Akademia WSB

Faculty of Applied Sciences

Field of study
Data Science / Management

MA THESIS

Rahul Rahul

Default payment prediction of Credit Card customers using machine learning

MA THESIS
written under the supervision of
dr Tomasz Kasprowicz

Approved ……………………………………………

Date and supervisor’s signature

Dąbrowa Górnicza 2020

Akademia WSB

Wydział Nauk Stosowanych
Kierunek studiów: Data Science / Zarządzanie

PRACA DYPLOMOWA MAGISTERSKA

Rahul Rahul

Default payment prediction of Credit Card customers using machine learning

Praca magisterska
napisana pod kierunkiem
dr Tomasz Kasprowicz

Pracę przyjmuję, dnia……

……………………………
podpis promotora

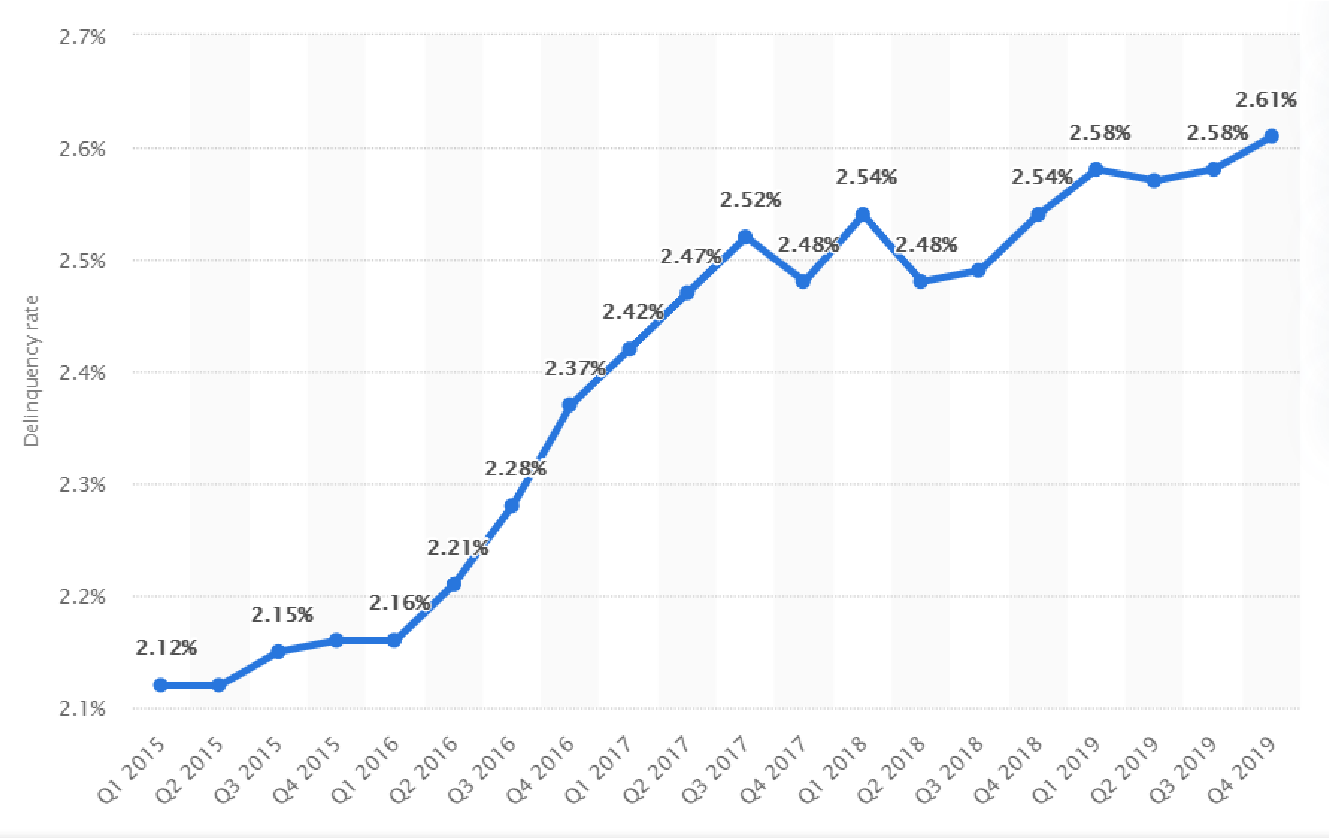
DĄBROWA GÓRNICZA 2020

TABLE OF CONTENTS

INTRODUCTION

According to association of banks in United States, there are total 374 million running credit card accounts in United states as of 2019. The credit card accounts are increasing at the rate of 2.5% from past few years. A study showed that 7 out of 10 customers in US have at least 1 credit card. As credit card accounts are growing, the delinquency rate on credit loans also increased in all banks and lending institutions in US from 2.12% in 1st quarter of 2015 to 2.61% in 4th quarter of 2019 as shown in figure below[1].

Figure 1: Delinquency rate of credit card loan in US



Source: https://www.statista.com

This results in the significant amount of money loss across all the institutions or commercial banks. So, it’s important for the lending institutes or banks to have a risk predictive model that can able to predict weather the customer do a default or not on credit card loan based on the past behavior of the customers.

[1] https://www.creditcards.com/credit-card-news/ownership-statistics/, access: 14.09.2020

In recent years, the usage of machine learning tools is greatly increased. Not long time ago machine learning used mostly for research purpose but in recent years it is favorite tool among data scientist developers. With the help of machine learning we can create this type of model. The primary aim of machine learning algorithms is to identify patterns in the historical data during the training process and based on these patterns try to predict the future values.

While selecting the most appropriate model or algorithm, there are many points that need to be considered. First of all, there is no algorithm that performs always better than the other. Also no algorithm performs always best for specific problem or one type of problem. Therefore, the best way is to implement various machine learning models on the same problem, find the results and then carefully evaluate and compare the results in order to select the best model.

CHAPTER 1

Machine learning and it's applications

1.1 Introduction

From the last few years machine learning is becoming increasingly popular all around the world in the area of predictive analytic s. Even though the machine learning term is not new, it was first invented by Arthur Samuel in the late 1950s. Frankly speaking, The primary aim of machine learning algorithms is to identify patterns in the historical data during the training process and based on these patterns try to predict the future values. Before 1990's the machine learning was mostly used for knowledge driven and for research. But from 1990's the big data driven companies started to deploy machine learning models more often in the data. The reason for machine learning popularity in recent years is, the companies all around the world are generating data more than ever before. And to find conclusive insights and make sense from the data, machine learning algorithms work as a great tool.

1.2 History of Machine learning

In 1950, first time machine learning terminology came in to existence when Alan Turing wrote an article and proposed a hypothesis (also known as Turing test) on Artificial intelligence. It states "The machine that succeed in convincing humans is not a machine, It has achieved an artificial intelligence."

In 1957, Frank Rosenblatt developed world first neural network. He used this neural network to create a model which classified inputs categories in to one of the two output class. This neural network was later called as a perceptron model.

In 1959, two scientist Bernard Widrow and Marcin Hoff created neural networks called Adeline and Madeline. Adeline was used in to identify binary patterns and Madeline was used in to remove echo in a telephone line. The Madeline had a real world application[2].

In 1967, the k-Nearest Neighbor algorithm was developed. It is one of the supervised learning algorithm and widely used in today real world applications.

[2] https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/history1.html, access: 14.09.2020

In 1980, Gerald Dejong introduced the idea of explanation based learning. In this approach, the algorithm analyze the data set and made some set of common rules. And based on these rules the algorithm get rid of the unnecessary information present in the data set[3].

Before 1990, the machine learning algorithms mostly used in the knowledge driven approach. But starting from 1990, the machine learning models started using more in data driven approach. The companies started creating applications that analyze the data in huge amount and draw conclusion and results from them. These results helps the companies to improve and give better services to the customers.

In 1997, a chess tournament conducted between the artificial intelligence powered computer named (deep blue) created by the International business machines corporation popularly known as IBM and the Russian chess grandmaster Gary Kasparov. The final result of the tournament shocked everyone. Gary Kasparov loss the tournament from the AI powered computer. This was the first time the AI beat the human intelligence.

In 2011, using a combination of natural language processing (NLP) and information retrieval techniques, IBM developed computer named Watson beat two human champions in a game of “Jeopardy”[4].

In 2016, The AI software developed by google defeated the European champion in a game “Go”. It was the major breakthrough in a field of artificial intelligence[5].

1.3 Categories in machine learning algorithms

There are many different categories of algorithms in machine learning but there are two which are most commonly used- supervised and unsupervised learning. We are also going to see a little details of reinforcement learning which is the third category in the later section. In starting we are going to discuss the supervised and supervised learning in detail and most importantly the supervised learning algorithms which we are going to use in our thesis research problem.

[3] https://roboticsbiz.com/machine-learning-the-complete-history-in-a-timeline/, access: 14.09.2020
[4] https://www.theguardian.com/technology/2011/feb/17/ibm-computer-watson-wins-jeopardy, access: 14.09.2020
[5] https://www.bbc.com/news/technology-35420579, access: 14.09.2020

Supervised learning: In supervised learning we try to find the relationship between the input values and the output value on a given data set and then based on this relationship predict the output of future input values. In supervised learning algorithms we already know in advance how our output will look like.

Supervised learning algorithms are categorized in to classification and regression problems. In the regression problems the output value that we try to predict are continuous in other words mapping the input variables into continuous output values whereas in classification problem the output are in discrete in another word mapping the input variables in to discrete categories.

