is the data processing, the second stage is the model training and testing.

Project framework

In this section, the approach to achieve the project objectives will be discussed. This approach forms the project framework. The first step in the research framework is to gather and read current scientific literature on three topics. These topics are machine learning theory, credit risk theory and statistical theory. The machine learning theory is used to create different models that are capable of classifying credit based on the likelihood of default. These models are first confronted with a data set to train the models and then confronted with a data set to get the model results. All three sources of literature are used to create assessment criteria to objectively compare the performance of the models. Part of the assessment criteria is the performance of a benchmark method representing the currently used credit scoring method. For this benchmark logistic regression is used. The last step in the project framework is to draw conclusions based on the confrontation between the model results and the assessment criteria.

Procedure Outline

The objective of this project is to train a classifier that can predict if an individual will experience financial distress in the next two years given the set of attributes listed above. This section outlines that process. Data preparation and feature selection are outline in Section 3; this details the process of creating and selecting features. This section also discusses the division of data into development, cross-validation and holdout sets. In Section 4, exploratory data analysis is done on the *development* data set. Section 5 presents a baseline performance, using Random Forests with default settings using a 10-fold cross validation on the *cross-validation* data set. Based on error analysis results, feature space redesign is performed in

Section 6; this includes a comparison of baseline and optimized performance. Finally, the optimized model is trained on both the *cross-validation* and *development* sets. This is then used to classify the instances in the *holdout* data set. The results of this are present in Section 7.

Data processing

As mentioned above, the first stage in conducting the experiment is to prepare the data for usage in machine learning.

Data Collection

The data sets used for this project has been retrieved from the 2011 Kaggle Competition: Give Me Some Credit and is obtained on ARCHE. In this project two different data sets will be used. Each of the data sets will be discussed separately. For each set a short summary will be given and the different variables are described. Variables such as Monthly Income, Number of Dependents, age, number of open credit lines and loans etc are in both datasets. Next the data sets will be prepared. After the data set has been prepared, the performance of the different resample methods is measured. The reason for using multiple data sets is to decrease the influence of the individual data sets. If a finding is supported by all two data sets it is more reliable.

2. Data Preparation

Before predicting probabilities of default, it is necessary to define borrowers default. We define a default event as a case of a payment overdue by more than 90 days according to the Basel II (III) Internal ratings-based (IRB) methodology (BCBS 2017).

Description of the Dataset

We use in this study, 2 data sets : test dataset and training dataset. The "testset" is constituted of 101503 observations and 12 variables while "trainingset" is constituted of 15000 observations with 12 variables.

Data Cleansing

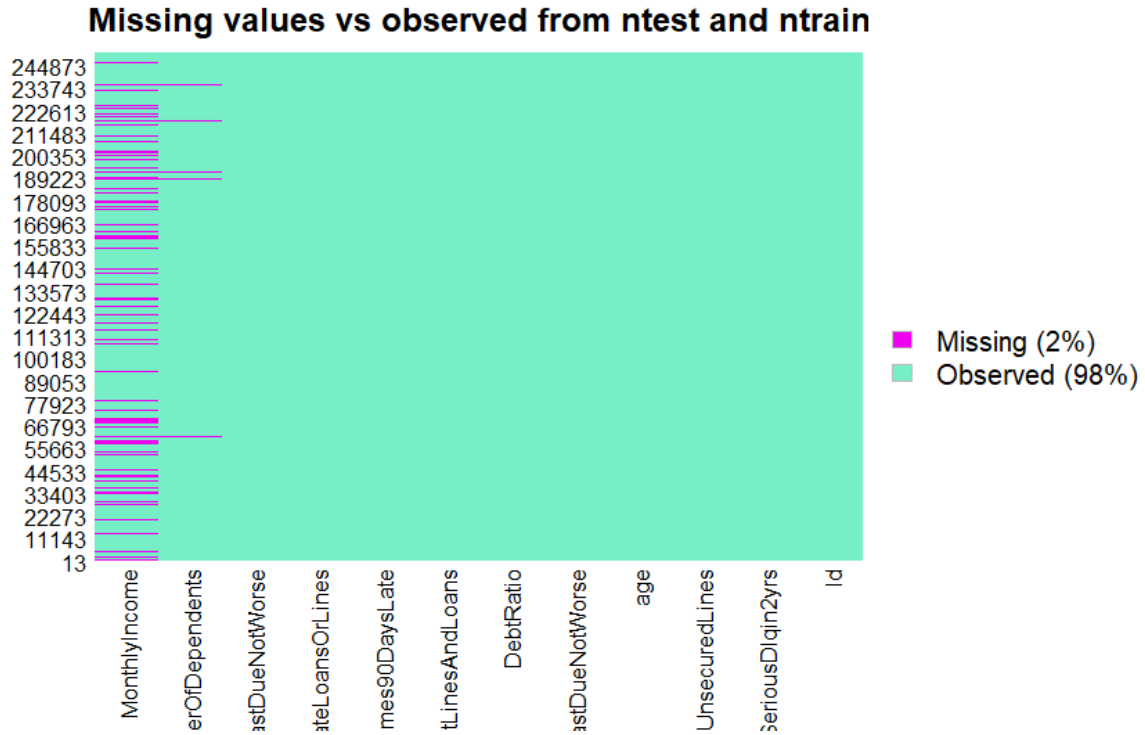
Before analysing a set data it is important to check as far as possible that the data seem correct. It is also important to identify features that may cause difficulties during the analysis. Three specific aspects are considered In this chapter : missing data, outlying values, and the possible need for data transformation.

Missing Observations

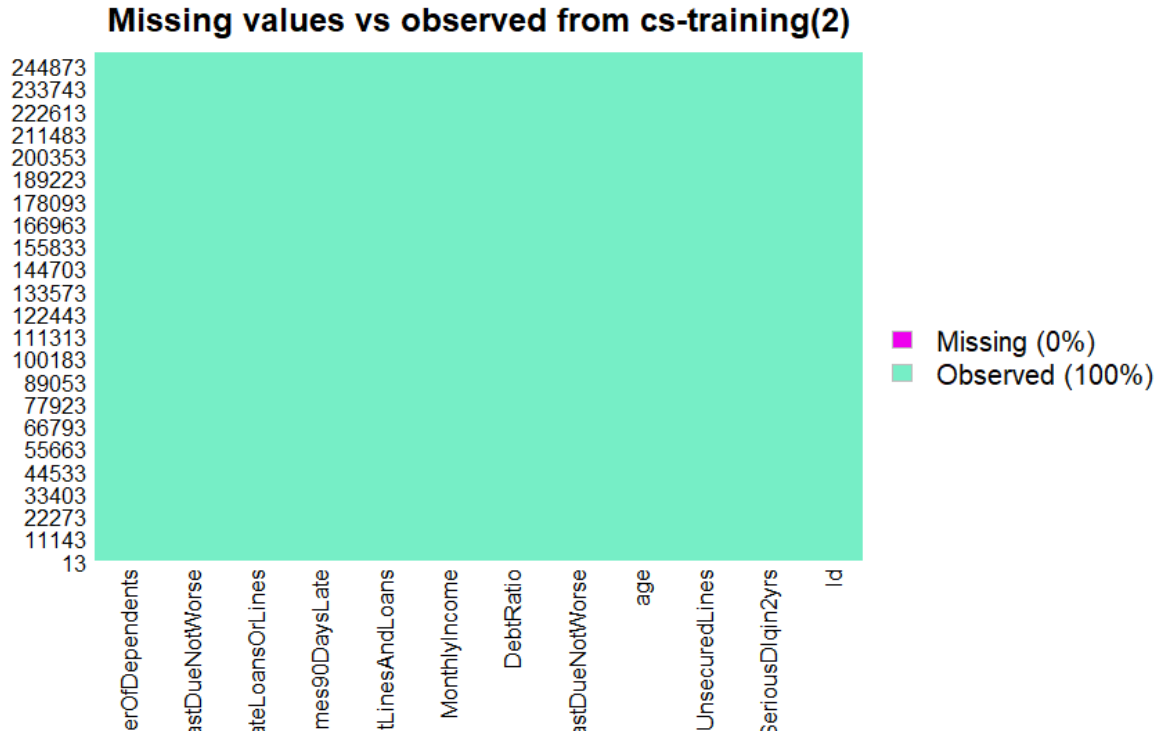
To visually explore the missing values, we plot the number of missing values for each variables in datasets. Graphs 1 represents the variables on the 'X' axis and the volume of the missing items on the 'Y' axis. In "testset", "DebtRatio" contain 20103 missing values and 'NT6089DPDNWorse' contain 2626 missing values, while in "trainingset", "MonthlyIncome" is variable with 29731 missing values. "NumbDependents" which is empty will be deleted in test dataset but its missings values (3924) in training dataset will be replace by median to solve missing problem. In the second step, since the missing values did not exceed 25% of the data, let's replace the missing values with the median of the variables.

. Graphs 2 is the new representation of missings values versus variables after replacing by median the variables with no observations.

Graph1: Training and test data missing values before cleaning



Graph2 : Training and test data missing values after cleaning



As we can we see, in graphs 2 missings values are no more.

Outlying values

Checking the data for continuous variables may reveal some outlying values that are incompatible with the rest of the data. Typically there may be one or two outliers for a few variables, although for most variables there will not be any. Outliers are particularly important because they can have a considerable influence on results of our statistical analysis. Because by definition, they are extreme values, their inclusion or exclusion can have a marked effect on the results of an analysis. We will identify outliers in our data in next section.

Variables description

- **Instance Number:** This variable contains the instance number data.
- **SeriousDlqin2yrs:** This is of binary type. Our algorithm is used to predict this. This depicts whether a person experienced 90 days past due delinquency or worse.
- **RevolvingUtilizationOfUnsecuredLines:** Total balance on credit cards and personal lines of credit except real estate and no instalment debt like car loans divided by the sum of credit limits. This is in percentage.
- **Age:** Contains the age of the borrower in years. It is of integer type. This column didn't contain any missing data.
- **NumberOfTime30-59DaysPastDueNotWorse:** This column contains number of times borrower has been 30-59 days past due but no worse in the last 2 years.
- **DebtRatio:** This field contains data in percentage form. It is obtained by dividing the sum of monthly debt payments, alimony, living costs with monthly gross income.
- **MonthlyIncome:** This column contained the information about the monthly income of an individual.
- **NumberOfOpenCreditLinesAndLoans:** This column contained information about the number of open loans such as car loans, house loans and lines of credit (ex. Credit card).

- **NumberOfTimes90DaysLate:** This column had information about the number of times an individual was late by 90 days or more in paying their bills.
- **NumberRealEstateLoansOrLines:** This column contained information about number of mortgage and real estate loans including home equity lines of credit an individual have taken.
- **NumberOfTime60-89DaysPastDueNotWorse:** This field contains the information about the number of times borrower has been 60-89 days past due but no worse in the last 2 years.
- **NumberOfDependents:** This column contained information about the number of dependents in the family excluding themselves.

Programming Language

We performed exploratory data analysis using the R software.

3. Statistical analysis

Descriptive Statistics Analysis

The descriptive statistics for the training data set variables are shown in tables Table 1. A structural analysis of the dataset gives a look at the makeup of the dataset based on its variables. All variables are either numeric or integers with the integer variables being categorical or auto-increment variables (ID) and the numeric variables being amounts. Table 1 present different statistics of our dataset.

	mean	sd	median	trimmed	mad	min	max	skew	kurtosis	se
SDl	1.05	0.21	1.00	1.00	0.00	1	2	4.30	16.49	0.00
RevUnL	5.75	229.63	0.15	0.27	0.22	0	50708	89.59	13043.27	0.46
age	52.34	14.78	52.00	52.01	16.31	0	109	0.19	-0.50	0.03
NT30.59	0.43	4.34	0.00	0.08	0.00	0	98	21.88	489.00	0.01
DebtRatio	349.56	1884.79	0.37	50.94	0.36	0	329664	90.93	13487.72	3.76
MIncome	6478.35	23035.41	5400.00	5646.92	2489.29	0	7727000	218.46	60930.00	45.93
NLLoans	8.45	5.15	8.00	7.96	4.45	0	85	1.22	3.18	0.01
NT90DL	0.28	4.31	0.00	0.00	0.00	0	98	22.34	502.85	0.01
NRELOL	1.02	1.12	1.00	0.88	1.48	0	54	3.21	49.20	0.00
N60.89	0.25	4.30	0.00	0.00	0.00	0	98	22.56	509.68	0.01
Ndepend	0.74	1.12	0.00	0.52	0.00	0	43	1.83	11.31	0.00

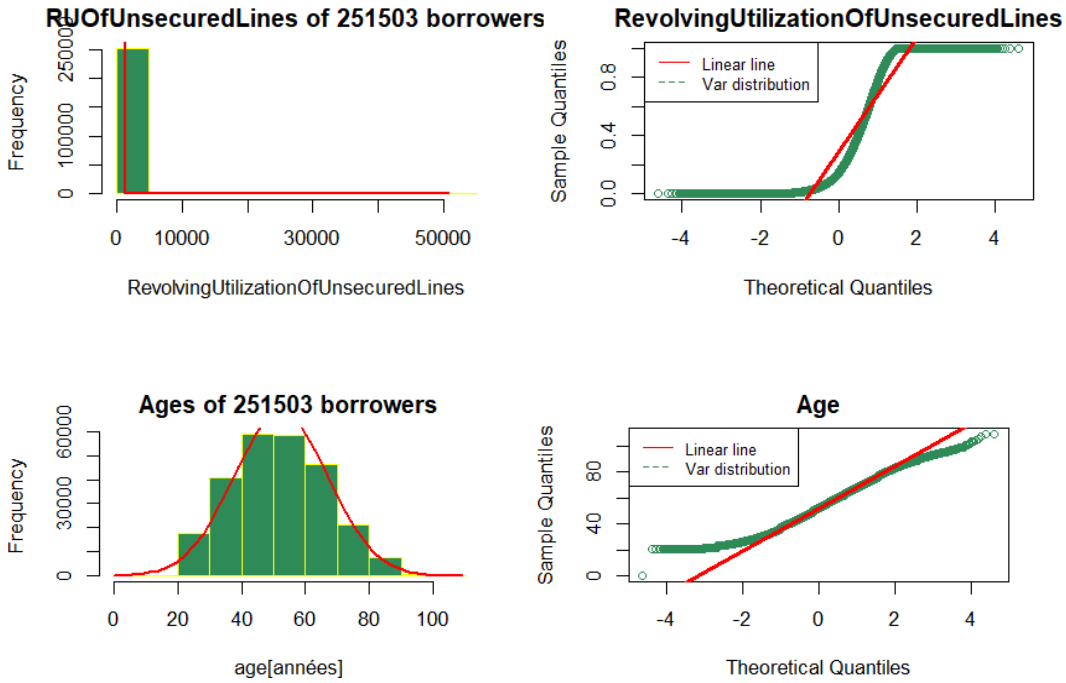
Table 1 : training statistic descriptive table