Example: The prediction of house cost based on the various features of house like house size, number of rooms, how far is the house from city center etc. Here the output will be in continuous value and so this a regression problem.

Now suppose we have given a dataset of lots of different mobile phone models and given various features of mobile phones such as, how much RAM does it have, the storage capacity of the mobile, battery power, the number of cores it have, the weight of the mobile and etc. And we want to classify all these phones in a different classes like weather a specific mobile phone belongs to budget range, or they belongs to mid-range, or they are part of expensive range phones. So by looking at the mobile features we can create a classification model which can help us to categorizes the mobiles in to different classes.

Some of the well-known supervised learning algorithms are regression models, k nearest neighbor, decision tree, Naive Bayes, random forest etc. We will see the detailed information of these algorithms in the next chapter.

The application of supervised learning algorithms are:

1. Spam detection: Supervised learning is used to classify an Email whether it is a spam or not. The Gmail has an algorithm that learn a fake keywords such as “you are a winner of” and so forth and put those messages with these keywords in spam folder.

2. Speech recognition: In speech recognition you teach the algorithm by your voice and it will be able to recognize you later and help in translating in to text and command. Google virtual

assistant and Siri are some of the well-known application in this domain which wake up only with your voice keywords.

3. Fraud and risk detection: Supervised learning are used in detecting the fraudulent transactions done by the users and also to access risk in insurance and financial companies and help them to minimize the risk.

Unsupervised learning: The unsupervised learning is an another form of machine learning. It is different from supervised in the sense that, in supervised the inputs and the output is already given for training the model and based on the relationship between the inputs and output value we try to predict the output of new data which have not been trained. So in supervised we already know how the output will look. But in unsupervised learning there is no output is given. The algorithms itself try to find the patterns among the data using some approach and try make some clusters and group based on some similarities. So any output can come which we can never thought in advance. K-Means clustering, Hierarchical clustering and Apriori algorithm are some of the well-known algorithms used in unsupervised learning.

Some applications of unsupervised learning algorithms are:

1. Market basket analysis: Apriori algorithm is most commonly used in market basket analysis. In this problem we tried to find the strong association rules with some probabilistic approach between the different set of products. For ex. If a customer is buying bread and butter then most probably he will buy the milk also etc. With the help of these strong association rules we can recommend products to the future customers who have same buying patterns[6].

2. Identifying accident prone areas: Unsupervised learning models used in identifying areas which are more prone to accident by analyzing the severity of accidents happened in order to introduce the safety measures.

3. Clustering documents: Suppose we have given millions of documents ranging from many different educational field. And suppose we have divide those documents and make a clusters so

[6] https://www.datacamp.com/community/tutorials/market-basket-analysis-r, access: 14.09.2020

that documents from same domain make a single cluster. Here we will use the clustering algorithm which will analyze the words in the documents and group all the documents together in a cluster which has a high number of words similarity between them.

Reinforcement learning

The third type of machine learning is the reinforcement learning. This learning is based on the reward and penalty. In reinforcement learning we have an agent and an environment. This agent can be anything and we see it later in an example what does it actually means. The agent performs some actions in an environment and in return he got the change state of an environment and the reward or penalty of the action the agent took. Then based on the change state the agent will make a new policy and try to perform the action in another way so that he got more reward for by learning from the past actions.

Let’s understand this concept with an example, the one application of the reinforcement learning is that it frequently used in the recent games. Suppose the AI bots is playing opposite to human in a shooting game. Here the bots are agents and the game is an environment. Suppose the bots performs some action in response to the human action. Either the bot get killed or the player selected by the human killed. If the bot get killed he received a penalty and if he kills the human player then he got a reward. So by every reward and penalty point the bot try to analyze the situation through reinforcement learning and decide which action he had to perform next time and which action he have to discarded.

CHAPTER 2

Methodology

2.1 Logistic regression

Logistic regression is a classic supervised learning statistical model and the specialized version of linear regression. It is powerful, fast and easy to implement. To understand the logistic regression, let's understand first what is linear regression.

Linear regression

Linear regression is a model which is used extensively in predictive modeling. There are two type of linear regression – simple and multiple. The simple linear regression is used when there is only one input variable in a model whereas the multiple linear regression is used when there are more than one input variables. The linear regression try to find the relationship by drawing a straight line between the input variables (also known as independent variables) and the output (also known as dependent or target) variable.

While performing simple linear regression we take 4 key assumptions[7].

1. Linearity: The relationship between the output variable also known as dependent variable and the independent variables will be linear.

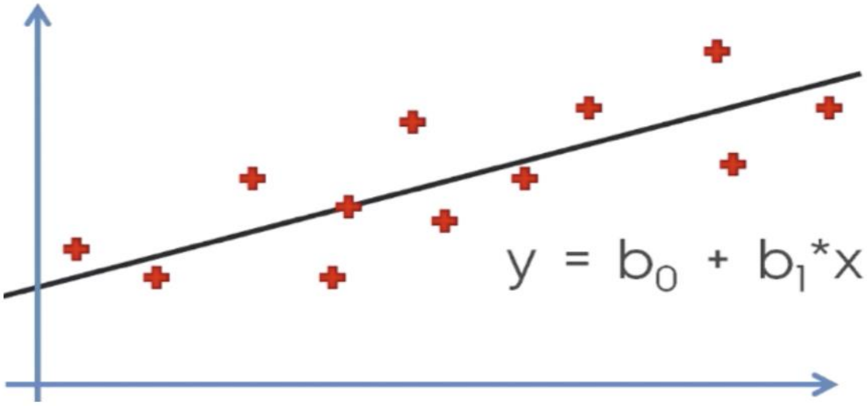
2. Independent: The observations should be independent to each other.

3. Homoscedasticity: The error between the observed value and predicted value is the same for every value of explanatory variable. The error between the observed value and predicted value also known as residual variance.

4. Normality: The residuals follow a normal distribution.

[7] https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html, access: 14.09.2020

Figure 2: Linear regression graph



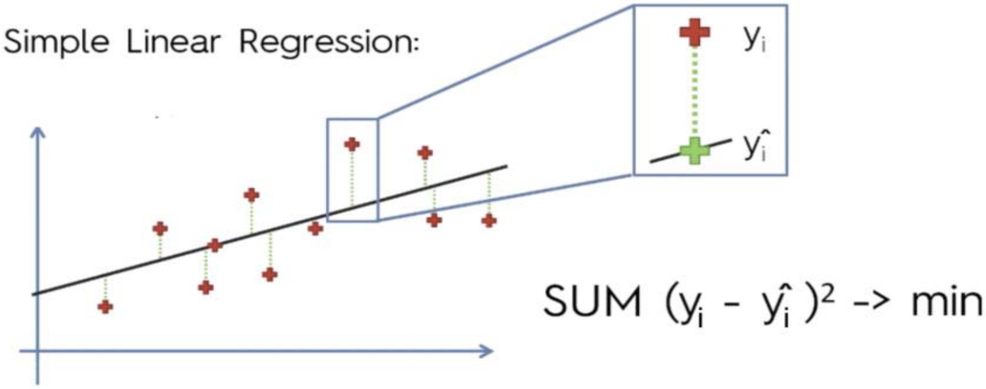
Source: https://www.udemy.com/course/machinelearning/learn/lecture/6270024#overview

The above graph is for the simple linear regression with one predictor variable x. The equation of the multiple linear regression is:

$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \ldots. + b_nx_n$

The red points in above graph are representing sample data. The black line is the best fitting straight line on the sample data. By best fitting we mean the sum of squared errors (SSE) between the observed value and the predicted value should be minimized as described in the graph below. The other method to minimize is mean squared error (MSE).

Figure 3: Sum of squared error in linear regression

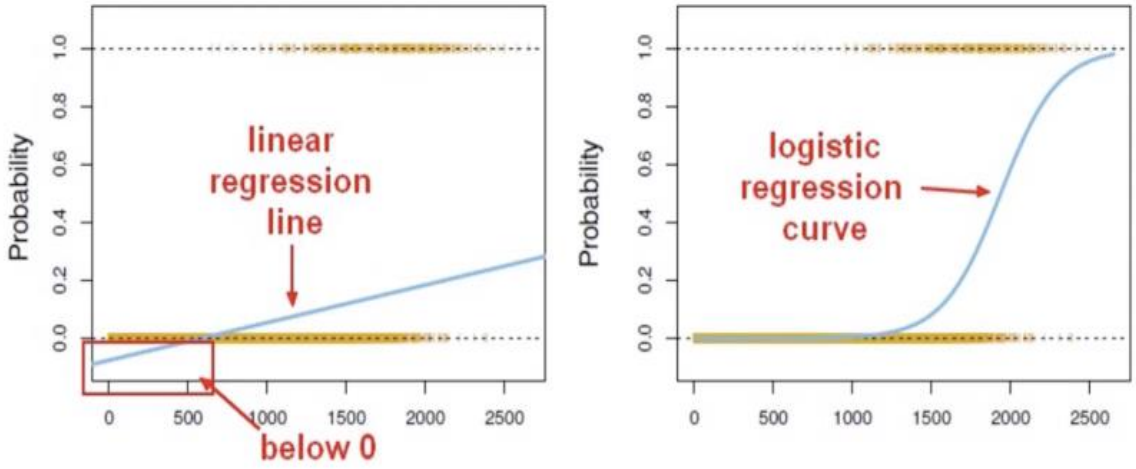


Residual: ***e*** *= y — ŷ (Observed value — Predicted value).*

Source: https://www.udemy.com/course/machinelearning/learn/lecture/5772186#overview

Now let’s talk about classification problem. In classification problem we predict the discrete output values. Let’s take an example of flipping the coin. There can be only two possible outcomes (head or tail). The 0 represent the head and the 1 represent the tail or vice versa. This is also an example of Bernoulli variable where the probabilities stay in between the 0 and 1 including both.

Figure 4: Difference between linear regression line and logistic regression curve



The linear regression line is below 0.

Source: https://datacadamia.com/data_mining/simple_logistic_regression

Linear regression deals with only continuous variables. Noted that the classification does not follow normal distribution which is violating the normality assumption of linear regression. Therefore the linear regression model is not suitable for classification problem. This problem can be solved by converting the linear regression in to logistic regression using logistic function (also known as Sigmoid function)[8].

Before we further discuss logistic regression, let’s understand the fundamental statistical term-probability and odds.

Probability: Let’s understand probability with a simple example. Suppose you have an unbiased die and you want to find the probability of getting an even number. While rolling a die there are total six possible random outcomes ranging from 1 to 6. The total number of even numbers are 3 (which are 2, 4 and 6). So, the probability of getting even number is 3 divided by 6 which is equal to 0.5.

[8] https://deepai.org/machine-learning-glossary-and-terms/sigmoid-function, access: 14.09.2020

Odds: The odds is defined by the following formula

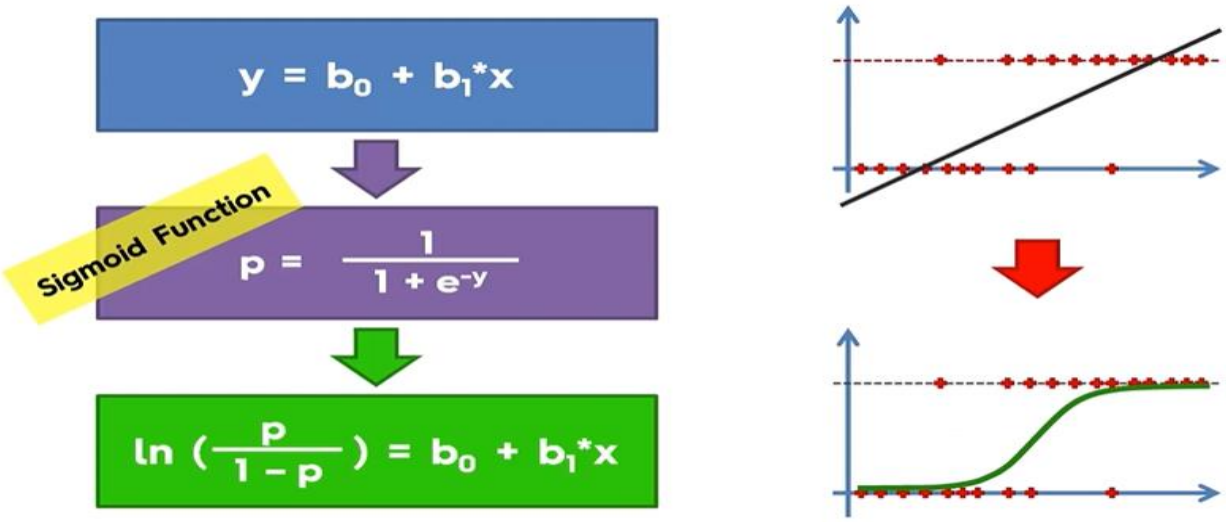
=> Odds = P / (1-P)

where P is the probability of an event happening in some random experiment and (1 – P) is the probability of not happening of that event.

In a rolling die example, the probability of getting even number is 3/6 which is equal to ½. The probability of not getting an even number is 1 – (½) = ½. So the odds are (½) / (½) which is equal to 1.

Let’s transform the simple linear regression to logistic regression using sigmoid function.

Figure 5: Converting linear to logistic regression using sigmoid function


$y = b_0 + b_1{*}x$
Sigmoid Function
$p = \frac{1}{1 + e^{-y}}$
$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1{*}x$

Source: https://www.udemy.com/course/machinelearning/learn/lecture/6270024#overview

With the help of sigmoid function we can transform the output of linear regression which is continuous value in between 0 and 1. The general understanding of logistic regression is saying if the probability is equal and greater that 0.5 than we assign that point to class 1 otherwise we assign to class 0.

The advantages of logistic regression algorithm is that it can easily extends to more than two class (0 and 1). This is also known as multi class logistic regression. Another advantage is that the model are quick to train and fast at classifying unknown records. The disadvantage is, it require large sample size to accurately train and predict future values.

The application of logistic regression are:

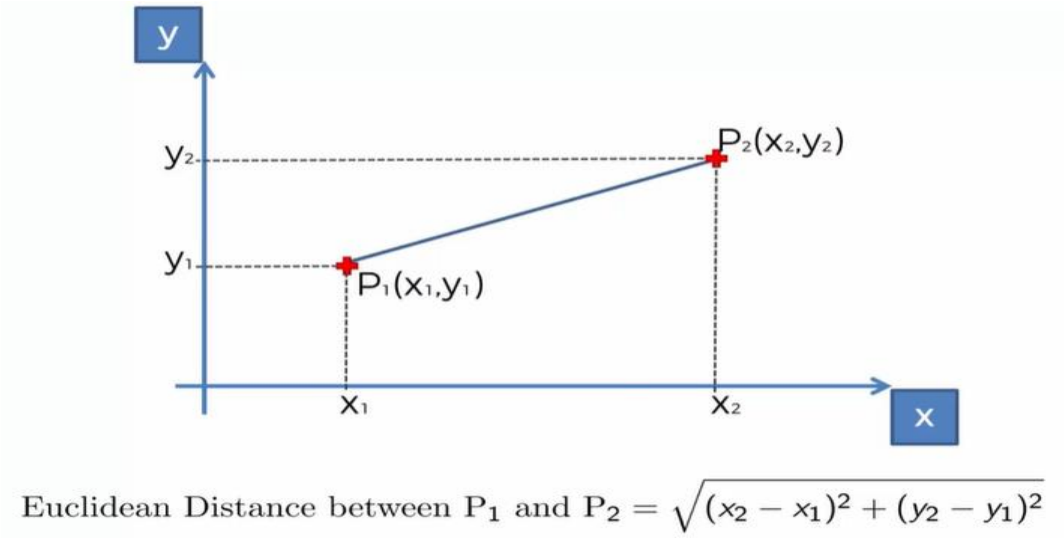
1. Medical diagnosis: Logistic regression had a many real world applications in the medical diagnosis. It can help in detecting the cancer at early stage by analyzing the various features of tumors and categorized in to whether the tumors is benign (not harmful or at the early stage) or a malignant (harmful or at advance stage). The other example is, it can also help in detecting the obesity in a person, weather a person is obese or not, by looking at various features of person such as the age, genotype and weight etc.

2. Weather forecast: Logistic regression can also be used in weather forecasting. The forecast can be whether it will be sunny, rainy, cloudy etc. And if we want to predict the temperature, then we will use the linear regression since, the output will be continuous value.

2.2 K-nearest neighbor classifier (K-NN)

K-NN classifier is the another algorithm for the classification model. This model works on the Euclidean distance concept.

Figure 6: Calculating distance between two points using Euclidean Distance


y
y₂
y₁
$P_2(x_2,y_2)$
$P_1(x_1,y_1)$
x_1
x_2
x
Euclidean Distance between P_1 and $P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Source: https://www.udemy.com/course/machinelearning/learn/lecture/5714404#overview

The steps in the model work as follow:

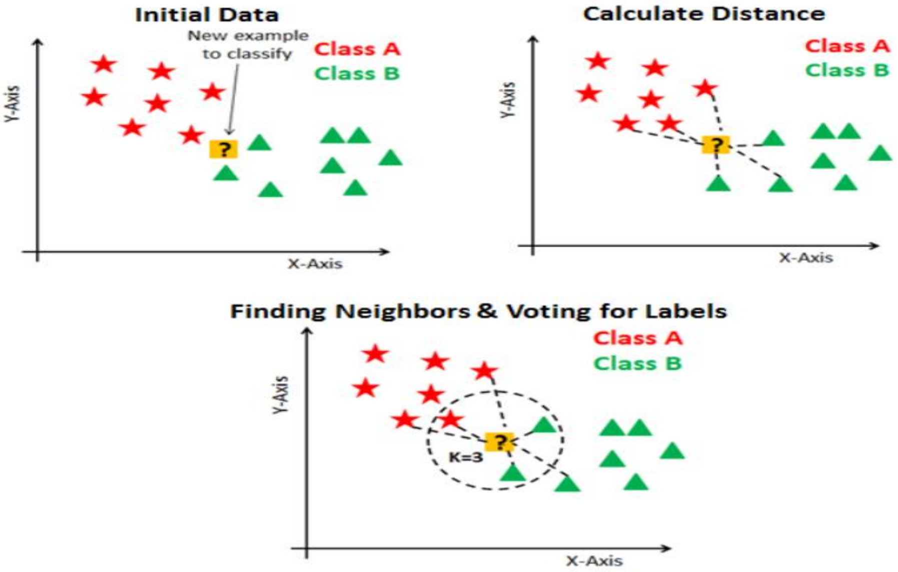
Step 1: Select K number of Neighbors for a model. This is an important step since, the model accuracy is depend on the value you will take.

Step 2: Find the K neighbors which are nearest to the new data point that you want to deploy. This can be done by calculating the distance between the each sample data points and the new data point using Euclidean distance and then select the K Nearest sample points around the new data point.

Step 3: Among this K nearest sample points, calculate the total number of point belong to each output category (class).