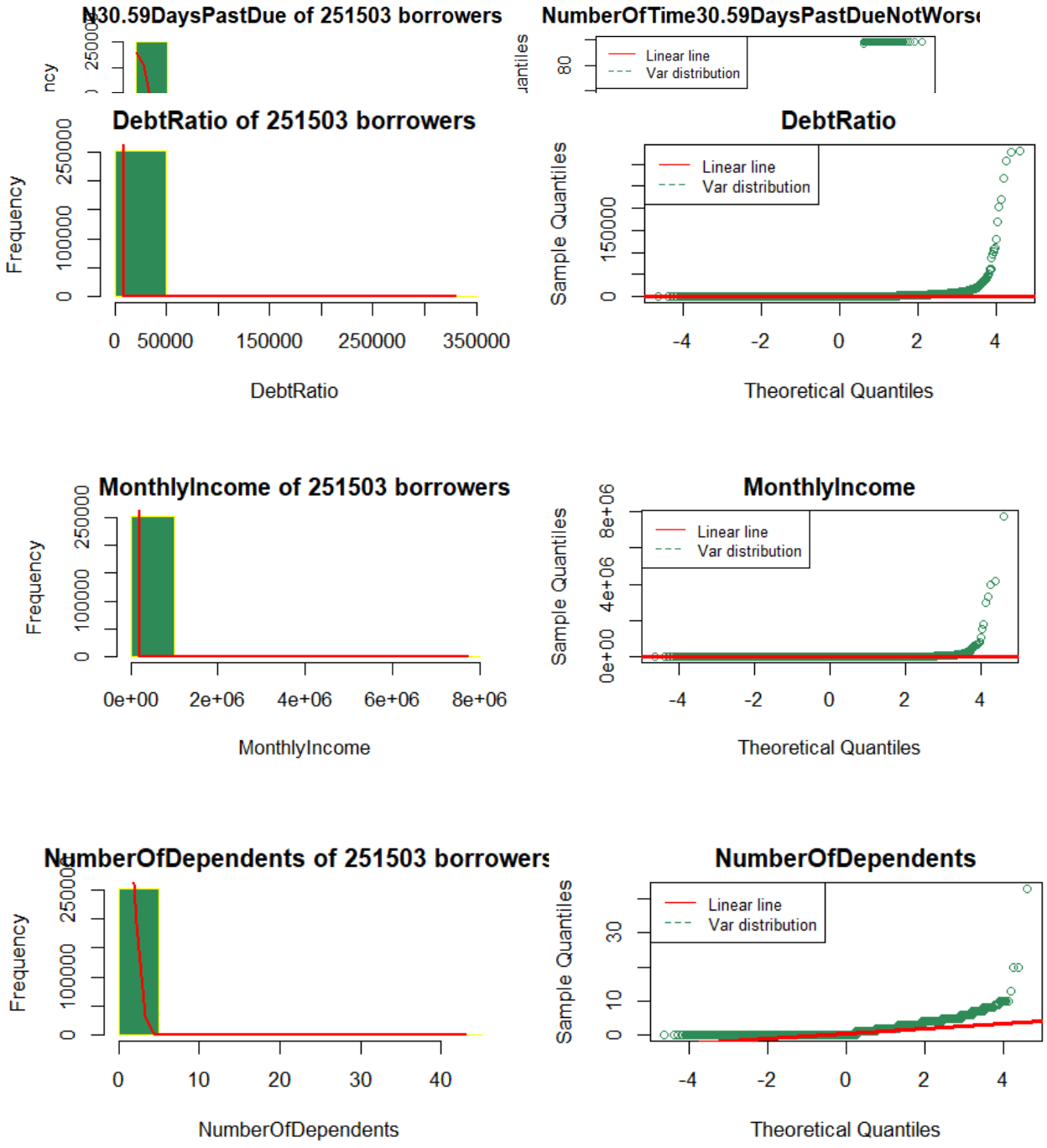
Source : ADANOU&BOUSSAA analyses (January, 2023)

"Skewness assesses the extent to which a variable's distribution is symmetrical. If the distribution of responses for a variable stretches toward the right or left tail of the distribution, then the distribution is characterized as skewed. A negative skewness indicates a greater number of larger values, whereas a positive skewness indicates a greater number of smaller values. As a general guideline, a skewness value between −1 and +1 is considered excellent, but a value between −2 and +2 is generally considered acceptable. Values beyond −2 and +2 are considered indicative of substantial nonnormality." (Hair et al., 2022, p. 66). According to our results in table 1 column 11, only two variables, "RevolvingUtilizationOfUnsecuredLines" and "DebtRatio" has a negative skewness indicating a greater number of larger values. The rest of

variables has a positive skewness so indicating greater number smaller values. In column 12, we can observe kurtosis values. It is a measure of whether the distribution is too peaked (a very narrow distribution with most of the responses in the center). “RevolvingUtilizationOfUnsecuredLines” and “DebtRatio” and “age” have a negative kurtosis. It indicates that those variables have a shape flatter than normal. In contrast, the rest of variables have positive value for the kurtosis indicating a distribution more peaked than normal. Analogous to the skewness, “age”, “RevolvingUtilizationOfUnsecuredLines ”, “DebtRatio” have both skewness and kurtosis close to zero, the pattern of responses is considered a normal distribution. We can see it graphicly on graphs 5.

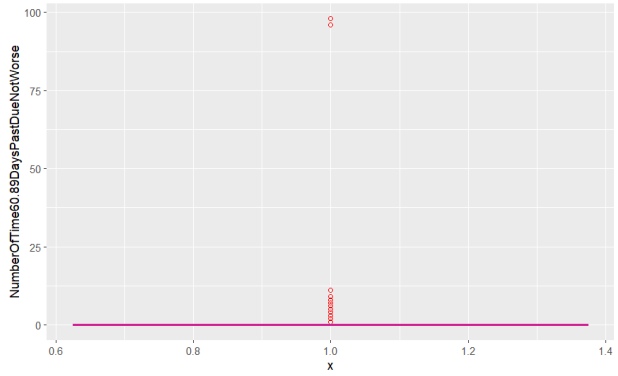
Graphic 5 : Variables distribution

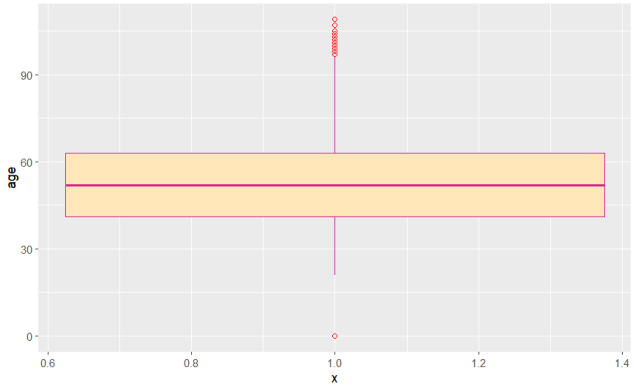


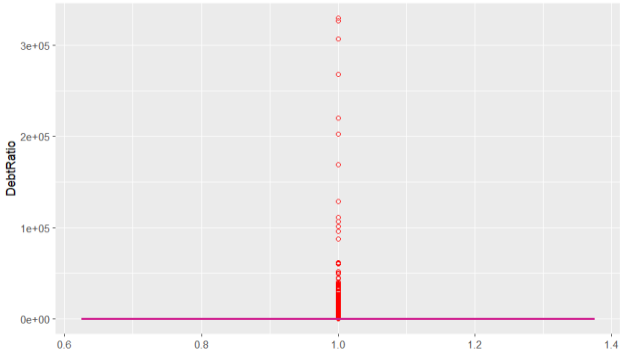


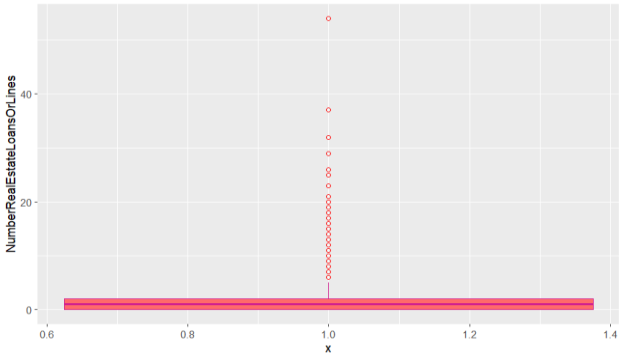
Source : ADANOU&BOUSSAA analyses (January, 2023)

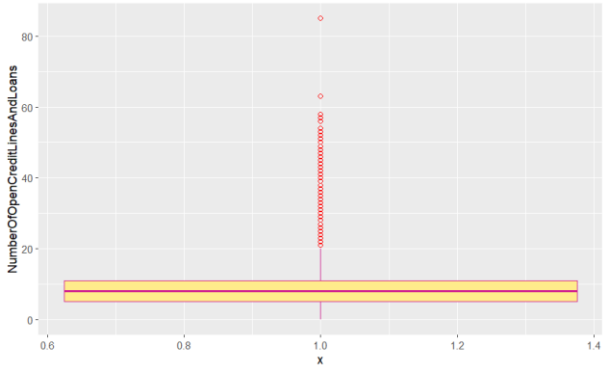
Outliers analysis :

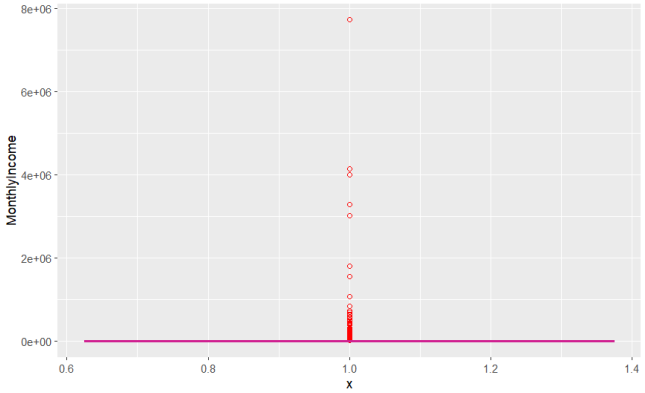


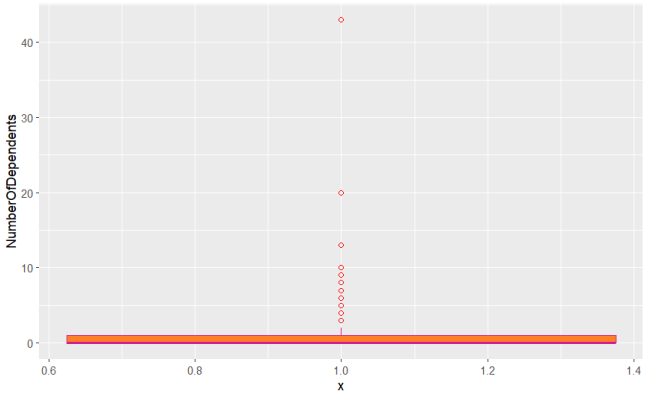


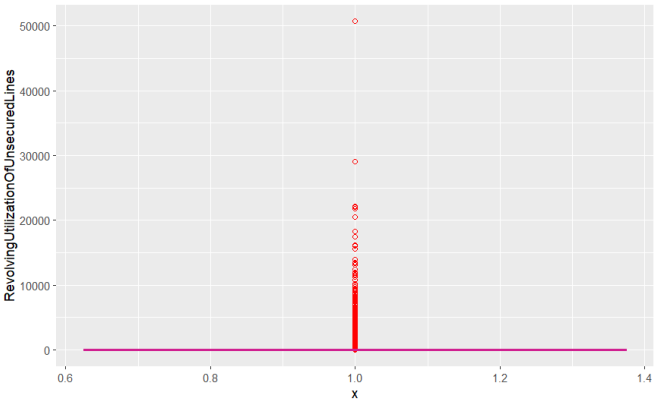












- First, we noticed RevolvingUtilizationOfUnsecuredLines field had some individuals with higher credit utilization greater than 100 percent. These are cases we will need to remove because no one can go over their max usage of credit. So we will replace those values greater than 100 percent with NA's.
- In the boxplot for Age, there are several outliers in the upper whisker and one in the lower whisker that needs to be removed. So we will replace those outliers with NA's as well.
- In the boxplot of NumberOfTime30.59DaysPastDueNotWorse, it's really difficult to find if there are cases of outliers, other than the two in the upper extreme, since there are no quartile boxes or whiskers to interpret. So, we will plot a histogram to see if we can get a better look at possible outliers. Noticed, the observations on the histogram displays values only less than approximately to 10.We can assume these values are outliers and replace them with NA.
- The DebtRatio field had some individuals with credit usage greater than 100 percent. These are cases we will need to remove since no one can borrow money than their max credit given to them. So we will replace those values greater than 100 percent with NA's.
- In the boxplot of MonthlyIncome, it's really difficult to find if there are cases of outliers, other than the one in the upper extreme, since there are no quartile boxes or whiskers to interpret. So, we will plot a histogram to see if we can get a better look of possible outliers. Noticed, the spread of the data is very skewed to the right and does not take shape of a normal distribution. Moro ever, we are going to replace the Monthly Income greater 14000 to NA, so we can reduce inaccurate classifications errors before using several machine learning techniques later.

- In the Number of Open Credit Lines And Loans boxplot, there are several outliers in the upper whisker and the data looks slightly skewed to the right. Let's take a look at the histogram to be sure they are outliers and data is slightly skewed to the right. The outliers in the boxplot could have caused it to not take the shape of a normal distribution, so let's remove them.
- In the Number Real Estate Loans Or Lines boxplot, there are several outliers in the upper whisker and only one at the very top. Let's take a look at the histogram to be sure they are outliers. Noticed, the data is slightly skewed to the right and does not have a bell-shaped curve. Moreover, the outliers in the boxplot could have caused it to not take the shape of a normal distribution, so let's replace those values greater than 7 them with NA.

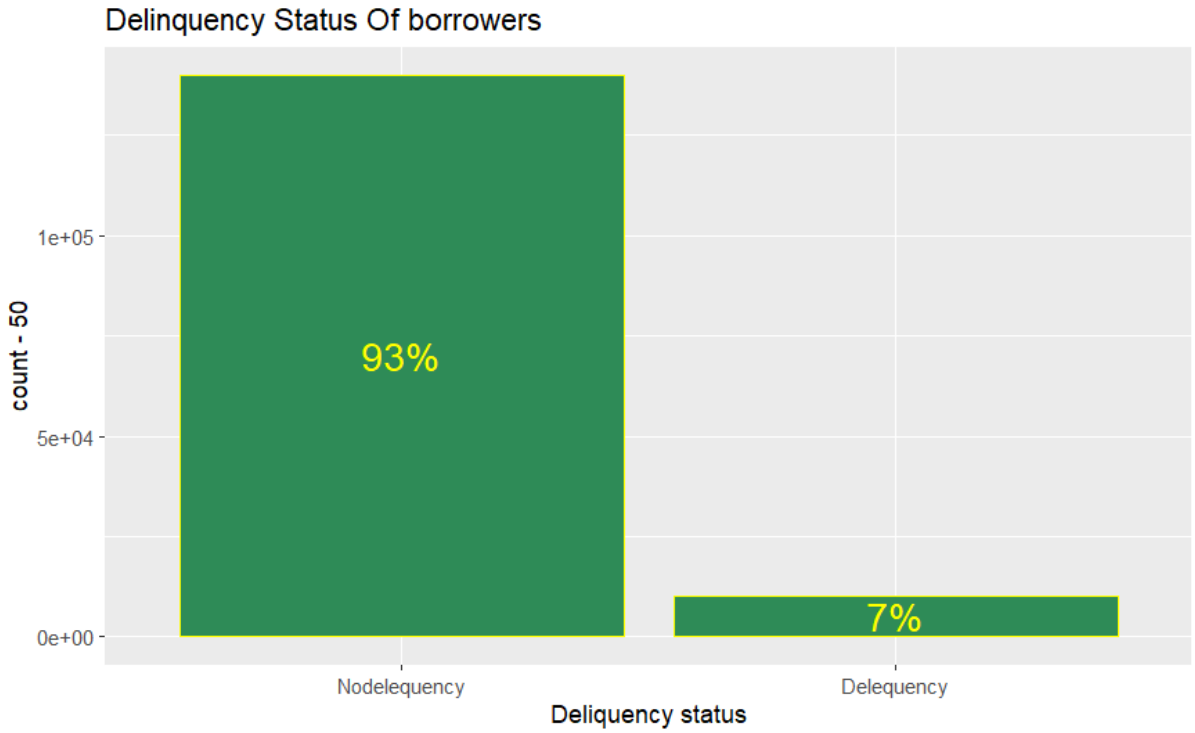
In the Number of Dependents boxplot, there are several outliers in the upper whisker. Let's take a look at the histogram to be sure they are outliers. Since it doesn't display the values greater than 5, let's assume these are outliers and replace them with NA.

By observing our results, MonthlyIncome and DebtRatio clearly have outliers. During data visualization other outliers like RevolvingUtilizationOfUnsecuredLines, NumberOfOpenCreditLinesAndLoans and NumberRealEstateLoansOrLines were also identified. Outliers can be identified and replaced using percentile method as well. That is what we do in next section.