Step 4: Assign the new point to the output category whose count is highest.

Figure 7: K-NN algorithm steps



Source: https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

The advantage of this algorithm is that it is easy to understand. Even a Naive person can understand the intuition behind the algorithm. Another advantage is that, this algorithm work just easily in a problem when we have to classify data in to more than two class as compared to other classification algorithms. The disadvantage are, during the testing K-NN model try to find the nearest neighbor for each new data which make it slow and also require computational power. This is the reason this algorithm also called lazy algorithm.

The applications of KNN are:

1. Recommender system: The KNN algorithm is used in the recommender system engine. The recommender system is used in the Amazon website, where the customer can see the recommendation of products on the down of page. These recommendations are the similar products from different brands that the customer is looking for. Apart from Amazon, the recommendation engines are also heavily used in the Netflix media. Each and every customer receive recommendation of movies and web series based on analyzing their past searching behavior. These behavior can be based on the genre, The actors in the movie, the time length, the rating etc.

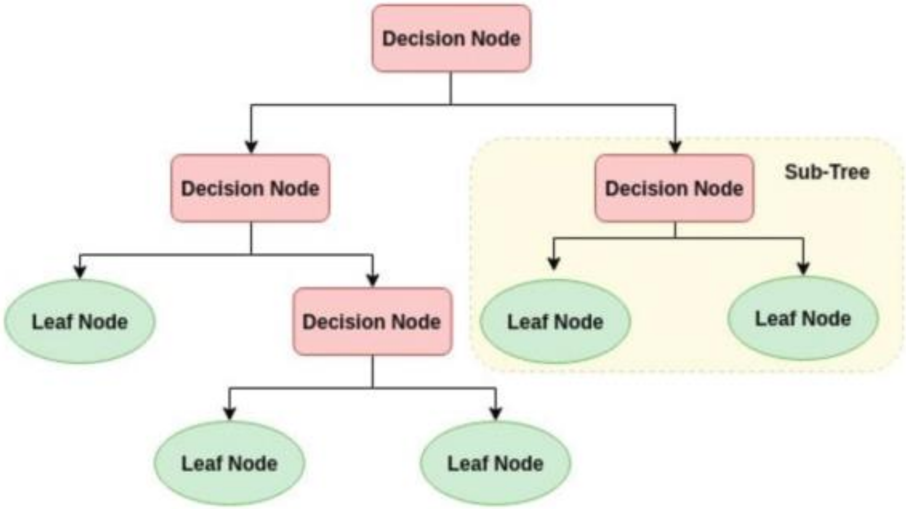
2. In Agriculture: The KNN is also used in the agriculture sector by predicting the weather forecast and estimating the soil and water parameters etc.

3. In finance: The KNN had a vast applications in financial domain. It can be used in the identifying the patterns in the stocks. Apart from this it can also be used in money laundering analysis, in the loan management, whether to give a loan or not (by analyzing the past data of customers).

2.3 Decision tree classifier

Decision tree (also known as classification and regression tree) is the oldest supervised learning concept which has still relevance today. Most of the machine learning researchers don’t recommend single decision tree for a model since it’s over fits a lot, but they recommend to use a group of decision trees to make an another algorithm know as Random forest which we will see later. This algorithm structure is resemble like a tree where root and internal nodes are attributes (or features) of a data set and the leaves are output class.

Figure 8: Decision tree



Source: https://devopedia.org/decision-trees-for-machine-learning

The steps in the algorithm work as follow:

1. First we have to select the attribute (column) for the root node. This attribute will be be selected among all the independent variables in the data set by some probabilistic approach.

2. Next, we divide the data in groups in such a way that each groups contains data from the same value of column.

3. Repeat both the steps above until you find the leaf (output class category) of each branch in a decision tree.

We take few assumptions before constructing the decision tree:

1. The values in columns in a data set is preferred to be discrete. If there are continuous values, then first we convert them in to discrete.

2. Not any attribute from the independent variables can be root node and the nodes at subsequent levels. There should be some criteria on the basis of which we assign the column to the root and other nodes.

Choosing the random attribute for root node and each level is not advisable since it may result to lower accuracy of model prediction. With the help of Information gain we can solve this selection criteria[9].

Before we discuss more about information gain let’s understand what is Entropy[10].

[9] Effective CRM using Predictive Analytics. Antonious Chorianopoulos. (2016). Wiley. p. 151. ISBN 978-1-119-01155-2
[10] https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8, access: 14.09.2020

Entropy is a matrix which measures the impurity of something. Impurity is the degree of randomness, means how random a data is. let’s understand this with an example.

Suppose you have 100 blue balls of same size in a bag and also you have 100 paper slips written as blue ball on each of them in another bag. Now the probability of selecting randomly one blue ball from bag and the corresponding paper slip from another bag is 1. So, in this case we say that the impurity is 0.

Now suppose you have 25 balls each of 4 different colors in a bag and the corresponding paper slips of equal amount in a different bag. Now the probability of selecting ball with a corresponding paper slip is not going to be 1. It will be less than 1. Any combination of ball and paper slip is possible. So, in this the impurity is going to be non-zero.

So coming back to entropy, The formula of entropy for binary classification (0 or 1) is,

Entropy = - (Prob(1) Log_2 Prob(1) + Prob(0) Log_2 Prob(0))

If total no. of 1 and 0 are in equal number, then
=> Entropy = 1

and, if all the no. are 1 or all are 0, then
=> Entropy = 0

The formula of Information gain is calculated as follow,

IG = E(Target attribute) – [(Weight Average) * E(each independent attribute)]

Where
E = Entropy
IG = Information gain

Let’s understand entropy and information gain concept by an example.

U	V	W	X	Y
4.8	3.4	1.9	0.2	Pass
5	3	1.6	0.2	Pass
5	3.4	1.6	0.4	Pass
5.2	3.5	1.5	0.2	Pass
5.2	3.4	1.4	0.2	Pass
4.7	3.2	1.6	0.2	Pass
4.8	3.1	1.6	0.2	Pass
5.4	3.4	1.5	0.4	Pass
7	3.2	4.7	1.4	Fail
6.4	3.2	4.5	1.5	Fail
6.9	3.1	4.9	1.5	Fail
5.5	2.3	4	1.3	Fail
6.5	2.8	4.6	1.5	Fail
5.7	2.8	4.5	1.3	Fail
6.3	3.3	4.7	1.6	Fail
4.9	2.4	3.3	1	Fail

In the above sample data set the U, V, W and X are the predictor attributes and the Y is the target attribute with two class pass and fail. This is also an example of binary classification. If the data need to classify in more than two class that it's a multi class classification problem.

Since the data is continuous we first convert it in to categorical by taking some random value.

U	V	W	X
>= 5.5	>= 3.4	>= 4.2	>= 1.3
< 5.5	< 3.4	< 4.2	< 1.3

The information gain of each attribute are calculated as follow:

- First calculate the entropy of each columns (U, V, W, X and Y) in a data set.
- Now the information gain of column U is calculated by subtracting the entropy of U from the entropy of output column. Same for column V by subtracting the entropy of V from the entropy of output column and so on.

Since the total number of pass and total number of fail are equal in target attribute Y, the entropy will going to be 1.

We can show it by Entropy formula also

$E(Y) = -((Prob(pass)*\log_2(Prob(pass)) + (Prob(fail)*\log_2(Prob(fail)))$
$= -((1/2)*\log_2(1/2)) + (1/2)*\log_2(1/2))$
$= 1$

Information gain of attribute U:

For attribute U >= 5.5 AND Y == Pass: 0/7

For attribute U >= 5.5 AND Y == Fail: 7/7

$E(0,7) = -((0/7)*\log_2(0/7) + (7/7)*\log_2(7/7)) = 0$

For attribute U < 5.5 AND Y == Pass: 8/9

For attribute U < 5.5 AND Y == Fail: 1/9

$E(8,1) = -((8/9)*\log_2(8/9) + (1/9)*\log_2(1/9)) = 0.5$

$E(Y, U) = Prob(>= 5.5)*E(0,7) + Prob(<5.5)*E(8,1)$
$= (7/16)*0 + (9/16)*0.5 = 0.28$

Information gain of U = E(Y) – E(Y, U) = 1 – 0.28 = 0.72

Information gain of attribute V:

For attribute V >= 3.4 AND Y == Pass: 5/5

For attribute V >= 3.4 AND Y == Fail: 0/5

$E(5, 0) = -((5/5)*\log_2(5/5) + (0/5)*\log_2(0/5)) = 0$

For attribute V < 3.4 AND Y == Pass: 3/11

For attribute V < 3.4 AND Y == Fail: 8/11

$E(3, 8) = -((3/11)*\log_2(3/11) + (8/11)*\log_2(8/11)) = 0.83$

$E(Y, V) = Prob(>=3.4)*E(5,0) + Prob(<3.4)*E(3,8)$
$= (5/16)*0 + (11/16)*0.83 = 0.57$

Information gain of V = E(Y) – E(Y, V) = 1 – 0.57 = 0.43

Information gain of attribute W:

For attribute W >= 4.2 AND Y == Pass: 0/6

For attribute W >= 4.2 AND Y == Fail: 6/6

E(0, 6) = 0

For attribute W < 4.2 AND Y == Pass: 8/10

For attribute W < 4.2 AND Y == Fail: 2/10

E(8, 2) = 0.72

E(Y, W) = Prob(>=4.2)*E(0, 6) + Prob(<4.2)*E(8, 2)
= 0.45

Information gain of W = E(Y) – E(Y, W) = 1 – 0.45 = 0.55

Information gain of attribute X:

For attribute X >= 1.3 AND Y == Pass: 0/7

For attribute X >= 1.3 AND Y == Fail: 7/7

E(0, 7) = 0

For attribute X < 1.3 AND Y == Pass: 8/9

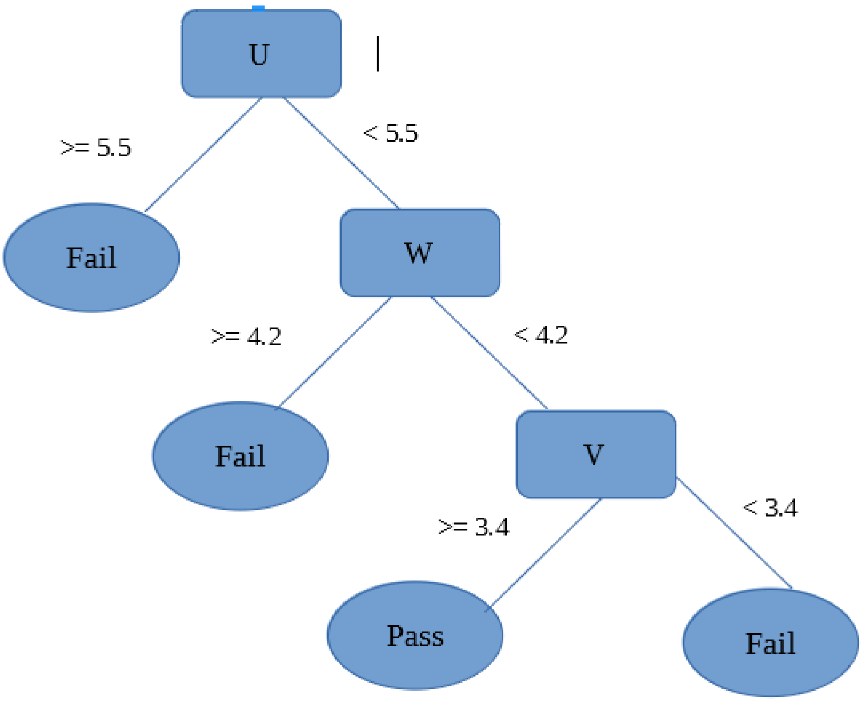
For attribute D < 1.3 AND Y == Fail: 1/9

E(8, 1) = -((8/9)*$\log_2$(8/9) + (1/9)*$\log_2$(1/9)) = 0.5

E(Y, X) = Prob(>=1.3)*E(0, 7) + Prob(<1.3)*E(8, 1)
= (7/16)*0 + (9/16)*0.5 = 0.28

Information gain of X = E(Y) – E(Y, X) = 1 – 0.28 = 0.72

Now we have information gain for all the attributes we can start building decision tree. The highest information gain value attribute will position on root node and the subsequent attributes position according to the sorted value. A branch with 0 entropy will convert to leaf node. And the branch with entropy greater than 0 need further splitting.



The advantages are, in best case the decision tree runs in logarithmic time at a time complexity of O(Log N), where N is the number of rows for training . This algorithm tries to solve the problem same like the humans take a decision at each step. So this make an algorithm unique among other classification problem in this aspect. The other advantages are, this algorithm can easily take care of any missing value in a dataset and also if some attributes which doesn’t contribute anything in an algorithm (attributes whose information gain is 0) will be discarded by decision tree.

The disadvantages are, overfitting is a huge issue in this algorithm. Overfitting can cause a problem while classifying the new data points to a correct class. Another disadvantage is, this algorithm try to use greedy approach while solving problem which may result in to model which is not best. Most of the disadvantages of this algorithm can be solved by Random forest classifier which we are going to discuss in our next section.

2.4 Random forest classifier

Random forest is an ensemble supervised learning algorithm. Ensemble learning model use multiple individual algorithms at the same time to obtain a prediction with an aim to have better prediction than the individual model. Ensemble learning give us better accuracy which mean lower error. It gives higher consistency by avoiding over fitting problem and also it reduces bias and variance errors. We can build random forest model by taking multiple random decision tree on a same data set[11].

The steps in the random forest model work as follow:

Step 1: Randomly select k data points from the data set.
Step 2: Apply the decision tree classifier on these data points.
Step 3: Select N no. of decision trees and repeat above 2 steps.
Step 4: For the new row in a data set, make each N decision tree to predict the output class and assign the row to the class who win the majority of vote.

The advantages of random forest are that this model can handle the missing data in a column with a greater accuracy as compared to other models. Another advantage is that this algorithms can be used both in classification as well as in regression problem. Over fitting is not a problem and also it has a power to handle huge data set with lots of columns.

The applications of random forest are:

- In object detection: This algorithm can be greatly used in detecting the specific object in a most complex conditions. For example in a traffic where we are trying to classify different vehicles such as cars, buses and trucks in their respective category this algorithm can be used.

- Remote sensing: This algorithm used in ETM device which is installed on a satellite. The ETM stands for Enhanced Thematic Mapper. The ETM is used to take an image of a surface of a planets with a greater accuracy.

[11] https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/, access: 14.09.2020

- Kinect game console: The well know application of random forest is that this algorithm is a used by Microsoft in developing the Kinect. The Kinect console track the body movement of a human by sensor and then try to recreate the same movement in a game[12].

- E-commerce: Random forest also used in recommender system on E-commerce websites. The recommender system suggests the related products to the customers based on their previous purchase history and also what the customer searched on website.

2.5 Naive Bayes classifier

Naive Bayes is a classification algorithm based on Bayes theorem. This model takes an assumption that the effect of features or explanatory variables are independent to the output class even when these features are dependent to each other. This assumption make the calculation easier and make this algorithm named Naive. This model works on the conditional probability. The equation of the conditional probability is[13],

$$P(J/K) = (P(K/J) * P(J)) / P(K)$$

where

P(J) = It denotes the probability of assumption that has been taken being true. This is also called prior likelihood.
P(K) = It is defined as the likelihood of evidence. Also known as marginal likelihood.
P(K/J) = It defined as the likelihood of evidence given that assumption is correct. Also known as likelihood.
P(J/K) = It defined as the likelihood of assumption with the evidence is there. Also known as posterior likelihood.

[12] https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/BodyPartRecognition.pdf, access: 14.09.2020
[13] https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/, access: 14.09.2020

The Naive Bayes algorithm work as follow:

1. At first we have to compute the prior likelihood of the given class category in the dependent variable.

2. Next we compute the conditional probability of each features or columns in the dataset to each class category of dependent variable.

3. Then the conditional probability would be multiplied that belong to same class category.

4. Next, we multiply the prior likelihood with the probability that has been calculated in step 3.

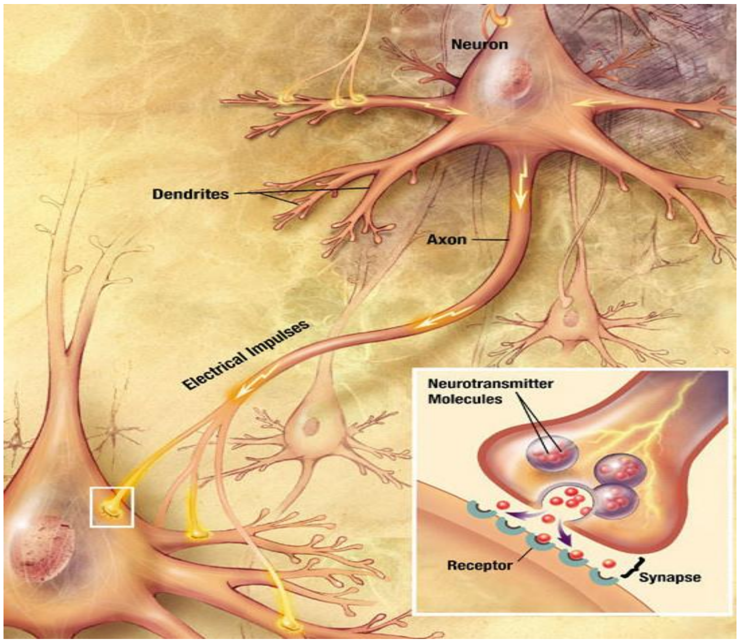
5. Next we see which class category in the dependent variable has the high probability, that class category belong the given row of the data set.

The advantages are that this algorithm can work efficiently on big data set. The disadvantage is the assumption that we took in the starting of algorithm. In reality, it is impossible to get an explanatory variables which are completely independent to each other’s.

2.6 Neural Network classifier

Neural network is a very powerful algorithm and can be used both in classification as well as regression problem. For understanding neural network let's go back and take a look in basics Neuroscience theory. Below is the diagram of Neuron[14].

Figure 9: Neuron Image



Source: https://en.wikipedia.org/wiki/Neuron#/media/File:Chemical_synapse_schema_cropped.jpg

The dendrites are the receiver of input signals in a neuron. The signal can comes from five sense organs or from another neuron. There are billions of neuron in the brain and are interlinked to each other. The neuron is a type of function where actually decision making take place. The output signals from neuron are transmitted through an axon in the form of electrical impulses. The axon from one neuron is not physically attached to the dendrites of another neuron instead they are attached through synapse. After signals passing through millions of neuron the brain takes decision.