Univariate analysis

SeriousDlqin2yrs

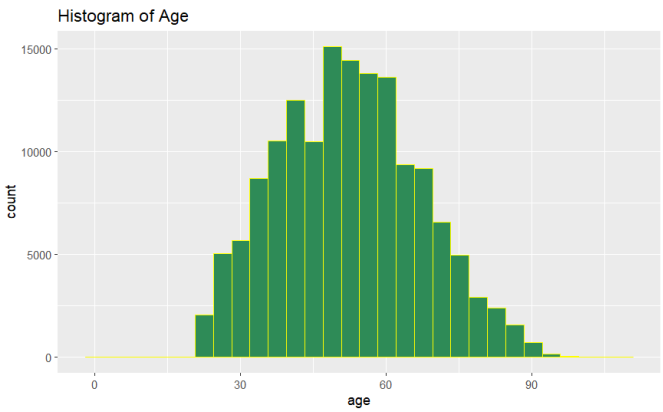
Graph 6 : Borrowers delinquencies

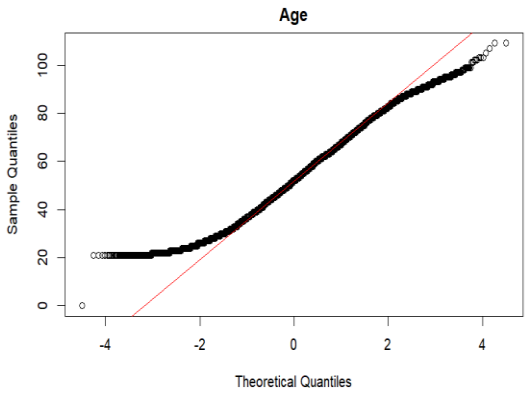


The following graph gives a good insight about the number of 0s and 1s in the datasets. We know that around 7% of borrowers experienced 90 days past due delinquency or worse in two years.

Age

Graph 7 : Age description

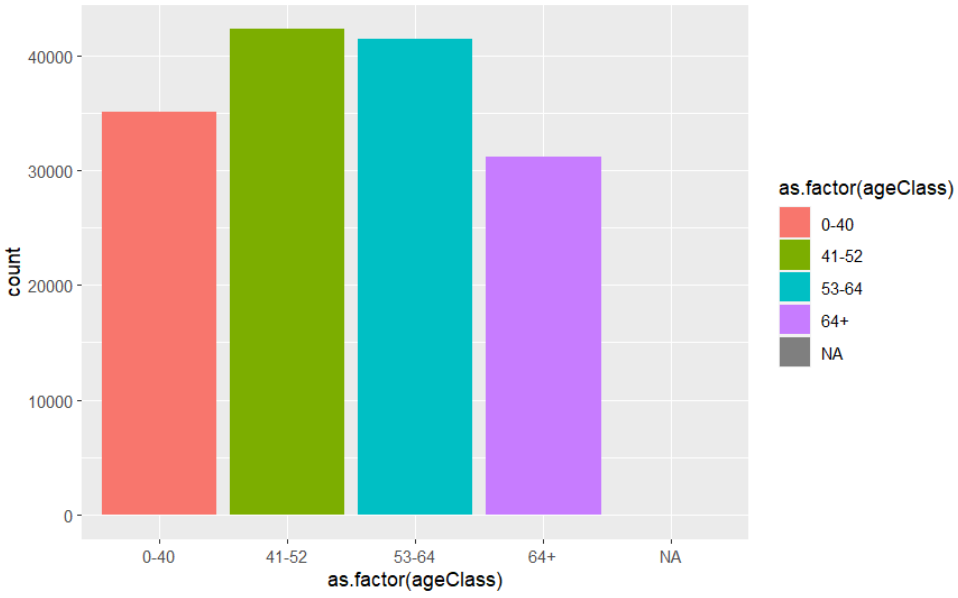




Looking at the borrowers age histogram, the sample has no tails beyond -4 or 4. This presents an interesting looking qq plot that is depicted upon. The light tailed distributions yield an s shape depicted in the qq plot. Approximately, from the values (-4, -2.5), the sample grows slower than the standard normal distribution. Therefore, It takes longer for the sample quantile to increase. This is shown by the concave up portion of the graph. From the value (-2.5, 2.5), the sample seems to grow at approximately the same pace as the standard normal distribution. Therefore their quantiles match in this region. Lastly from the value (2.5,4) the sample grow faster than the standard normal distribution. Therefore the sample reaches its highest quantile before the standard normal distribution. This why our variable age looks flat at the top. The sample has reached its highest quantile, but the standard normal has not and still needs to increase a little to reach it. In the histogram, the skewness of Age, looks normally distributed.

Then we regroup borrowers age by classes.

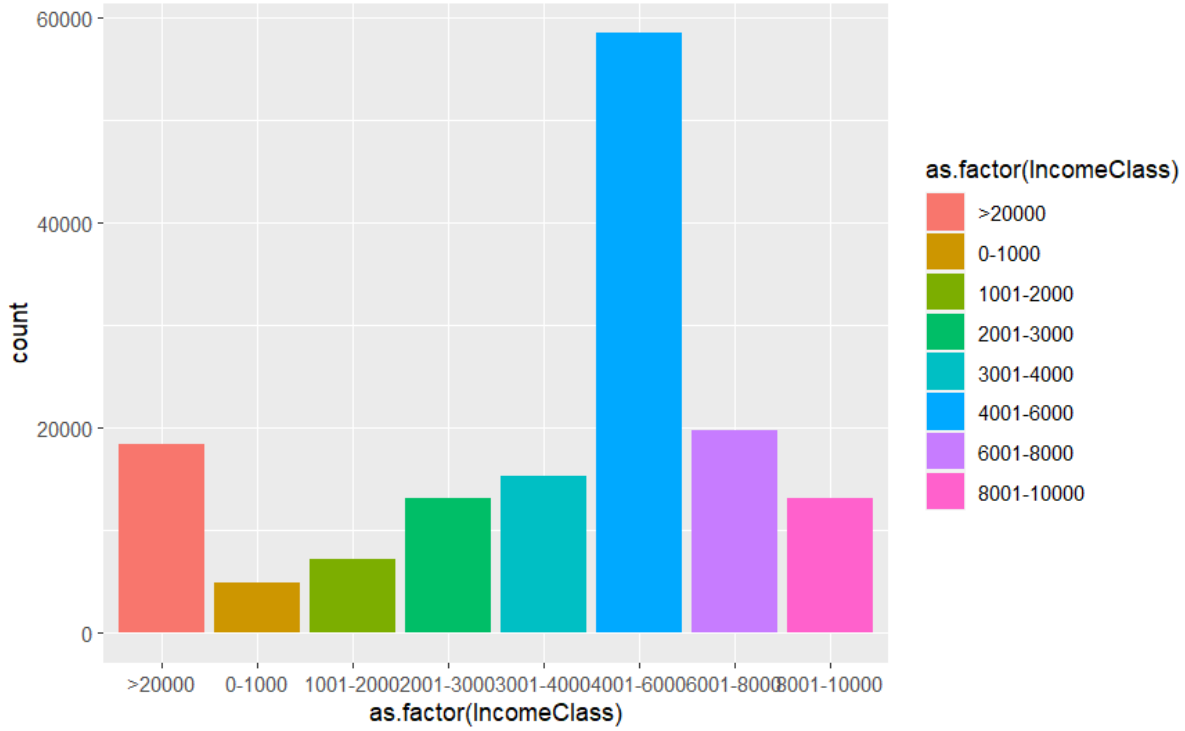
Graph 8 : Age analysis



Most Borrowers in our data are between 41 and 52 years old. Then follow those aged from 53 to 64 years old.

Borrowers monthly Income

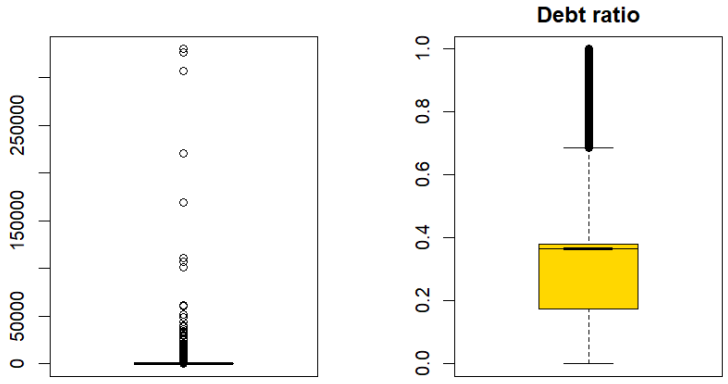
Graph 9 : monthly Income description



In the same data base, we make a class of borrower's monthly income after analyse its distribution. It reveal that most of our borrowers own between 4001 and 6000 unitary monetary. Less paid borrowers (0-1000) are less represented.

Debt ratio

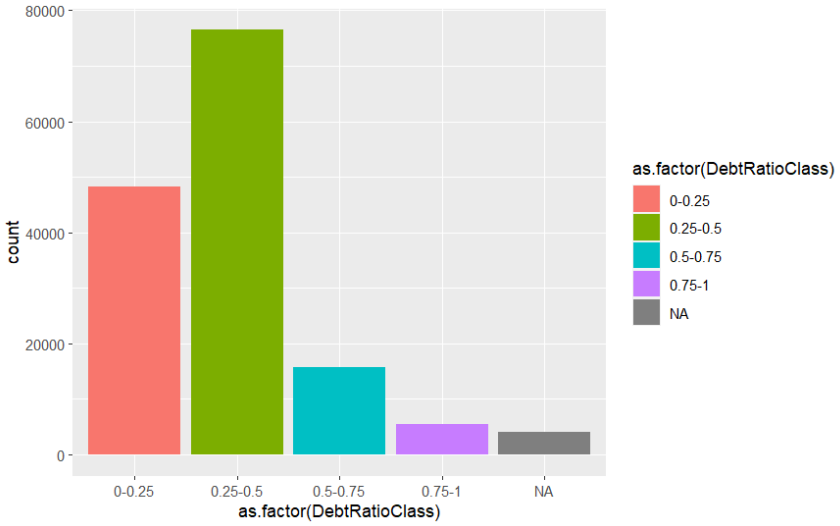
Graph 10a : Debt Ratio description



with the boxplot on left, we remark that borrowers with debt ratio greater than 100k seems weird. we remove them to simplify our study. Now debt ratio is between 0 and 1. At the right, we can see the new boxplot.

The boxplot for Debt ratio is slightly disturbing in that the median is close to the upper quartile and the lower whisker is shorter than the upper one, which would be suggesting positive skewness. Also there is an outlier and Pearson's correlation is sensitive to these as well as skewness.

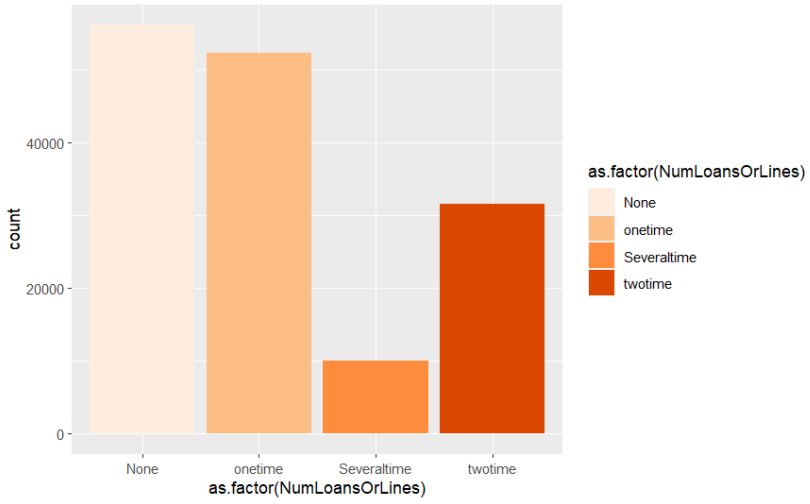
Graph 10b : Debt Ratio analysis



Debt ratio between 0.25 and 0.5 is the most represented.

NumberRealEstateLoansOrLines

Graph 11 : NumberRealEstateLoansOrLines analysis

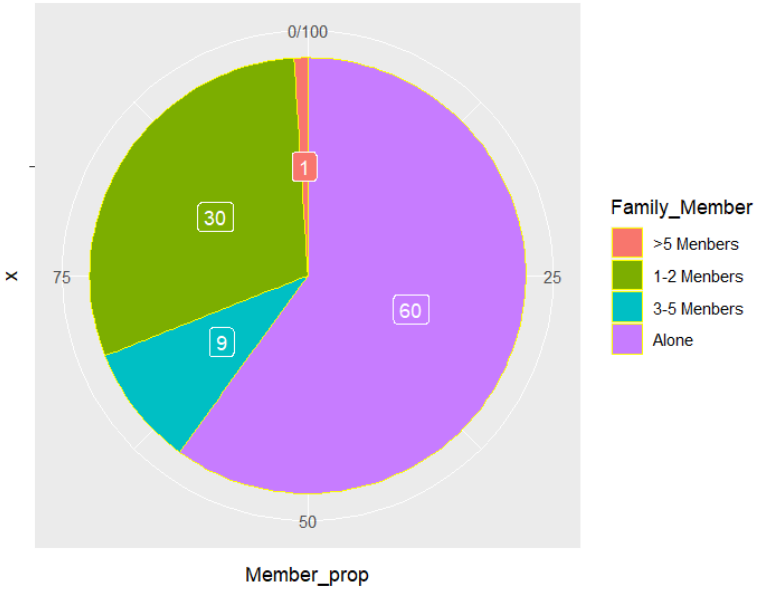


In this data base, many barrowers never subscribe for mortgage and real estate loans including home equity lines of credit an individual have taken. Those who subscribe it have done it one time. Borrowers with multiple number for mortgage and real estate loans including home equity lines of credit an individual have taken are fewest.

NumberOfDependents

This variable, gives information about borrowers number of dependents in the family excluding themselves. By analysing this variable in the training data set, we observe that :

Graph 12 : NumberOfDependents analysis



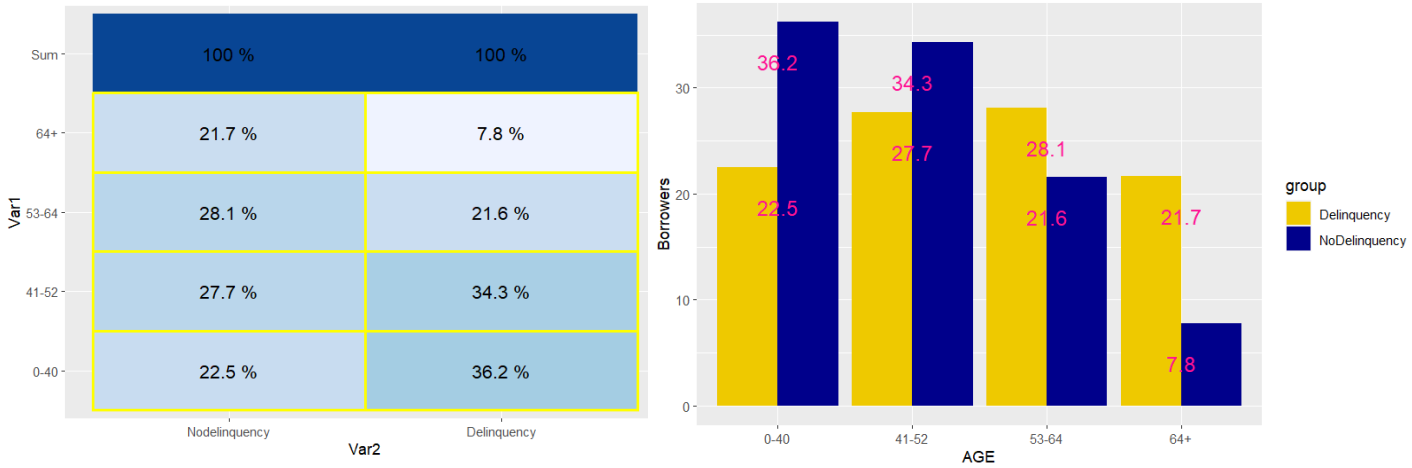
60% of borrowers don't have any family members while 30% have 1 to 2 members in their family excluding themselves. Family with more than five members are fewest.

Bivariate analysis

age variable X SeriousDlqin2yrs (delinquency)

we know that "SeriousDlqin2yrs" depicts whether a person experienced 90 days past due delinquency or worse but we don't know how old are those who experienced this. Here we analyse it.