[14] https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/index.html, access: 14.09.2020

Artificial neural network in a machine:

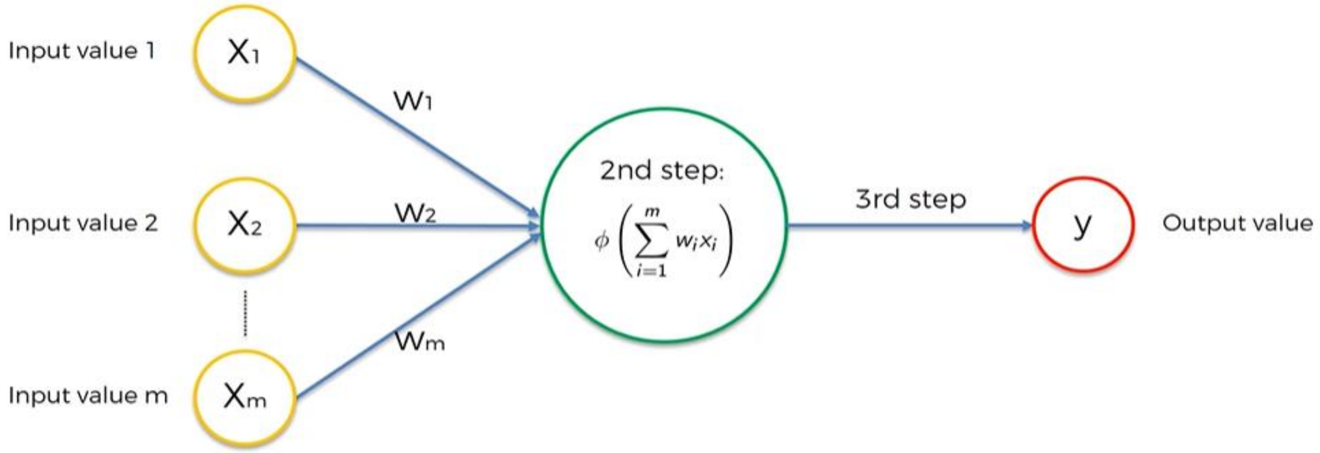


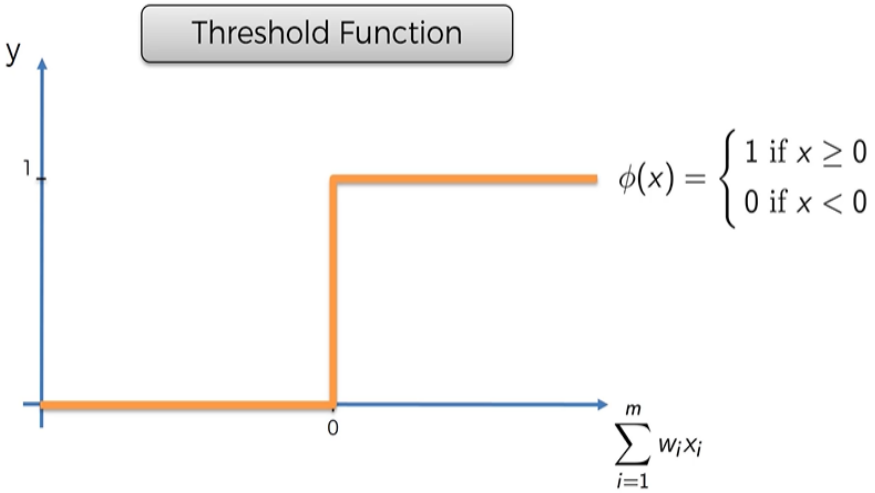
Figure 10: Neuron workflow

Source: https://www.udemy.com/course/machinelearning/learn/lecture/6760384#overview

The above diagram is the basic diagram of Neuron. The $X_{1,}$ X_2 …. X_m are input values of independent variables. This is also called an Input layer. W_1, W_2 …. W_m are the weights assigned to each input value. These weights are important since the learning rate decide which weight should be given more importance and which one should least. Then there is a functional node as described in green circle. There are four type of function mostly used: threshold function, sigmoid function, rectifier function or the hyperbolic tangent function.

Threshold function

Figure 11: Threshold function graph

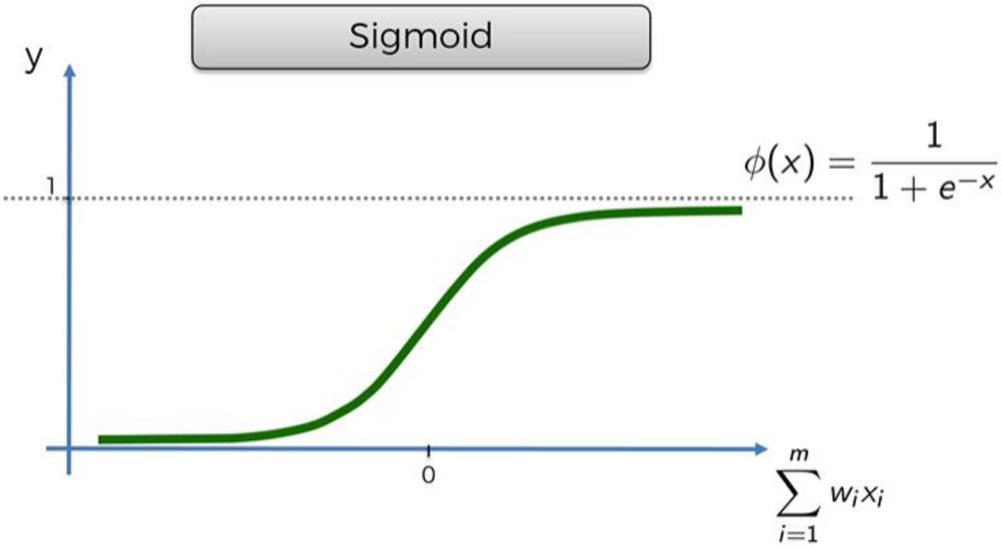


Source: https://www.udemy.com/course/machinelearning/learn/lecture/6760384#overview

Threshold function is simple and straight forward. As we can see in the above image, If weighted sum of inputs is greater than or equal to zero the output is going to be 1 and if the weighted sum is less than zero the output will be 0.

Sigmoid function

Figure 12: Sigmoid function graph

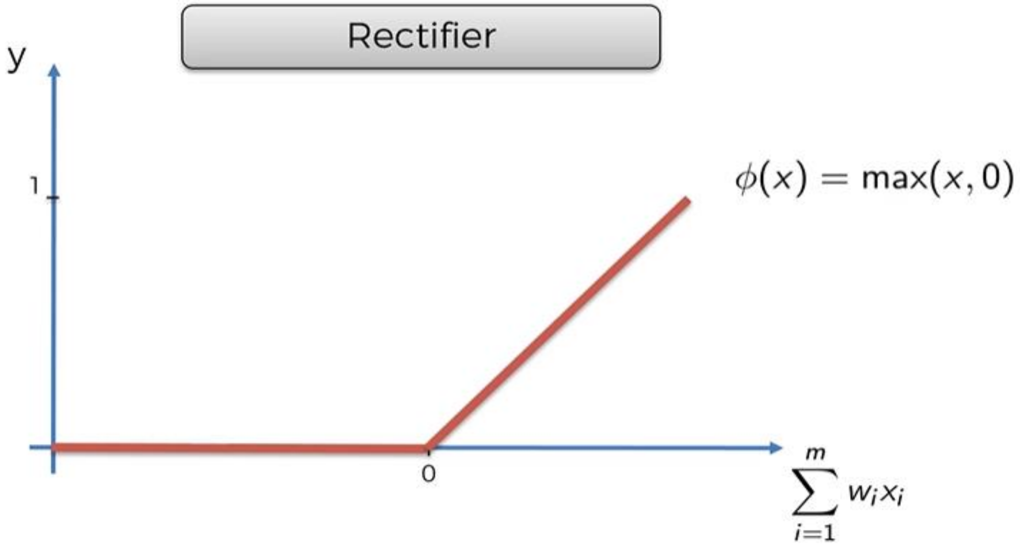


Source: https://www.udemy.com/course/machinelearning/learn/lecture/6760384#overview

We saw this function in the logistic regression also. With the help of this function we can transform any continuous value in between 0 and 1. And after specifying some threshold we can divide the values in a classes. If the value is less than some specified threshold we can assign the value to class 0 and if it is greater than and equal to threshold value than we can assign it to class 1. This function is very useful in final or output layer in a neural network.

Rectifier function

Figure 13: Rectifier function graph



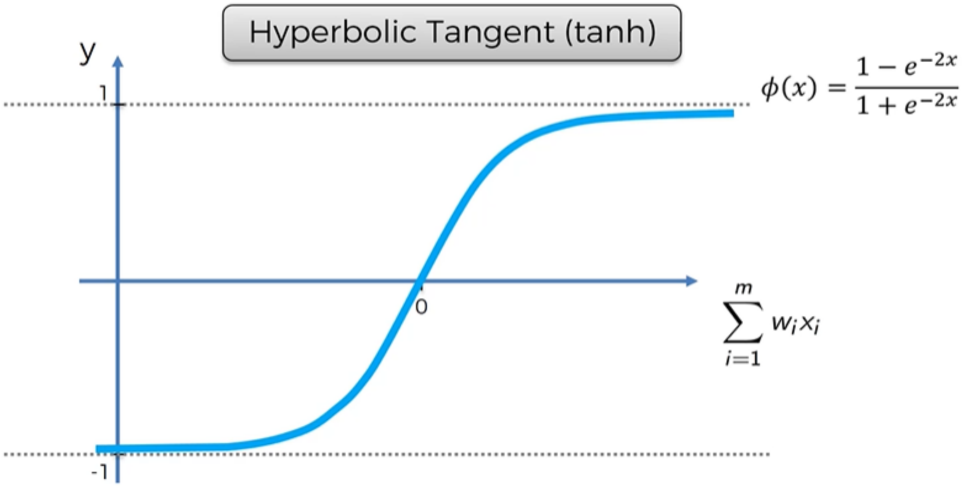
Source: https://www.udemy.com/course/machinelearning/learn/lecture/6760384#overview

This function mainly used in the hidden layers of neural network. This function will give the output of the maximum value between weighted sum or zero. If the weighted sum is greater than 0, the output will be equal to the weighted sum. And if the weighted sum is less than or equal to 0, than the output will be zero. This combination is quite common in ANN classifier when rectifier function is used in the hidden layers and the sigmoid function in output layer[15].

[15] http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf, access: 14.09.2020

Hyperbolic tangent function

Figure 14: Hyperbolic tangent function graph



Source: https://www.udemy.com/course/machinelearning/learn/lecture/6760384#overview

As seen in the above image, the hyperbolic tangent function is creating a smooth curve very similar to the curve that we saw in sigmoid function but hyperbolic tangent function generate the output in the interval from -1 to 1, which is different from the sigmoid function which restrict the output between 0 to 1

For the binary classification problem the Sigmoid or the Threshold function are the best. After passing through the functions the predicted output generated.

The steps in the Neural network algorithm using stochastic gradient descent work as follow[16]:

1. First we have to randomly assign any small values to the weights in a neural network.
2. Now take the first row of the dataset and input each attribute value in a unique node of the input layer.
3. Now we have to propagate forward (from left to right). The inputs to the Neuron functions in a second layer is the weighted sum of output values of the first layer. In the same way the inputs in the third layer is the weighted sum of output values of second layer. It goes on till we found the predicted value.
4. After predicting the value we have to calculate the error between the actual and predicted value.

[16] https://iamtrask.github.io/2015/07/27/python-network-part2/, access: 14.09.2020

5. After calculating the error we have to go back and update the weights according to how much they are contributing to the error. This will be decided by learning rate.
6. Continue the steps from 1 to 5 for all the rows in a dataset. This process is also known as reinforcement learning.
7. After all the rows in a data set pass through the neural network, the first cycle is complete. Do more cycle in order to train model better and achieve greater accuracy.

CHAPTER 3

Research Methodology

3.1 Objective of thesis

The delinquency rate of a credit card loan is one of the major source of concern among banks and lending institutes. Every year these institutions bear a heavy loss in the form of non-payment of loan back to them. The objective of thesis is to build a most accurate and robust model using machine learning algorithms for the banks or lending institutions that can able to classify the customers in to whether they will do default payment or not in the next month based on the demographic variables and the past payments history of the customers. This model will help the banks and lending institutes in the decision making during the credit card loan application.

3.2 Data set information

The data set for training and testing is collected from the UCI website. There are total 30,000 rows and 25 attributes. Each row represent the information of a single customer. The attributes consists of the demographic variables of customers and past payment history from April to September of year 2005[17].

Attributes information

The data set contains total 25 columns. The description of the columns is given below.

The “ID” column contain unique number in serial order from 1 to 30,000. The “LIMIT_BAL” column contain the information of amount of credit given to the customers in dollars. In “SEX” column the information is stored on the scale of 1 and 2. “1” value if gender is male and “2” for female. In “EDUCATION” column the information is stored on the scale of 1 to 6. “1” value if customer is graduate, “2” is assigned to university, “3” is assigned to high school, “4” is others, “5” and “6” are unknown. In “MARRIAGE” column the information is stored on the scale of 1 to 3. “1” value if person is married, “2” if person is single and “3” for others. The “AGE” column contains the age of customers.

[17] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients, access: 14.09.2020

The columns “PAY_0” to “PAY_6” contains the information of the status of repayment from April to September. The Scale used in the attributes from PAY_0 to PAY_6 of the repayment status is described as follow: “-1” indicates, pay duly; “1” indicates, payment delayed for 1 month; “2” indicates, payment delayed for 2 months, “3” indicates, payment delayed for 3 months, “4” indicates, payment delayed for 4 months, ………………… “8” indicates, payment delayed for 8 months and above.

The columns “BILL_AMT1” to “BILL_AMT6” contains the information of the amount of bill statement in dollars from April to September. The columns “PAY_AMT1” to “PAY_AMT6” contains the information of the previous payment in dollars from April to September. The last column “default.payment.next.month” contains the information whether customer will do a default payment or not in next month.

Out of total 25 attributes, the first 24 attributes are used as an explanatory variable or Independent variables and the last attribute “default.payment.next.month” will be used as target variable or the dependent variable. Since there are only 2 possible outcomes (0 = Will Pay, 1 = Will Default) in the target variable, this problem will be described as a binary classification.

3.3 Exploratory Data Analysis

First, let’s see first few rows of data to get a broad understanding of the data set.

Table 1: Overview of Data Set

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	...	PAY_AMT6	Default.payment.next.month
1	20000	2	2	1	24	2	...	0	1
2	120000	2	2	2	26	-1	...	2000	1
3	90000	2	2	2	34	0	...	5000	0
4	50000	2	2	1	37	0	...	1000	0
5	50000	1	2	1	57	-1	...	679	0

Source: own calculations based on research

As we can see in the above table, The column “ID” is randomly generated column by the lending institute and it does not have any effect on the customer ability to pay back the bill. So, I will remove this column. After that, I will check if there is any missing data in columns.

Table 2: Missing values in the Data Set

	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	...	PAY_AMT6	Default.payment.next.month
Total	0	0	0	0	0	0	...	0	0
Percent	0	0	0	0	0	0	...	0	0

Source: own calculations based on research

As we can see in above table, There is no missing value in the entire data set. Next, we will check the statistics of the first 10 attributes.

Table 3: Statistical result of first 10 columns

	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6
count	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000
mean	167484.32	1.6	1.85	1.55	35.48	-0.02	-0.13	-0.17	-0.22	-0.27	-0.29
std	129747.66	0.49	0.79	0.52	9.22	1.12	1.20	1.20	1.17	1.13	1.15
min	10000	1	0	0	21	-2	-2	-2	-2	-2	-2
25%	50000	1	1	1	28	-1	-1	-1	-1	-1	-1
50%	140000	2	2	2	34	0	0	0	0	0	0
75%	240000	2	2	2	41	0	0	0	0	0	0
max	1000000	2	6	3	79	8	8	8	8	8	8

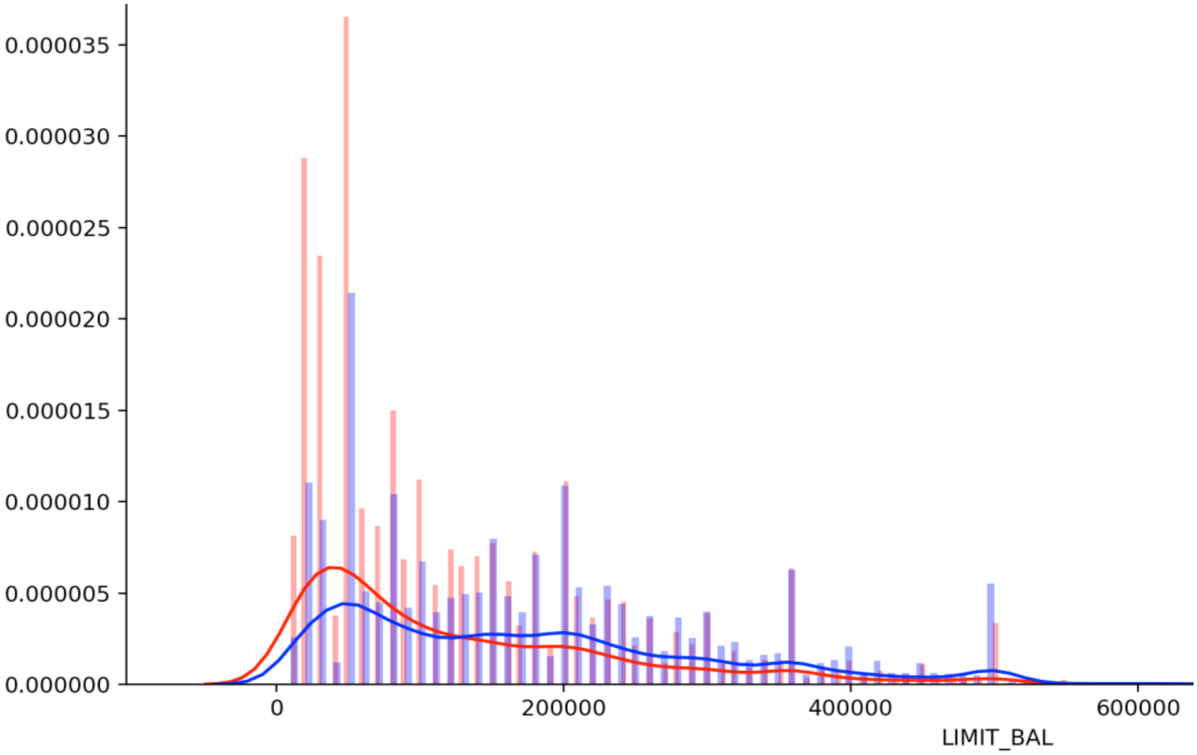
Source: own calculations based on research

In column “EDUCATION” we see the minimum value is 0 which is not defined. And also 5 and 6 value are unknown. So I am assuming 0, 5 and 6 values as same as 4 which is described as “other” in attribute information. In column “MARRIAGE” we see a minimum value 0 again which is not defined. So, I am assuming the 0 value as same as 3 which is described as “other” in the attribute information.