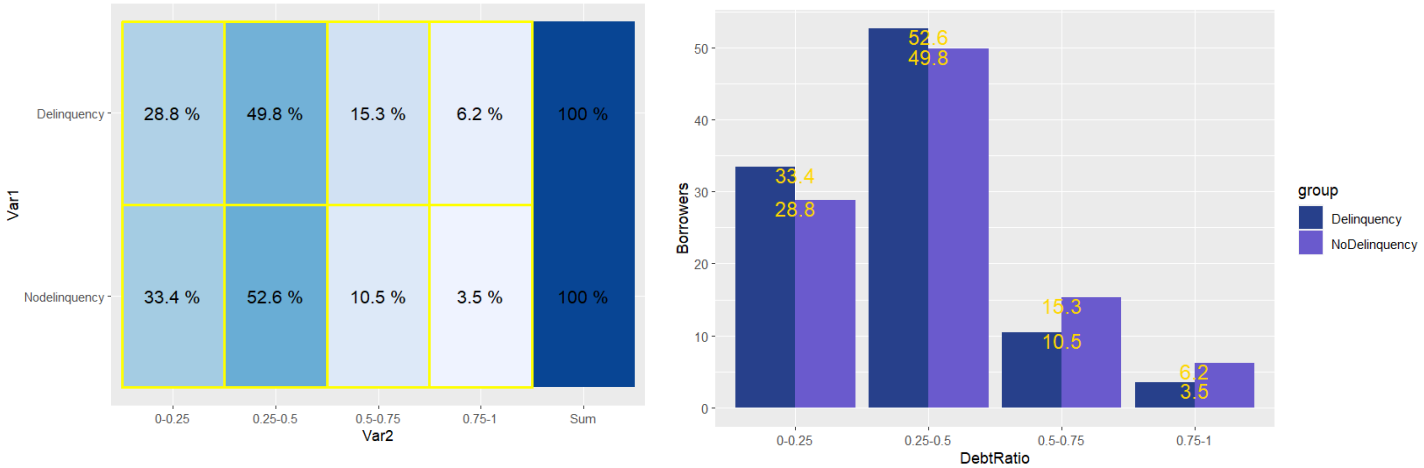
Graph 12 : Age X SeriousDlqin2yrs analysis



According to results below, on graph on right, histogram in blue represent borrowers without delinquency and those in gold, show borrowers who experienced delinquency. As we can see borrowers between **41-52 (27.7%)** and **53-64 (28.1%) years old**, are those who experienced 90 days past due delinquency or worse. That could be explained by the fact that borrowers from 41 to 64 years in their ambition to achieve their various projects (house, children school fees, preparing for retirement) has difficulty to repay their loans on time.

DebtRatio variable X SeriousDlqin2yrs

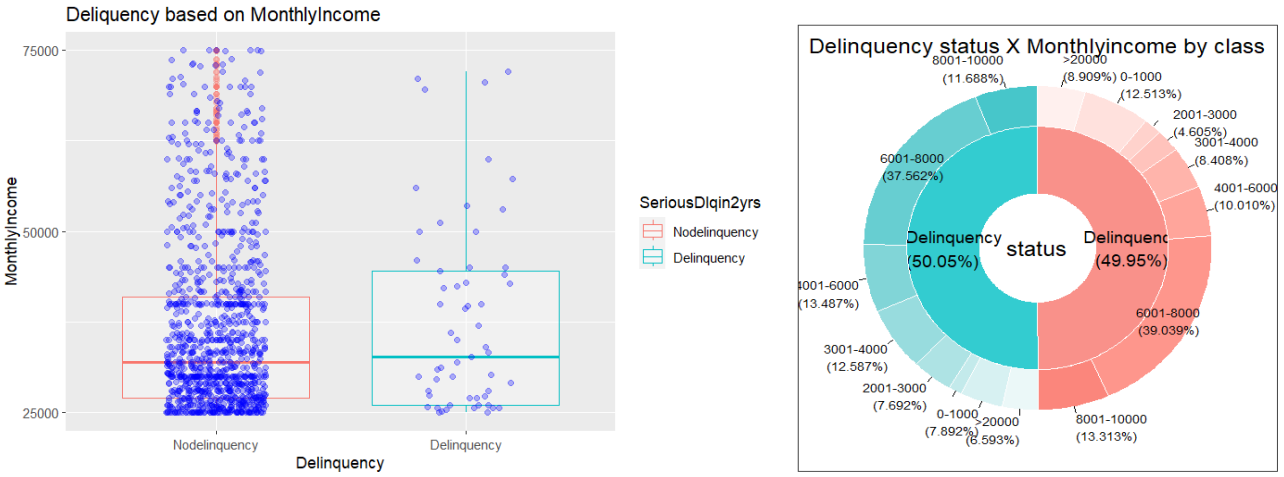
Graph 13 : DebtRatio X SeriousDlqin2yrs analysis



Borrowers with debt ratio between 0.25 and 0.5 are those who experienced 90 days past due delinquency or worse.

MonthlyIncome variable X SeriousDlqin2yrs

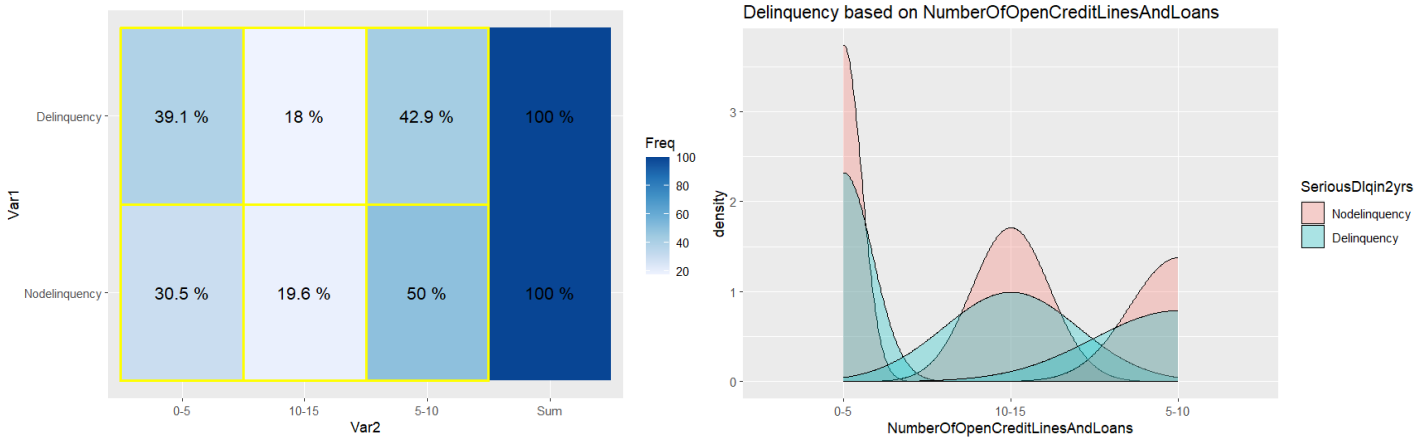
Graph 14 : Income X SeriousDlqin2yrs analysis



Borrowers with 6001 to 8000 (40%), 8001 to 10000 (13.31%) monetary unity, monthly income are those who experienced 90 days past due delinquency or worse. While, borrowers with 2001 to 3000 monetary unity experienced less 90 days past due delinquency or worse. This can be explained by the fact that civil servants who earn these salaries have the ease of borrowing large sums of money from the bank because they rely on their high salaries while those who earn less than them can only borrow a limited amount. and they are inclined to pay back as soon as possible.

NumberOfOpenCreditLinesAndLoans X SeriousDlqin2yrs

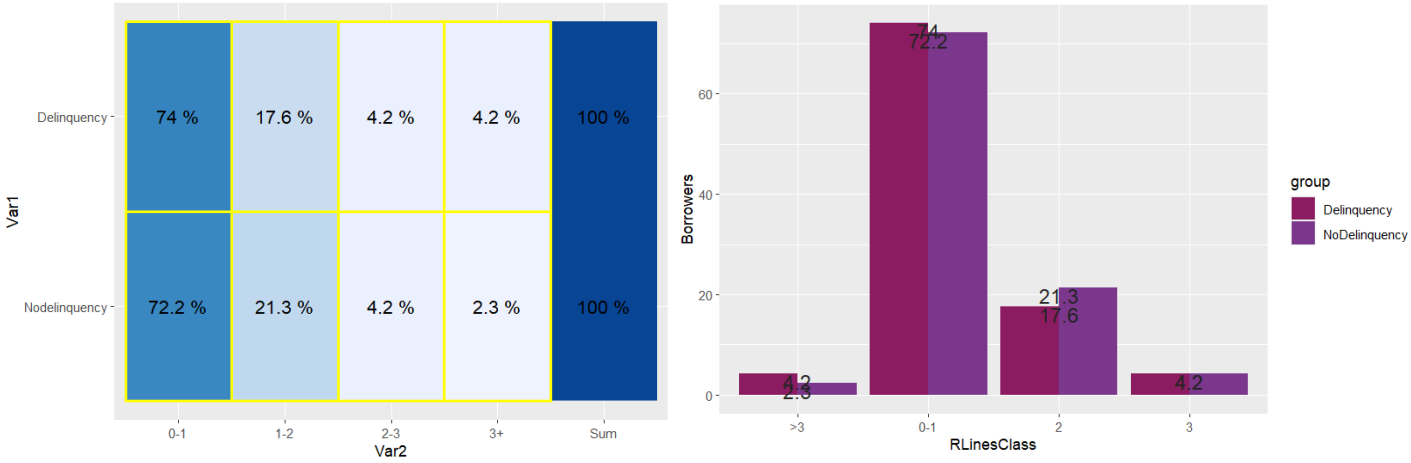
Graph 15 : NumberOfOpenCreditLinesAndLoans X SeriousDlqin2yrs analysis



Borrowers with 5 to 10 open loans such as car loans, house loans and lines of credit (ex. Credit card) experienced 90 days past due delinquency or worse. We can conclude that more the number of open loans increase, more the borrowers chances to have delinquency increases.

NumberRealEstateLoansOrLines and SeriousDlqin2yrs(delinquency)

Graph 16 : NumberRealEstateLoansOrLines X SeriousDlqin2yrs analysis



Borrowers with none or 1 mortgage and real estate loans including home equity lines of credit an individual have taken experienced more 90 days past due delinquency or worse.

Descriptive statistic conclusion

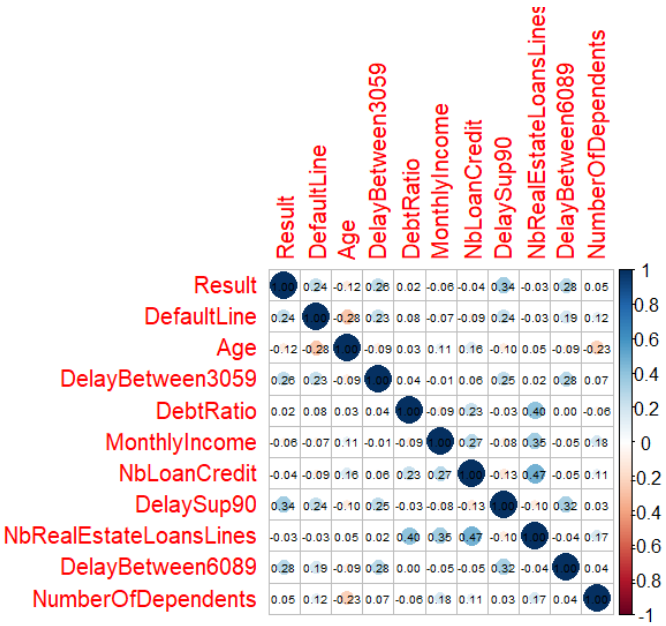
In conclusion with this statistic analysis, we now know that borrowers who experienced more 90 days past due delinquency or worse are those who are 41 to 64 years old. They are debt ratio is between 0.25 and 0.5. They own between 6000 to 10000 as monthly income, and have 5 to 10 open loans such as car loans, house loans and lines of credit, with none or 1 mortgage and real estate loans including home equity lines of credit an individual.

Multivariate analysis : spearman correlation matrix

Before building any complex model we need to have an overview of the data relation to see how our different variables interact together. Correlation finds the trends shared between two variables. When with the increase in value of one variable is associated with the increase in values of another variable, then these two variables are said to be positively correlated. On the other hand, when with the increase in value of one variable is associated with decrease in

value of another variable then these two variables are said to be negatively correlated. If the correlation is Zero, then that means there is no pattern visible between two variables. The total selected features are shown in Table-2 with their correlation percentage:

Table 2 : feature correlation

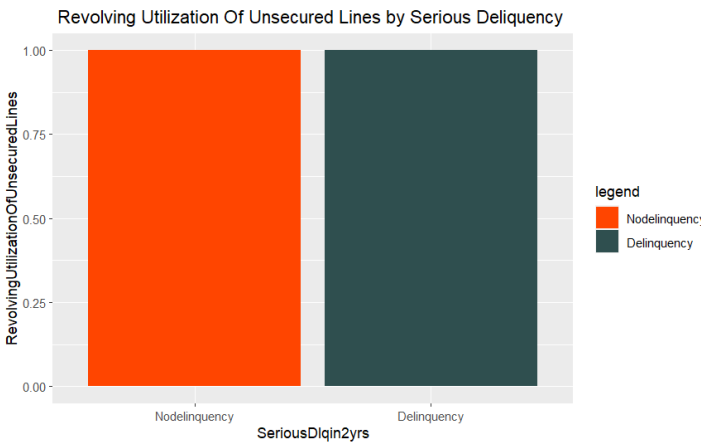


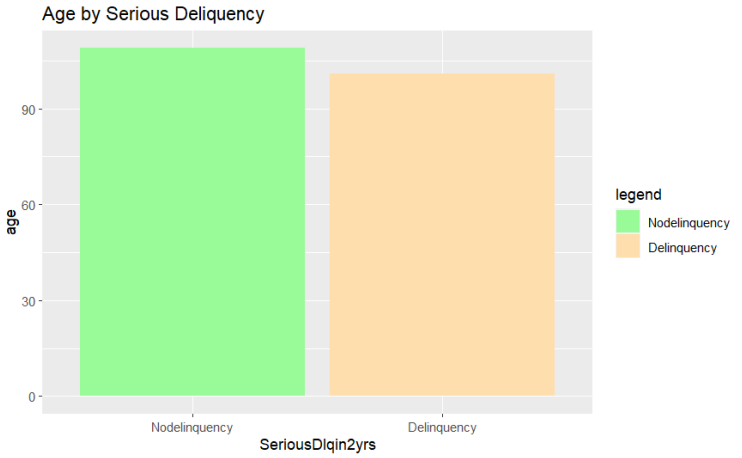
There is a positive relationship between “SeriousDlqin2yrs” and dependent variables like RevolvingUtilizationOfUnsecuredLines (0.24), NumberOfTime30.59DaysPastDueNotWorse (0.26), DebtRatio (0.02), NumberOfTimes90DaysLate (0.34), NumberOfTime60.89DaysPastDueNotWorse (0.28) and NumberOfDependents (0.05). it means the probability to experienced 90 days past due delinquency or worse increase when those variables increase while an increasing in age, MonthlyIncome, NumberOfOpenCreditLinesAndLoans, NumberRealEstateLoansOrLines have the opposite impact on SeriousDlqin2yrs. An increase in those variables, decrease respectively from 0.12, 0.08,0.04, and 0.03 the probability to experienced 90 days past due delinquency or worse.

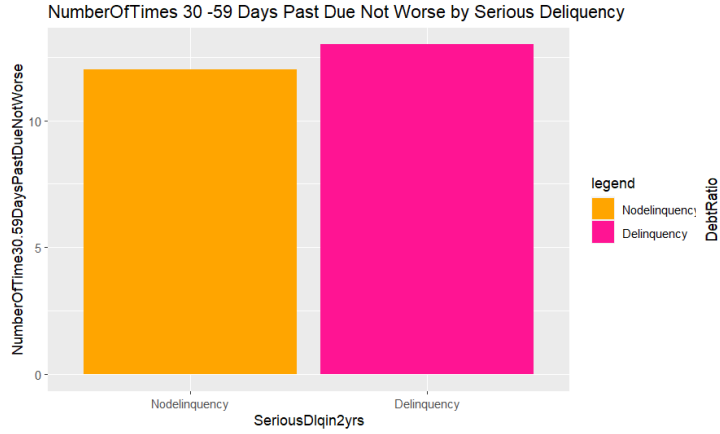
Multivariate analysis conclusion

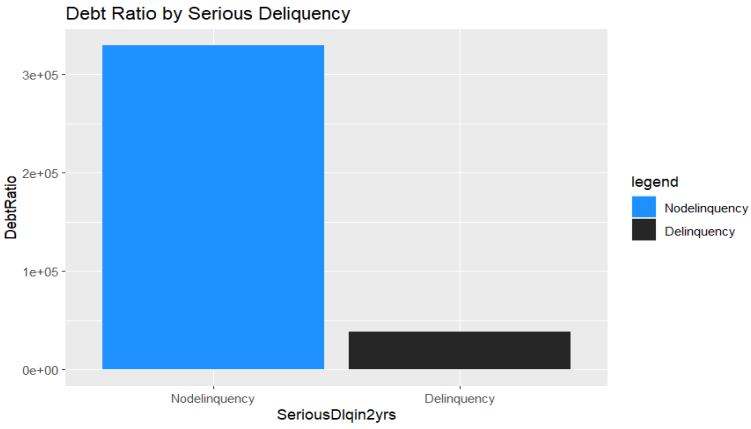
Explain variable (SeriousDlqin2yrs) and explanatory variables (RevolvingUtilizationOfUnsecuredLines, NumberOfTime30.59DaysPastDueNotWorse, DebtRatio, NumberOfTimes90DaysLate, NumberOfTime60.89DaysPastDueNotWorse and NumberOfDependents) are positively correlated while SeriousDlqin2yrs and MonthlyIncome.

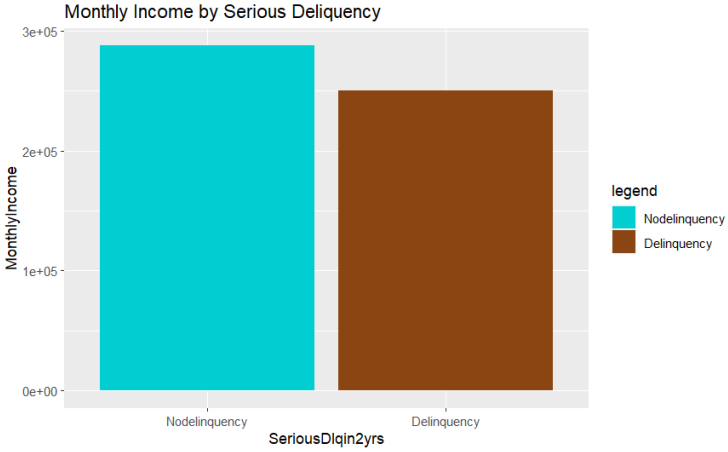
Groups mean t-test

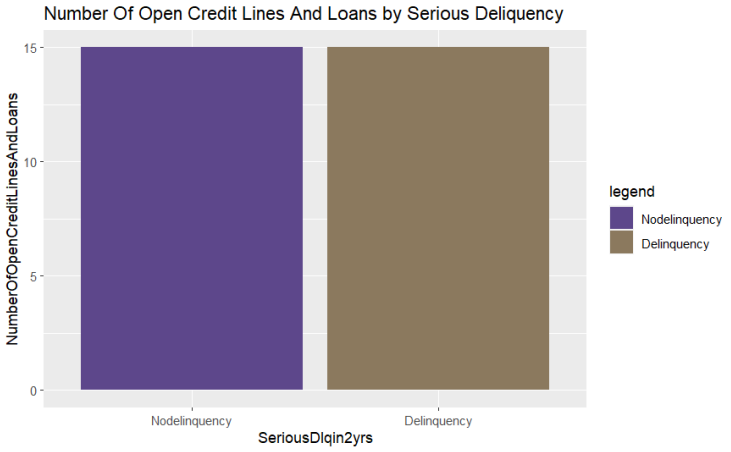


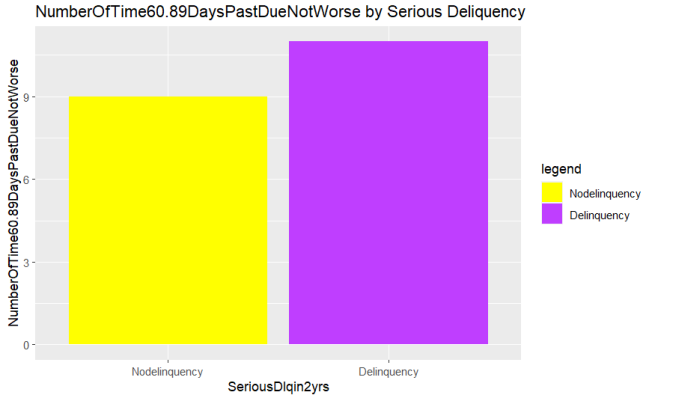


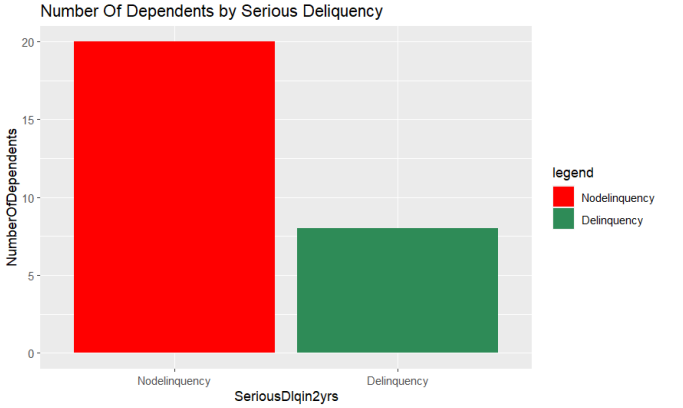












From the bar plot, there were an equivalent number of individuals who were seriously delinquent according to Revolving Utilization Of Unsecured Lines, Monthly Income, Number Of Open Credit Lines And Loans ,Debt Ratio, Number of Dependents and age except NumberOfTimes 30 -59 Days Past Due Not Worse where more number of individuals were Seriously delinquent than were not delinquent according to NumberOfTimes 30 -59 Days Past Due Not Worse.

```
        Welch Two Sample t-test

data:  RevolvingUtilizationOfUnsecuredLines by SeriousDlqin2yrs
t = -78.462, df = 11156, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Nodelinquency and group
Delinquency is not equal to 0
95 percent confidence interval:
 -0.3043552 -0.2895187
sample estimates:
mean in group Nodelinquency   mean in group Delinquency
                  0.2805296                   0.5774666
```

According to the t-test, there is a significant difference within the mean of the groups of those variables.

4. Modelling

In this chapter, the parameters of the different machine learning algorithms will be determined. This is done in two sections, first the parameters will be determined for use with the first data set, then for use with the second data set. Note that all of the model training happens on the training set only. The testing set will only be used to determine the final performance.

Methodology

In this section, we study approaches for predicting qualitative responses (default or not), 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.

Machine Learning Method and Machine Learning Algorithms:

Logistic Regression

Logistic regression using in credit risk prediction started in the late 1960s. one of the first study was conducted by Ewert (1969) with classification accuracy of 82%. Logistic regression was replaced in early 1890s by logistic regression by Ohlson (1980) and Wiginton (1980), who concluded that it performs better than discriminant analysis. Similar findings have been also confirmed in the later studies such as in the study of Altman and Sabato 2013 in their study of default prediction for SME companies Cultrera's and Brédart's (2016) study of SMEs in Belgium, to name a few.

We perfume this model because it is considered as a superior model in statistical analysis, as many of the conceptual and computational challenges of linear regression, such as possibility of negative possibility and possibility with larger than one, can be taken into account in the model. It also has several advantages over discriminant analysis, normal distribution of the input variables is not required and therefore also qualitative variables can be included in the model(Hand & Henley, 1997, p. 533).

The first of the machine learning algorithms discussed, is Logistic Regression. It has been proposed by Cox (1958), making it one of the older machine learning algorithms. The primary idea of Logistic Regression is to use techniques developed for linear regression to model the probability of a sample belonging to a certain class. This is done using a linear predictor function, Equation 2.9, which is a linear combination of m feature values and m + 1 regression coefficients.

$$f(i) = \beta_0 + \sum_{i=1}^{m} \beta_i x_i$$

Logistic regression is different from other forms of regression due to the way the linear predictor is linked to the probability of a certain outcome. In linear regression, it is not possible to determine a closed form equation to determine the coefficients. Instead other methods like maximum likelihood estimation are used. In this method an iterative process is used during which in each iteration the coefficients are slightly changed to try to improve the maximum likelihood. In this research project two methods to fit the model are taken in to account. These methods will not be discussed in depth since it is not within the scope of this research project. The first method is Liblinear, it used a coordinate descent algorithm to find suitable values for the coefficients. The second method is saga which uses a stochastic average gradient descend. The second method usually is faster on large data sets.

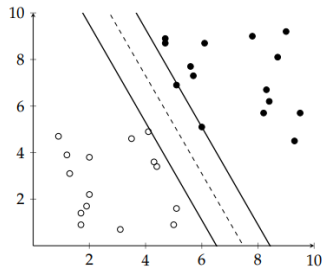
In the loss function that is minimized it is usual to include a regularization term. Such a term is to penalize complex models and favor models which are simpler. With Logistic Regression two types of regularization are commonly used, L1 and L2. The first of these is a regularization that favors sparse models, or models where a large fraction of the coefficients is zero. L2 is used as regularization term when a sparse model is not suitable. When the data set contains highly correlated features, L1 should be used as regularization term. It picks a single of the correlated features and sets the coefficient of the other features to zero. L2 would simply shrink the

coefficient of all correlated features. Usually a parameters is added to the algorithm which can be used to determine the strength of the regularization.

Support Vector Machine (SVM)

The next type of model discussed in this chapter is the Support Vector Machine, or SVM (Shmilovici et al., 2009)This model will only be discussed on a high level, it is one of the more complex models, the exact mathematical derivation falls not within the scope of this research project. SVM works by constructing a hyperplane in a high dimensional space. It is constructed in such a way that it splits the classes of the training samples in such a way that the distance to the closest sample of either class is maximized. The plane is used as a separation border for classifying new samples. A hyperplane is a plane whose dimension is one less than its ambient space. In other words, points in a three dimensional space are separated by a two dimensional plane. Points in a two dimensional space are separated by a one dimensional line. An example of the application of SVM is shown in Figure SVM1. As can be seen in the graph, the dashed line splits the classes in such a way that the distance to the closes sample of either class is maximized.

FIGURE 2.11: SVM generated maximum margin hyperplane in a two class two dimension problem.



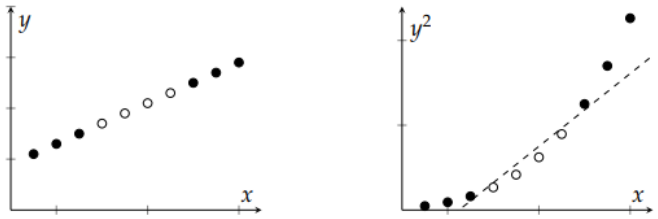
Source : Michiel Cornelissen (2018)

The above description clearly has a large flaw, what will happen when the samples are not linearly separable? Without deviating from linearity, the solution is to add a penalty for samples on the wrong side of the separation. This penalty is multiplied with a certain weight to determine the trade-off between the margin around the hyperplane and the samples on the wrong side. This weighted penalty is added to the hyperplane margin and finally minimized. Later a more

advanced method was created to separate samples using a nonlinear classifier. This method is achieved by applying the so called 'kernel trick'. The general idea of the method is to map the samples to a higher dimensional space. An example showing the benefit of this transformation is shown in Figure 2.12. By mapping data from a {x, y} space to three dimensions, {x, y, y 2}, it becomes possible to separate the two classes with a line. Each type of transformation is called a kernel, some examples are polynomial, Gaussian and hyperbolic.

FIGURE 2.12: Mapping data to a higher dimension to achieve linear separability.

(A) Two dimensional $\{x, y\}$ space. (B) Two dimensions of the $\{x, y, y^2\}$ space.



Source : Michiel Cornelissen (2018)

We choose this model because many studies have confirmed SVM as a successful method for default prediction. Study of Fan & Palamiswami (2000) concluded that SVM outperformed neural networks and linear discriminant classifiers, Moula et. al. (2017) showed that the SVM model is marginally superior to CART, Yu et al. 2008 reported that SVM performed slightly better than logistic regression, Shin et al. (2005) concluded that SVM performed better a back-propagation neural network and Chen (2011) found that in comparison to other prediction models, the SVM had high accuracy and performed well for both, short and long-term predictions.

Artificial Neural networks

An Artificial Neural Network is a network of interconnected neurons. This section is based on (Zhang, 2010). Each neuron receives an input, processes these signals and produces an output signal. A Neural Network consists of several layers of neurons. The leftmost layer consists of the input neurons, the rightmost neurons are the output neurons. The layers in between are

called hidden layers. The neurons are connected in such a way that the output of one neuron is the input of all neurons in the next layer. The exceptions are the input and output neurons. Input neurons have no predecessor and are used as input for the network. Output neurons have no successor and function as network output. In classification problems with two classes only a single output neuron is used. Depending on the value of the output neuron the sample is assigned to either class. When more than two classes exist, each class has a corresponding output neuron and the sample is assigned to the output neuron with the highest value.

Neural networks became popular in the academic research in 1990's. Neural networks are designed to emulate the process of the human brain. Neural networks can be described as multi-stage information processing. where at each stage the hidden correlations of the predictor variables are identified. The complexity of the model makes it extremely difficult to interpret and the model can therefore be considered as a black box model. (Anandarajan et aI., 2004).

We choose to perform this model because this more complex model has been utilized to improve the prediction accuracy of credit risk modelling, a comparison of the prediction accuracy in the academic research reveal, that the average overall classification performance of neural network models is similar to logistic models (Aziz & Dar, 2006).

Random Forest

Another popular ensemble model is the random forests classifier. The basic principle of this classifier is to train multiple Decision Trees and have those together make a classification (Breiman, 2001). Each of those trees is trained on a subset of the training data drawn with replacement. The training procedure is similar to how a normal Decision Tree is trained except for one difference. At each split in the tree a random selection of features is selected, from which the feature for the split is selected. Usually the square root of the number of available features is used for how many features have to be drawn (Hastie, Tibshirani, and Friedman, 2009). The reason for this random feature selection is to decrease the correlation between the individual trees. Given a feature set X = xi , ..., xn and corresponding labels Y = yi , ..., yn, for each tree in the random forest a random subset Xr and Yr is drawn with replacement. To each of the sets of random samples a Decision Tree is fitted. At each split in the tree a random subset of features is selected on which the split can be based. For a classification with p features the most used number of features considered for a split is $\sqrt{p}$ or log2 (p). This tree construction process results in N separate Decision Trees which are combined in a single classifier. This can be done either by letting each classifier cast a vote or by averaging the probabilistic predictions. Michiel Cornelissen (2018).

Decision Tree

A Decision Tree probably is one of the best known classifiers due to its logic structure. Decision Tree classifiers are extensively described by Rokach and Maimon (2009). A Decision Tree consists of connected nodes which form a rooted tree, meaning that the tree has a single root node as starting point. All following nodes have a single incoming edge, if the node also has outgoing edges it is called an internal node. Each of the internal nodes splits the data set according to a certain logic. In classification this split is usually based on the value of a certain feature. Nodes that do have incoming edges but no outgoing edged are called leaves. Leaves are assigned to a label based on which label is most appropriate. After a tree has been

constructed, the classification is done by starting at the root node and following through the internal nodes until a leave has been reached. Decision trees utilize hierarchical decision-making process, which is similar to human behavior in real-life decision-making process and can be used for both, classification and regression. One key advantage of decision tree is that trees can also interpret categorical data. (Joshi 2020, 53-63). Classification & Recession tree (CART), also known as Recursive Partitioning Algorithm, is a data mining method, that utilizes decision trees in classification. Study by Frydman et al. (1985) concluded, that CART outperformed MDA in most sample and holdout comparisons and is good at giving additional, easily interpreted information of the relationship of the predictor variables.

Cross Validation

Naive Bayes Classifier

Naïve Bayes Classifier is one of the most popular and powerful algorithm for classification task. If a dataset has millions of instance with many attributes, then the Naïve Bayes classifier is the suggested one. The foundation of Naïve Bayes classifier is Bayes theorem. Bayes Theorem works on conditional probability. A conditional probability is something like, probability of an event (A), given that another (B) has already occurred [12]. The assumption of Naïve Bayes classification on Bayesian probability called class condition independence.

Naïve Bayes classifier predicts the probability of each instances of a class, and the class with highest probability is counted as most likely class, the process of determining the class with highest probability is called Maximum A posteriori (MAP). Predicting credit score is a classification problem and its need Gaussian Naïve Bayes Classification, Gaussian Naïve Bayes gave powerful output in classification.

Gradient Boosting Machines (GBM)

The Gradient Boosted Trees Operator trains a model by iteratively improving a single tree model. After each iteration step the Examples are reweighted based on their previous prediction.

The final model is a weighted sum of all created models. Training parameters are optimized based on the gradient of the function described by the errors made.

We choose it because XGBoost is a fast implementation of Gradient Boosting, which has the advantages of fast speed and high accuracy. For classification, XGBoost combines the principles of decision trees and logistic regression, so that the output of our XGBoost model is a number between 0 and 1. For the remainder of the paper we refer to XGBoost as GBT.

Bagging & Boosting

There are two paradigms of ensemble methods, that is, sequential ensemble methods, wh learners are generated sequentially, with Boosting as a representative, and parallel ensemble methods where the base learners are generated in parallel, with Bagging as a representative. The basic motivation of sequential methods is to exploit the overall performance can be boosted in a residual of parallel ensemble methods is to exploit the independence between the base learners, since the error can be reduced dramatically by combining independent base learners.

Boosting

Boosting refers to boosting performance of weak models. It involves the first algorithm is trained on the entire training data and the subsequent algorithms are built by fitting the thus giving higher weight to those observations that were poorly predicted by the previous model. The general boosting procedure is quite simple. Suppose the weak learner will work on any data distribution it is given and take the binary classification task as an example; that is, we are trying to classify instances as positive and negative. The training instances in space X are drawn i.i.d. from distribution D, and the ground-truth function is X2 and X3, each takes 1/3 amount of the distribution, and a learner working by random guess has 50% classification error on this problem. We want to get an accurate (e.g., zero error) classifier on the problem, but we are unlucky and only have a weak classifier at hand, which only has correct classifications in spaces

X1 and X2 and has wrong classifications in X3, thus has 1/3 classification error. Let's denote this weak classifier as h1. It is obvious that h1 is not desired.

Bagging

It is also called Bootstrap Aggregating. In this algorithm, it creates multiple models using the same algorithm but with random sub randomly with random with replacement sampling technique (i.e. bootstrapping). This sampling method simply means some observations appear more than once while sampling. After fitting several models on different samples, these models are aggregated by using their average, weighted average or a voting method. The bagging and boosting algorithms are suitable means to increase efficiency of algorithms, however, the loss of simplicity of this classification sche disadvantage (Machova, Puszta, Barcak, & Bednar, 2006).

Performance measure

AUC & ROC

The Receiver Operating Characteristic curve (ROC) is the most widely used metric to depict the discriminatory power of the classification model. This metric was selected as performance metric in this study, as it does not rely on threshold settings, can be used to evaluate a model with probability output and works with imbalanced classifications. (Gong, 2021). ROC represents the relation between true positive and false positive predictions of the models at different thresholds. The ROC curve is constructed by plotting the fraction of the false positive predictions on the x-axis against the fraction of the true positive predictions on the y-axis. (Kotu & Deshpande, 2014)) Figure 3 illustrates various ROC curves. The point in top left corner represents perfect model, where all cases all cases are classified perfectly, and dotted line represents a model with no prediction power. For interpretation purposes the area under the ROC curve can be calculated (ROC AUC). A model with perfect classification has the ROC AUC of 1 and model with no discriminatory power would have an ROC AUC of 0,5. (Kotu &

Deshpande, 2014) It should be noted that ROC AUC does not take into account the shape of the ROC curve. In Figure 3, the curves A and B represent ROC curves of two different models with the same AUC measure. The steeper curve A, can be considered better in default prediction, as it allows using higher cut off to avoid false positives (predicted to default, but did not) with smaller impact of misclassifying true positives (predicted to not default, but defaulted). This is important, as misclassifying non-defaulters as defaulters, can be costly as it results in a loss of sales.

Precision-Recall curve

Precision-recall curves (PR curves) represents the relation between mode's precision and recall at different classification thresholds. Precision-recall curves are recommended for highly imbalanced datasets and can give more informative picture of models' performance than ROC curves. Precision is the ratio of correct positive predictions to the total positive predictions and recall is the ratio of correct positive predictions of all of the total positives in the dataset. (Davis & Goadrich, 2006). High precision is favourable when there is high cost for false alarms, and high recall is beneficial if there is low cost for false alarms and all potential positive cases need to be identified. The formulas for Precision and recall are presented below (Davis & Goadrich, 2006): PR curves show the relation between model's precision and recall at different classification thresholds. A classifier without discriminatory power would be a horizontal line proportional to the number of positive samples in the dataset. Similarly, to ROC curve, the PR AUC can be used as a metric in model evaluation (Tharwat, 2020).

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} = \frac{TP}{P}$$

Confusion matrix

The first method to analyze the performance of a classification algorithm is by using a confusion matrix. It is a method to visualize the accuracy using a table. For the purpose of explaining the confusion matrix a classifier that classifies instances as positive or negative is assumed. Four

fields have to be calculated to fill in the matrix, true positive, false positive, true negative and false negative. These fields are simple to calculate. For example, true positive is the number of occurrences correctly classified as positive and false negative are the occurrences incorrectly classified as negative. For the following example, assume a classifier that classifies every instance as positive. This would result in 100% of the actual positive instances being classified as positive. If the accuracy would be purely evaluated on the actual positive instances being classified as positive such an algorithm appears to be performing perfect. To prevent this situation, the confusion matrix can be used. Since the column with the negative predictions is empty, it is clear that the algorithm is not working.

Accuracy

Using the values calculated for the confusion matrix it is possible to calculate several interesting statistical measures. The first of which is the accuracy, shown in (2.5). The abbreviations used in this section are identical to those given in 2.1 The accuracy is used to calculate the fraction of total predictions that is correctly classified. A random classifier will get on average half of the classifications correctly. Values above 0.5 indicate the model has a higher accuracy as random guessing. A perfect prediction has accuracy 1.0.

$$Accuracy = \frac{TP + TN}{P + N}$$

F1- score :

As mentioned in the previous section, precision and recall are often combined. One of the statistics resulting from such a combination is the F1-score. As can be seen in Equation 2.8, the F1-score is equal to the harmonic mean of the recall and precision. A disadvantage of the F1-score is that it does not take the true negatives into account.

$$F1 - Score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

5. Empirical results & interpretations

Logistic Regression

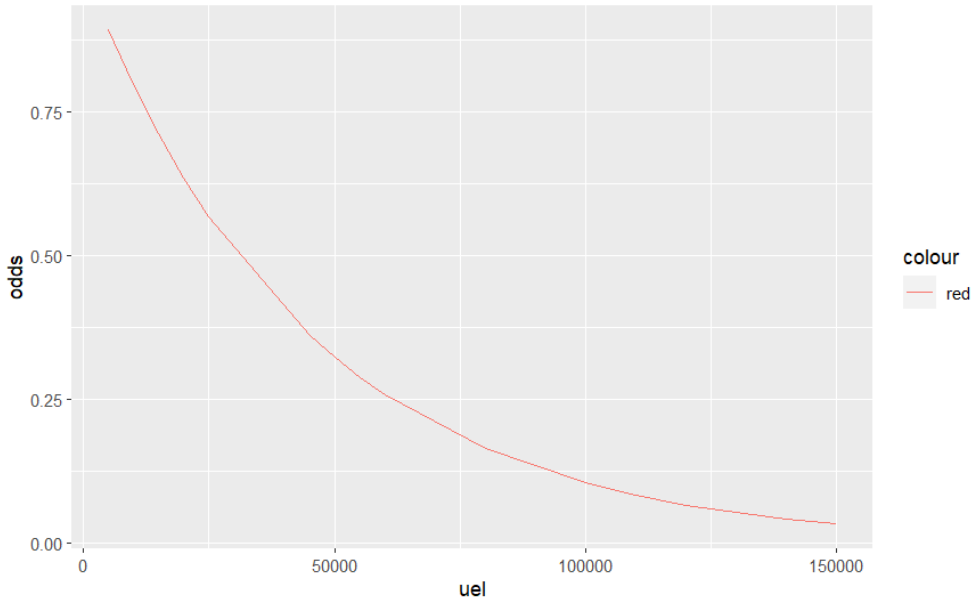
From the summary of our model logistic model, we can see that, judging by the significance parameter asterisk, "RevolvingUtilizationOfUnsecuredLinesis" an insignificant variable.

Interpretation of logistic regression

The coefficient of fico of -0.0123, we interpret that for a unit (100) increase in fico, the log odds in favor of a person default is decrease by 0.0123 units. However, this interpretation is not quite hard to digest. Another way is to take the antilog of the logit. Remember earlier we define :

$$\frac{p_i}{1-p_i} = \exp(\beta_1 + \beta_2 X_i)$$

so we know that the odds is e−0.00004.047∗50000=0.8356e−0.00004.047∗5000=0.817 which we can conclude that for people with income of $50,000 per year, on average 17 out of 20 person are likely to default their loan. Here I visualize the simulation of how the odds for default is decreasing as the income increasing. We can clearly see that the odds is decreasing exponentially as the income increasing. It might be better to not just interpret a single *Xi* but simulate for any values and visualize the effects to the odds.



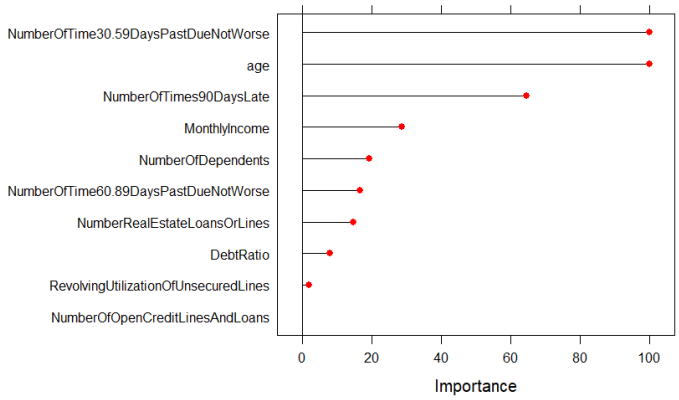
Graph above is the visualization of the effects of increasing the monthly income to the odds of default. It shows that as monthly income increasing, the odds of default is decreasing exponentially. This is an intuitive result as we expect people with higher income will likely to pay their loan, vice versa.

$$log_e \frac{P(Target = 2(Delinquency))}{P(Target = 0(Nodelinquency))}$$

= -1.319 - 0.00003680 X1 -0.002842 X2 + 0.5034X3 -0.00002753X4 -0.007680X5 + 0.4681X6 + 0.006833X7 – 0.9395X8 + 9.552X9.

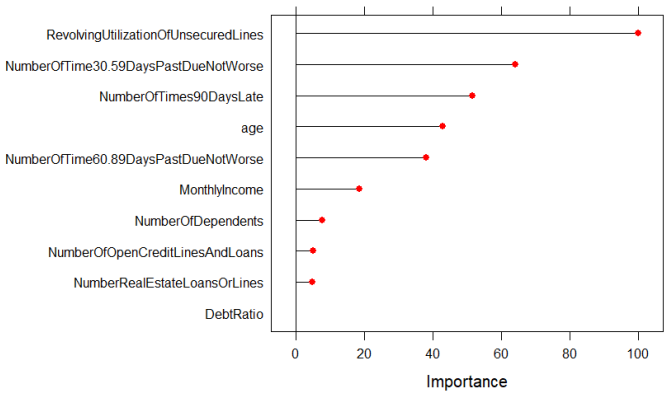
In the model above, *c* = -1.319 , and *p* = *P{Y=1}* is the probability of experiencing delinquency. Let's pick *NumberOfTime30.59DaysPastDueNotWorse* and see how it impacts the chances of experiencing delinquency. Increasing the *NumberOfTime30.59DaysPastDueNotWorse* by 1 unit will result in a 0.5034 increase in *logit(p)* or *log(p/1-p)*. Now, if *log(p/1–p)* increases by 0.13, that means that *p/(1 — p)* will increase by *exp(*0.5034*)* = *1.65*. This is a 65% increase in the odds of experiencing delinquency (*assuming that the variable RevolvingUtilizationOfUnsecuredLines remains fixed*).

Variable importance



According to our logistic model, first, NumberOfTime30.59DaysPastDueNotWorse, second age and finally NumberOfTimes90DaysLate are variables which impact seriously the fact that a borrower will experienced delinquency.

Support Vector Machine (SVM)

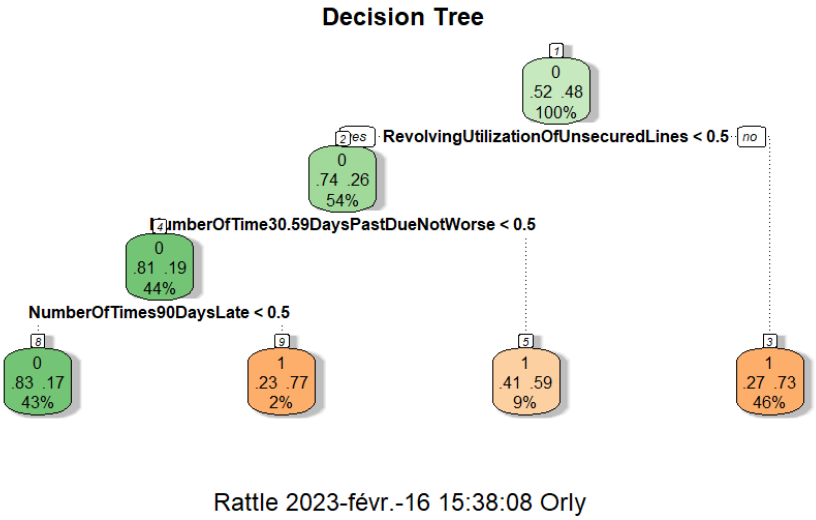


Our support vector machine (SVM), show us that RevolvingUtilizationOfUnsecuredLines, NumberOfTime30.59DaysPastDueNotWorse, NumberOfTimes90DaysLate and age are those variables wich have great importance in predicting the probability of default.

Bagging

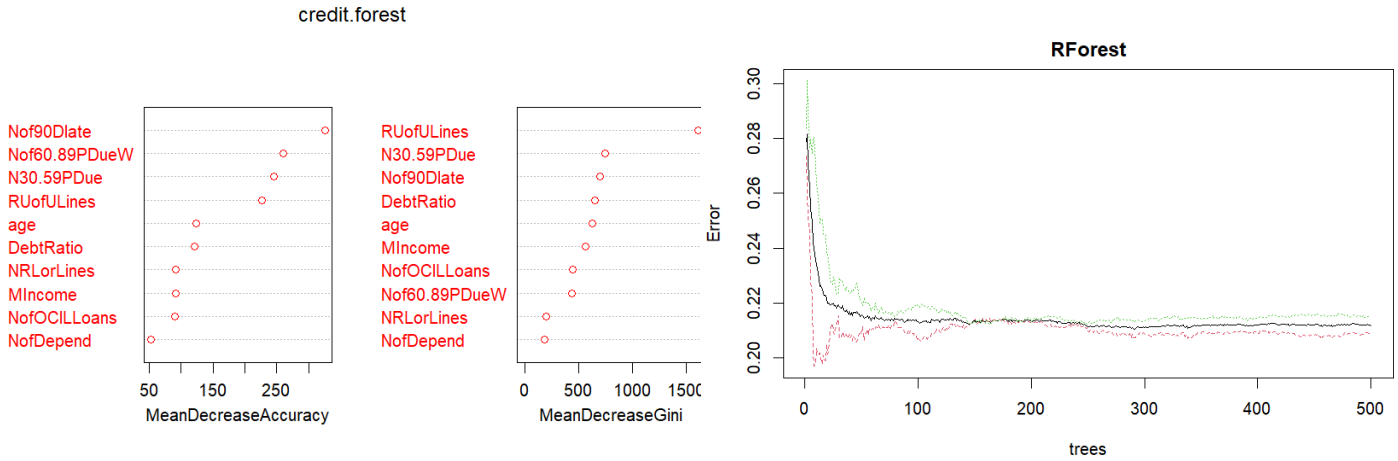
The resulting tree presented in Figure can be considered simple and comprehensive. Decision trees can help to interpret the underlying relationships between the variables. From the tree we can learn that RevolvingUtilizationOfUnsecuredLines and Number Of Time30.59 Days Past Due Not Worse have a relationship: longer balance on credit cards and personal lines of credit except real estate and no instalment debt like car loans result in higher default risk even with number of times borrower has been 30-59 days past due but no worse in the last 2 years. This of course logical, as indebted companies who have difficulties to pay invoices clearly have solvency challenges.

Figure 4: A Decision Tree model for the loan default problem



Random Forest

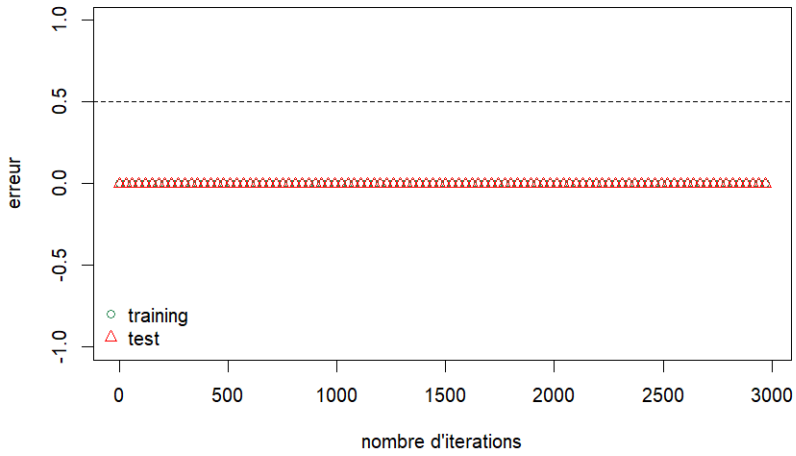
Random forest model was built using MASS r-package's rf method. The hyper parameter optimization of mtrees was tested with values between 1 to 15, and ROC AUC was used to select the optimal hyperparameter. The final value used for the model was mtry = 6.



Random forest (RF), show us that NumberOfTimes90DaysLate, NumberOfTime30.59DaysPastDueNotWorse, and age are those variables wich have great importance in predicting the probability of default.

Boosting

Now, we propose to compare bagging to boosting using as base predictor strains which have the advantage of being poorly tuned algorithmically. Boosting can be done, for example, using the R package gbm or ada. We begin by inspecting the influence of the number of iterations of the procedure on the error by examining the training sample or the test sample.



The figure above clearly illustrates the problem of over-fitting: the training error decreases continuously while the test error approaches 0.3 corresponding to the proportion of bad payers in the data set. Thus, if B is too high, the performance of the bagged predictor approaches that of a random draw. The performance metrics of the AdaBoost for both datasets are presented below in next chapter.

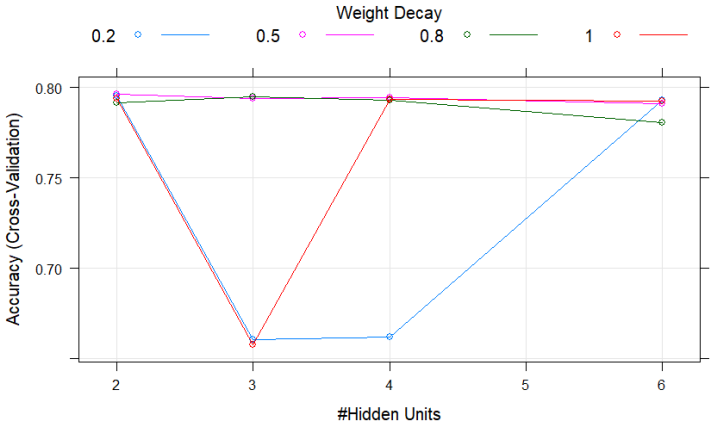
Naïve Bayes

The Naive Bayes classifier has no parameters which can be analyzed. The only parameters is the type of distribution, this can be Gaussian, Multinomial and Bernoulli. However, Bernoulli requires binary features and is thus not applicable. Multinomial is only suitable for non-negative features which makes it not usable for this data set. Therefore the conclusion is that only Gaussian can be used.

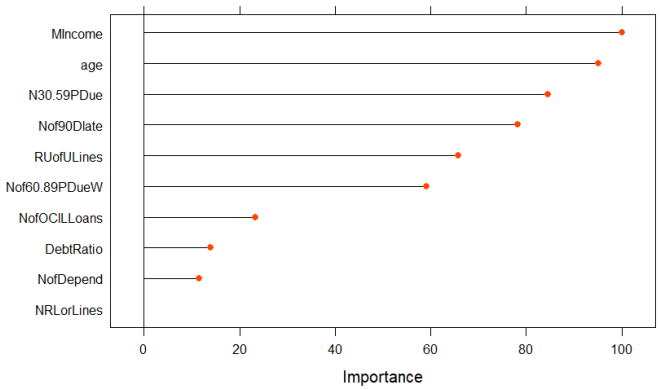
We will use the naive Bayesian algorithm which is one of the simplest methods in supervised learning based on Bayes' theorem. It is little used in relation to decision trees or logistic regressions but it is easy to estimate parameters and it is fast.

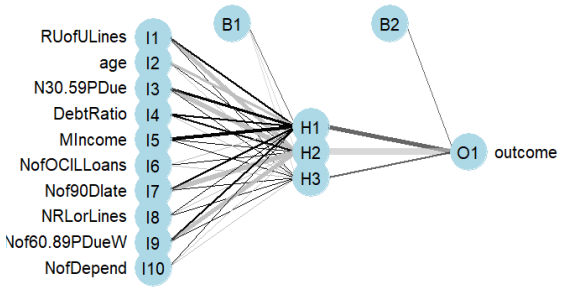
Neural networks

Neural Network was built using r-package nnet with nnet method. The hyperparameter size was tested with values between 1 to 10 and decay between 0,1 to 0,5. Disabling the over-sampling improved model performance significantly and the ROC-AUC with the hold out test sample improved from 85,6 to 92,7 (in next chapter). Therefore the selected model was trained without over-sampling. ROC was used to select the optimal model using the largest value. The optimal model performance was reached with hyperparameter size set to 8 and decay set to 0.5.



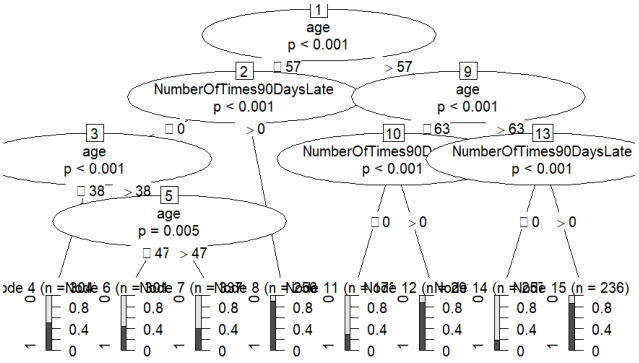
Variable importance in the Neural Network model presented in Figure. The model found statistical significance in 10 predictors and Number Of Time 30.59 Days Past Due Not Worse had the highest relative significance.

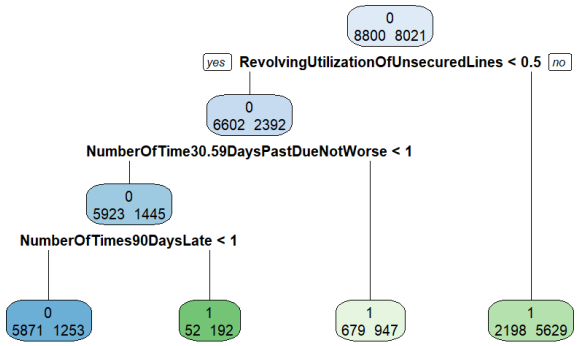




Decision tree CART

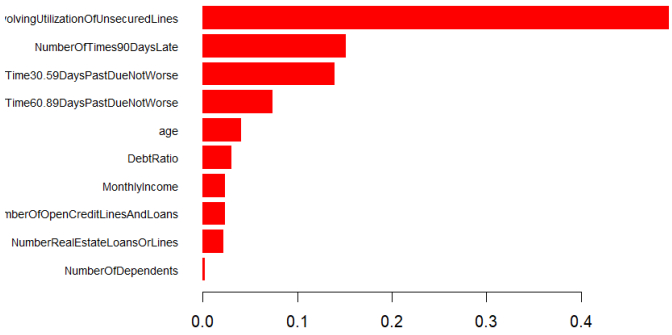
Classification & Recession tree (CART) was built using r-package rpart. ROC value was used to select the optimal pruning hyperparameter called cp. The complexity parameter (cp) in rpart controls the minimum improvement needed at each node of the model. The best performance was reached with cp of 0,01.





XG Boost

The performance metrics for evaluating prediction power for XGBoost were the ROC curve and confusion matrix with the threshold value obtained from the Youden's index. The results on the test set are presented below.



RevolvingUtilizationOfUnsecuredLines, NumberOfTime30.59DaysPastDueNotWorse, NumberOfTimes90DaysLate, NumberOfTime60.89DaysPastDueNotWorse have great importance in predicting the probability of default.

6. Performance measure

6.1. Confusion Matrix

Confusion Matrix is an additional measure of predictability performance. A confusion matrix is a table which is often used to describe the performance of a classification model on set of test data, whose true value is known. We can easily calculate algorithm's accuracy for a particular test data through confusion matrix. Here:

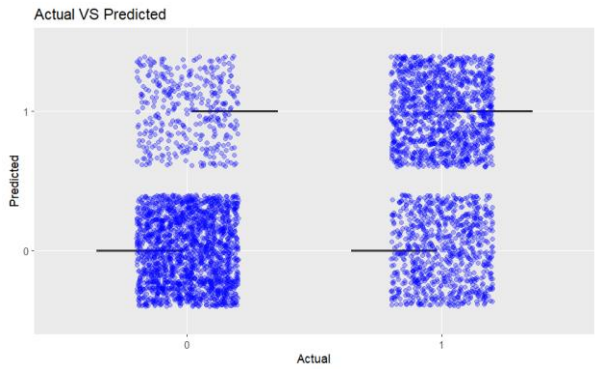
- True Positive: If a default is correctly classified and also labelled as a default then it is said to be 'True positive'.
- False Positive: if a non-default is wrongly classified and is labelled as a default then it is termed as 'False Positive'.
- True Negative: If a non-default is correctly classified and also labelled as a non-default then it is said to be 'True negative'.
- False Negative: if a default is wrongly classified and is labelled as a non-default then it is termed as 'False negative'.

Table : Confusion matrix

	0 = positive	1 = negative
0 = positive	True Positive	False positive
1 = negative	False negative	True negative

Sometimes accuracy can be decisive, to ensure the accuracy of a model we use confusion matrix. Here basically we test the model's classification results against the actual observed classification.

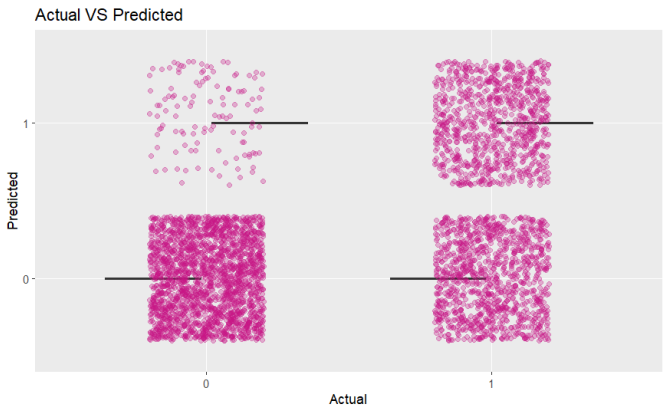
Logistic Regression



	0	1
0	1742	687
1	458	1659

The result above tells us that, regarding to the confusion matrix, our model correctly predict off the test dataset that 1830 individuals (true positive) paid their loan while 1742 people default (true negative). The overall accuracy is calculated by $\frac{\text{correctprediction}}{\text{totalobservations}}$ or equal to $\frac{1742+1659}{1742+1659+687+458}$ wich results 0.7481. That means that given data points of 4205 observations from our test dataset, our model has correctly predict 3016 outcome. But the confusion matrix suggest that our model has false negative of 370 data, which means that our model predict 370 persons will default but they actually paid the loan, and 819false positive which our model predict not default but they actually default.

Support Vector Machine (SVM)

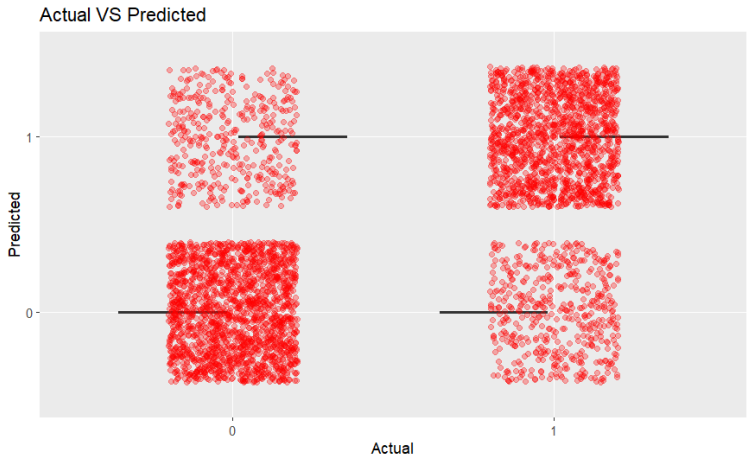


	0	1
0	2081	1110
1	119	895

The SVM confusion matrix suggest that our model has false negative of 119 data, which means that our model predict 119 borrowers will default but they actually paid the loan, and 1110 false

positive which our model predict not default but they actually default. The SVM overall accuracy is 0.7077.

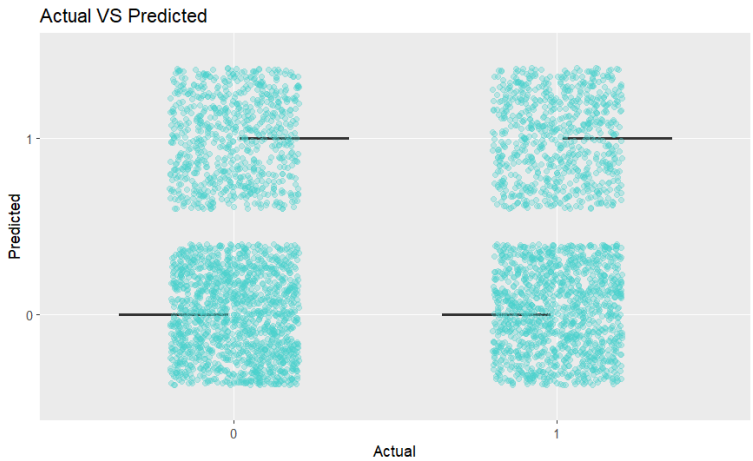
Random Forest



	0	1
0	1789	497
1	402	1508

On the other side, our random forest model correctly predict off the test dataset that 1789 individuals (true positive) paid their loan while 1508 people default (true negative). The SVM confusion matrix suggest that our model has false negative of 402 data, which means that our model predict 402 borrowers will default but they actually paid the loan, and 497 false positive which our model predict not default but they actually default. The SVM overall accuracy is 0.7862.

Boosting

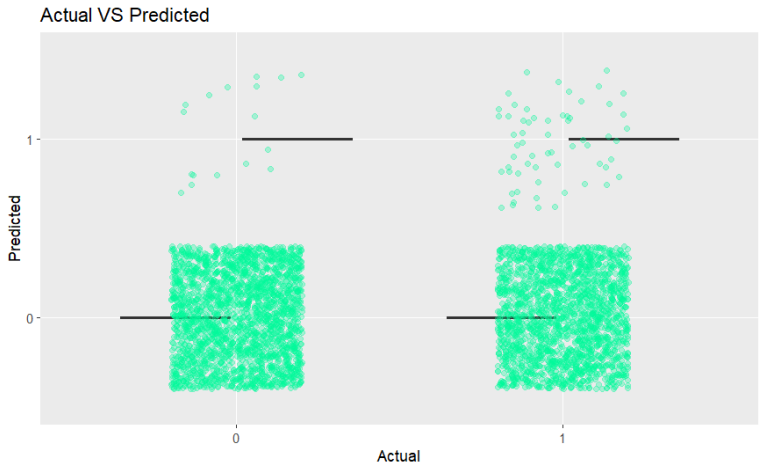


	0	1
0	1710	940
1	490	1065

The Boosting overall accuracy is 0.65992. The Boosting model correctly predict off the test dataset that 1710 individuals (true positive) paid their loan while 1065 people default (true

negative). The SVM confusion matrix suggest that our model has false negative of 490 data, which means that our model predict 490 borrowers will default but they actually paid the loan, and 940 false positive which our model predict not default but they actually default.

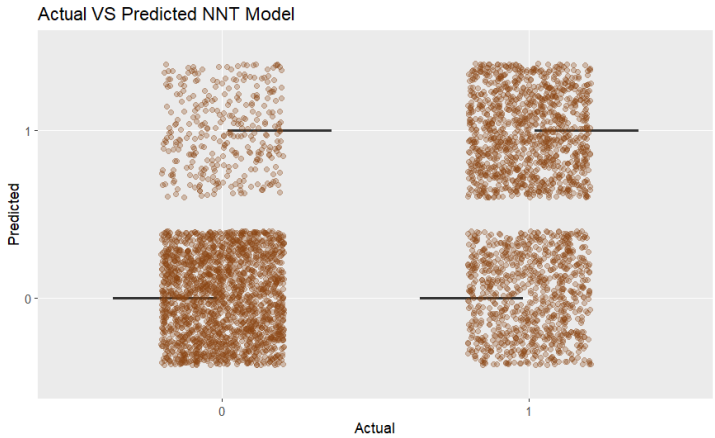
Naïve Bayes



	0	1
0	2183	1944
1	17	61

Naïve Bayes overall accuracy is 0.5337. The Naïve Bayes model correctly predict off the test dataset that 2183 individuals (true positive) paid their loan while 61 people default (true negative). The SVM confusion matrix suggest that our model has false negative of 17 data, which means that our model predict 17 borrowers will default but they actually paid the loan, and 1944 false positive which our model predict not default but they actually default.

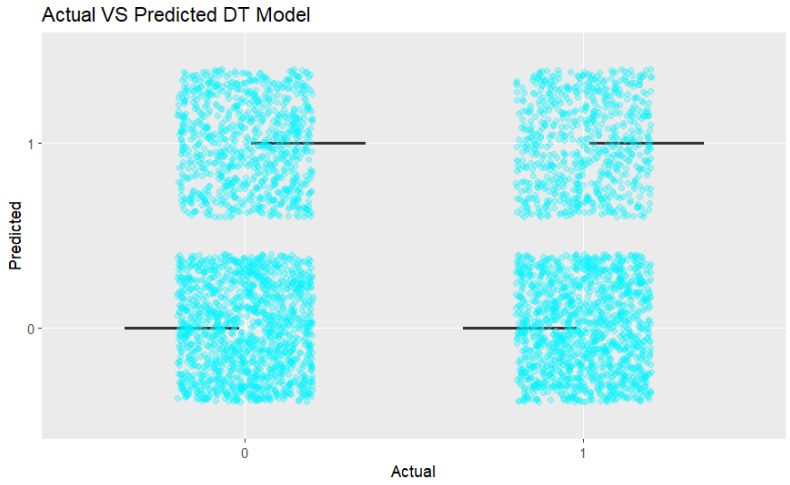
Neural networks



	0	1
0	1876	876
1	324	1129

With 0.71462 of accuracy, the confusion matrix suggest that our model has false negative of 324 data, which means that our model predict 324 persons will default but they actually paid the loan, and 876 false positive which our model predict not default but they actually default.

Decision tree CART



	0	1
0	1754	922
1	446	1083

Decision tree CART overall accuracy is 0.6746. The model correctly predict off the test dataset that 1754 individuals (true positive) paid their loan while 1083 people default (true negative). The DT CART confusion matrix suggest that our model has false negative of 446 data, which means that our model predict 446 borrowers will default but they actually paid the loan, and 922 false positive which our model predict not default but they actually default.

Model Evaluation Approach

Optimizing model performance while upholding a clearly defined standard of explainability is the primary goal of this exam project. The models are evaluated under the headings of Model Performance and Model Explainability. The following subsections will detail the criteria that the derived model must meet for it to be suitable for deployment.

Evaluation of Model Performance

In the context of this study, model performance is defined as the model's ability to correctly predict potential defaults. This does not encapsulate the model's transparency and the model performance measures discussed in this section do not account for the model's transparency. To assess the performance of the model the following performance measures were derived:

Precision

Precision is the number of true positives divided by total positive number. Precision can be declared as the exactness of a model. It is the proportion of true positives among the predicted positives.

Therefore, Precision = $\frac{\text{True Positives}}{\text{False positives + True Positives}}$.

Recall

Recall is another number in confusion matrix, it represented by true positive is divided by true positives and false negatives. Recall = $\frac{\text{True Positives}}{\text{True Positive + False Negative}}$ We can call recall as sensitivity or true positive rate of a data. It is the proportion of positives correctly predicted.

Accuracy

It is the proportion of correct classifications in the evaluation data

Area Under receiver operating Curve (AUC)

The Receiver Operating Curve (ROC) measures the model's classification ability subject to varying decision boundary thresholds. The ROC plots the true-positive rate to the false-positive rate. The area under the curve (AUC) aggregates the performance measures given by the ROC curve. The class imbalance (discussed in Section 3.2.3) is considered when considering what performance measures must be favoured. Given that the proportion of defaults tends to

outweigh non-defaults, accuracy alone is not a sufficient measure to assess model performance. This is related to the misclassification cost imbalance seen with loan defaults. A false negative is significantly more costly than a false positive. Therefore, Recall and AUC measured are favoured for this research project. The precision and accuracy measures are only considered in the model selection phase when the Recall and AUC measures do not yield a conclusive optimal model.

Evaluation of Model Explainabilty

Explainability is subjective and cannot be quantified using a numeric measure. It is because of this that a definite standard of explainability was defined in Section 2. The transparency goal of the resulting solution will be to meet this standard. If the model cannot be adequately explained to this standard, then it will not be considered for deployment.

Evaluation

Analysis of model using on test dataset : Model Performance

The model performance subject to the measures discussed in Section above are summarized in the following table : Each model has the same performance profile. The implication that business with varying risk appetites will choose different models is not an issue when choosing between the above models as each model delivers the same risk profile. This means evaluating the models subject to a varying misclassification cost as seen in Chen et al. (2021) is not necessary and likely will not yield any significant results.

Table 3: Comparison of Model Performance Between Algorithms

Model	Accuracy	Error rate	Exec-Time	Recall	Precision	F1-score
Logistic R	0.7481302244	0.2827586	2.53 secs	0.8318	0.6908	0.7547
SVM	0.7408710955	0.2922711	43.67 secs	0.9459	0.6521	0.7720
Bagging	0.6725840000	0.3274160	0.75 secs	-	-	-
RF	0.8042234932	0.2137931	-213.48 secs	0.8172	0.7834	0.8
GB	0.6907171139	0.3400713	45.96 secs	0.7772	0.6452	0.7051
NB	0.5037395513	0.4663496	0.15 secs	0.9922	0.5289	0.6900
NNT	0.8013638363	0.2853746	88.16 secs	0.8527	0.6816	0.7576
DT	0.7012758469	0.3253270	1.37 secs	0.7972	0.6554	0.7194
XGBOOST	0.0002199736	0.9997622	0.04 secs	0.5	1	0.6666

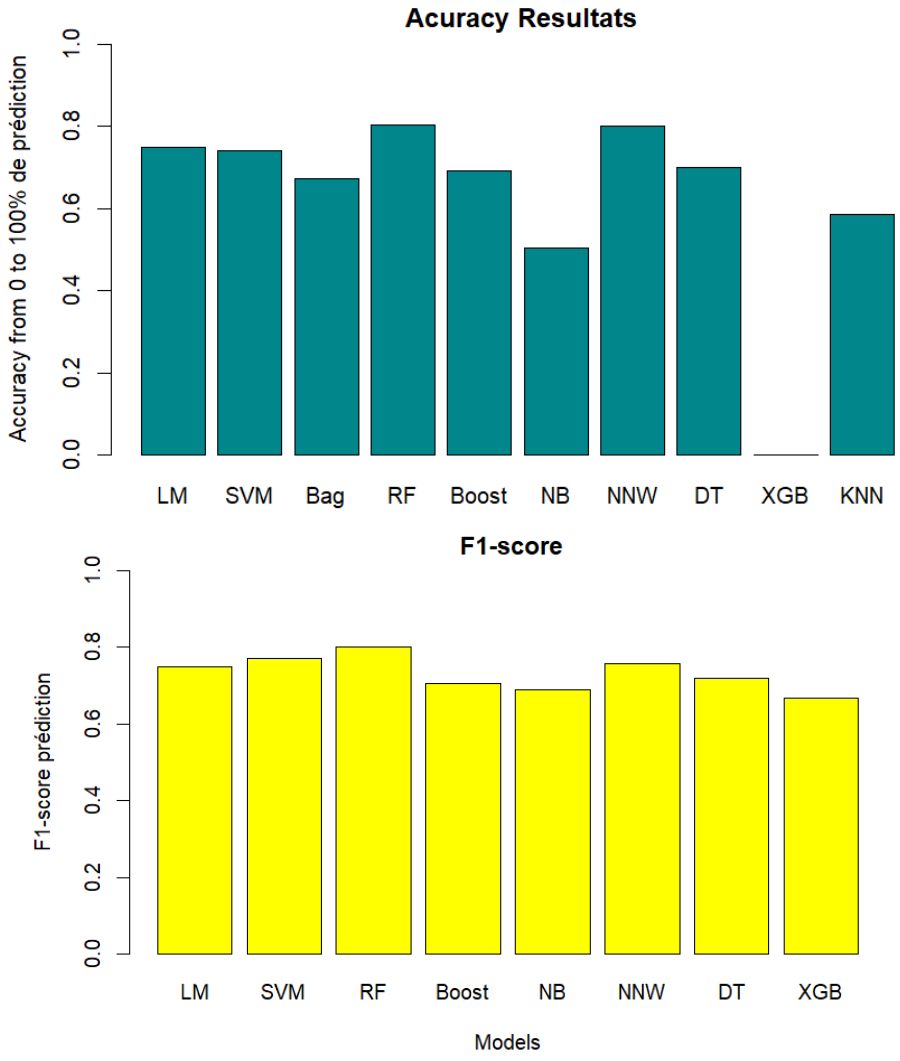
When applying decision tree (DT) model on our data, more than 5 trees were created due to the number of categorical features present. As a matter of fact, one of the major benefits of this kind of method is its simplicity to understand and interpret. However, they are highly biased in favor to categorical variables. This model resulted on the least prediction rate with an accuracy of 0.6746730083. On the other hand, Random forest is constructed of a multitude of decision trees and outputting the class that is the mode of the classes (classification) of the individual trees. We use the R package 'randomForest' to apply the model on our dataset. This model performance comparing to the previous models, with an accuracy of 0.8042234932. it is following by neural network (0.8013638363) and logistic model (0.7481302244), and support vector machine (0.7408710955). XGBoost 0.0002199736 have fewest accuracy.

The logistic regression model predicted 83.18% of the defaulted loans correctly and 59.15% of the non-defaulted loans. Accuracy is also presented in table 3, although it is not the most reliable metric for this research. The support vector machine regression model predicted 94.59% of the

defaulted loans correctly and 44.64% of the non-defaulted loans. Accuracy is also presented in table 3, although it is not the most reliable metric for this study.

In this section, we present the results of our classification models

We choose accuracy to identify the best model because probably this is the most common evaluation metric for classification problems. It performs better for equal number of observations in each class. Graph 8 displays each model's accuracy and a comparison between them. Random Forest had the highest prediction rate, with an accuracy of 0.7862068966 with the best score (0.8), and NNT had the second place with an accuracy of 0.7146254459, followed by the Logistic R model 0.7172413793. XGBOOST and Naïve bayes had the least prediction rate with an accuracy of 0.0002378121 and 0.5336504162 respectively, as illustrated on graph 11.



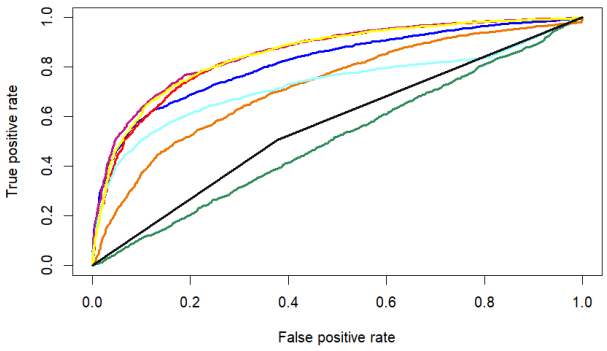
6.2. AUC

Logistic R	0.8126635
RF	0.8499422
Boost	0.8642674
DT	0.7747837
NB	0.7230685
NNT	0.7282723
KNN*	0.579058*

As can be seen from the table, the Boosting's perform best. Interestingly, all machine learning techniques except the NNT, NB and K-NN perform about equally well. The KNN performs much worse, which is expected, since it is a very simple machine learning technique. Indeed, the decision tree is so simple that one cannot blame it for the same black-box properties as the other techniques. Also, one would expect the Random Forest to outperform the Decision Tree, since the former is a more advanced and general version of the latter. As seen from the table, The K-NN algorithm has relatively low predictive power. The RF performs slightly better when it is deep.

Since the classifiers perform roughly equally well, with most classifiers achieving a ROC AUC between 0.72 and 0.86, it is difficult to clearly see the difference. Still, one might observe that the Boosting (light marron) consistently has a higher TPR (Sensitivity) for a given FPR (1-Specificity) than the other classifiers. Also, it is clear from the figure that the neural network (light green), and the K-NN (light black) perform worst.

6.3. ROC



Conclusion

This exam work aimed to explore, analyse, and build a machine learning algorithm to correctly identify whether a borrower, given certain attributes, has a high probability to default on a loan. This type of model could be used by Lending bank to identify certain financial traits of future borrowers that could have the potential to default and not pay back their loan by the designated time. From the result above, we know the following: Random Forest had the highest prediction rate, with an accuracy of 0.7862068966 with the best score (0.8), and NNT had the second place with an accuracy of 0.7146254459, followed by the Logistic R model 0.7172413793 while XGBOOST and Naïve bayes had the least prediction rate with an accuracy of 0.0002378121 and 0.5336504162 respectively. Based on Random forest, the most important variables are Number Of Time 30.59 Days Past Due Not Worse, Number Of Times 90 Days Late, Number Of Time 60.89 Days Past Due Not Worse, Revolving Utilization Of Unsecured Lines, and age. For age around 25 to 30 the value of the loan may be lowered The results can be improved by better data preparation. Hence, the Random Forest model appears to be a better option for such kind of data. Banks must be careful when identifying potential borrowers who fit certain criteria. For example, borrowers who do not own a home and are applying for a small business or wedding loan, this could be a negative combination that results in the borrower defaulting on a loan. Random forest model prediction is closed to testresponse base.

References

Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Shmilovici, A., Kahiri, Y., Ben-Gal, I., & Hauser, S. (2009). Measuring the efficiency of the intraday forex market with a universal data compression algorithm. *Computational Economics*, *33*(2), 131–154.

Zhang, G. P. (2010). Neural networks for data mining. *Data Mining and Knowledge Discovery Handbook*, 419–444.

Fan, A., & Palaniswami, M. (2000, July). Selecting bankruptcy predictors using a support vector machine approach. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium (Vol. 6, pp. 354-359). IEEE

Moula, F. E., Guotai, C., & Abedin, M. Z. (2017). Credit default prediction modeling: an application of support vector machine. Risk Management, 19(2), 158-187.

Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. Expert systems with applications, 28(1), 127-135.

Chen, M. Y. (2011). Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. Computers & Mathematics with Applications, 62(12), 4514-4524.

Breiman, Leo (2001). “Random forests”. In: Machine learning 45.1, pp. 5–32. Chawla, Nitesh V., Nathalie Japkowicz, and Aleksander Kotcz (2004). “Special issue on learning from imbalanced data sets”. In: ACM Sigkdd Explorations Newsletter 6.1, pp. 1–6.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). The Elements of Statistical Learning - Data Mining, Inference and prediction. en. Springer. URL: //www. springer.com/us/book/9780387848570 (visited on 02/01/2018).

Maimon, Oded and Lior Rokach (2005). "Introduction to supervised methods". In: Data Mining and Knowledge Discovery Handbook. Springer, pp. 149–164.

Frydman, H., Altman, E. I., & Kao, D. L. (1985). Introducing recursive partitioning for financial classification: the case of financial distress. The journal of finance, 40(1), 269-291

Anandarajan, M., Lee, P., & Anandarajan, A. (2004). Bankruptcy Prediction Using Neural Networks. In Business Intelligence Techniques (pp. 117-132). Springer, Berlin, Heidelberg.

Aziz, M. A., & Dar, H. A. (2006). Predicting corporate bankruptcy: where we stand?. Corporate Governance: The international journal of business in society. Bank for International Settlements (2009). Enhancements to the Basel II framework. https://www.bis.org/publ/bcbs157.htm

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240).

Tharwat, A. (2020). Classification assessment methods. Applied Computing and Informatics.