In columns from “PAY_0” to “PAY_6” we see minimum value -2 which is not defined. The -1 in these columns represents the payment done by a customer on time. 1 represents payment delay by a customer one month, 2 represents payment delay by a customer two month and so on. So I can safely assume any negative values in these columns as 0 means all those customers pay duly in a given month. I will change the name of column “PAY_0” to PAY_1, just it make it more consistent with other columns.

The column “LIMIT_BAL” (amount of credit given to the customers) can be a good criterion of weather customer will do a default or not. The customers with good credit score have given more credit line amount than in comparison to the customers whose credit score is not good. On the other hand, the customers who defaulted in past will find hard to raise credit line.

Figure 15: Graph of number of non-defaulters and defaults in a specific LIMIT_BAL group

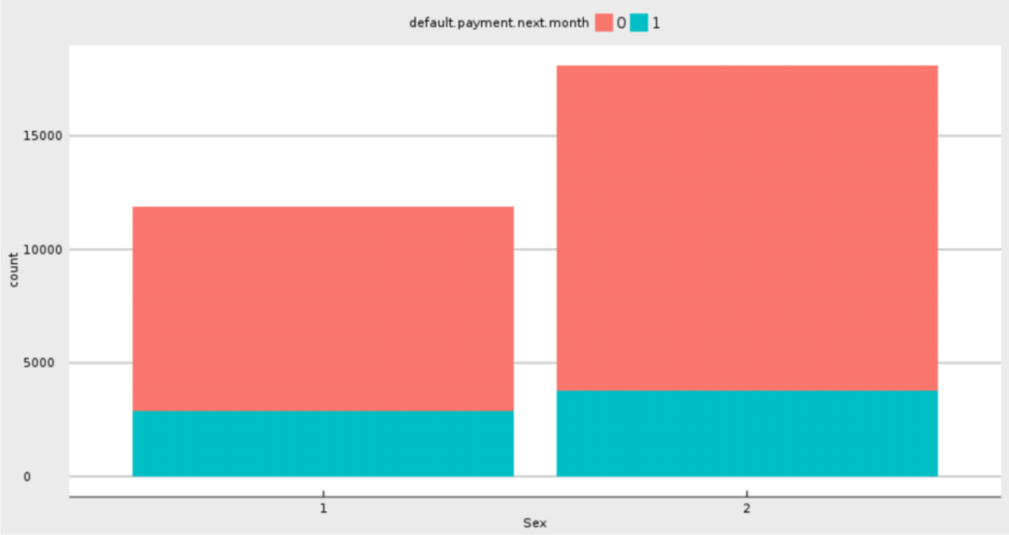


Source: own figure based on research

In the plot above, the pink line is showing the total number of defaulters in a specific LIMIT_BAL group and the blue line are number of non-defaulters. This plot clearly showing the relation between the number of defaulters and the amount of credit given. The customers are doing more default when the credit limit is below the average than the customers who credit limit is above the average. On some of the low credit limit, the number of defaulters are almost double than non-defaulters. As we go beyond the average credit limit the number of non-defaulters getting increase and in every credit limit group the number of non-defaulters are more than the defaulters.

Next I am checking that if there is any impact of gender on default payment.

Figure 16: Gender distribution in default payment

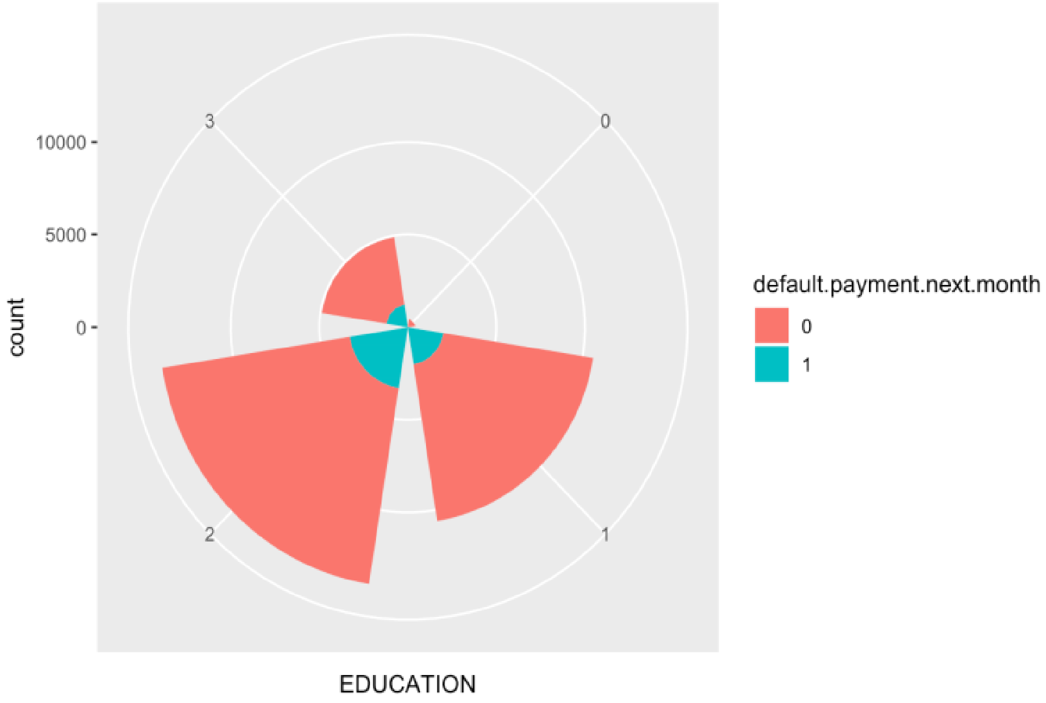


Source: own figure based on research

As we can see in plot, female are more than male customers. Subsequently, the female defaulters are slightly higher than the male defaulters. Well it is hard to find any conclusive reason that the gender had an impact on the default payment. We are still going to add the SEX column in a data set to train the model.

Next I am checking whether education of customer has any impact on default payment.

Figure 17: Education column vs default payment next month



Source: own figure based on research

As we can see in chart, the education has no significant impact on the default payment. The customers irrespective of education, showing almost same proportion of default payment. Well in starting I assumed, the customers with higher degree tend to pay their due loan on time. If people have higher degree then most likely they have a job and are earning more than in comparison to the people who are less educated.

3.4 Data preprocessing

In this data set there are 3 columns which are categorical features namely “EDUCATION”, “SEX”, and “MARRIAGE”. We have to convert all these variables in to dummy variables before we put our data in to our classifications algorithms. The reason is, the values in these columns are representing category. Let’s understand this with an example. In “MARRIAGE” column the 1 value represent that person is “married”, 2 represent that person is “single” and the 3 value represent “others” means that person can be divorced or maybe he/she don’t want to tell his marital status etc. Here 1, 2 and 3 are not greater or less than each other. There is no order between these numbers. These numbers are just representing marital status of a person. But if we put this data in to model without modified, the model will think it as Numerical value and not the categorical value.

So to avoid this problem, we can create a separate column of each category known as dummy variable in the data set. The “MARRIAGE” column will get replaced by 3 columns namely “married”, “single” and “others”. The information stored in these columns as follow: If a person marital status is “single”, then we assign 1 value to column “single” and 0 value to other two columns. If there is another person who is “married” than we assign 1 value to column “married” and 0 value to “single” and “others”. Like the same way we will assign the values to other customers[18].

Feature Scaling

Feature scaling is one of the important feature that we have to perform on our data set before putting the data in to classification algorithms. In feature scaling, we re-scale all the columns in a data set except the output column so that the data in this columns follow normality. This method is also called data standardization technique.

$$\boldsymbol{\mu} = 0 \text{ and } \boldsymbol{\sigma} = 1$$

Here, $\boldsymbol{\mu}$ represent the Mean and $\boldsymbol{\sigma}$ represent standard deviation. In standard normal distribution, the curve is symmetrical and the mean is centered at 0 and the spread is determined by the standard deviation of 1. The standard score **z** is determined by this formula[19]:

$$\mathbf{z} = (\mathbf{x} - \boldsymbol{\mu}) / \boldsymbol{\sigma}$$

[18] https://stattrek.com/multiple-regression/dummy-variables.aspx, access: 14.09.2020
[19] https://www.geeksforgeeks.org/ml-feature-scaling-part-2/?ref=lbp, access: 14.09.2020

3.5 Machine learning models analysis

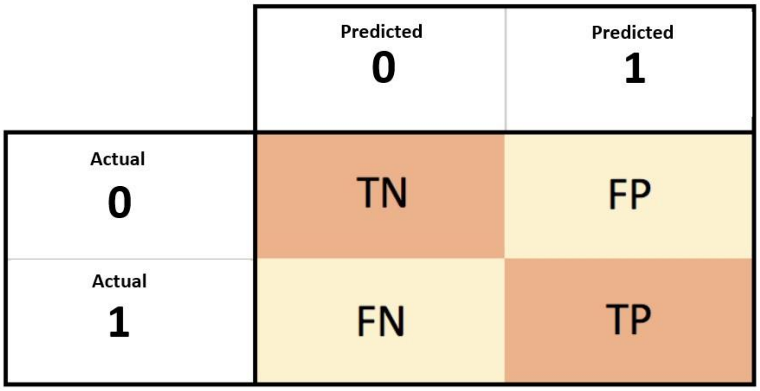
In order to test my models performance, I will divide the whole data in to 80% – 20% split. The 80% data I will use in training the classification models and the rest 20% data for testing the model performance. As we know there are total 30,000 rows in our data set. The 24,000 rows will be used in training the model and the rest 6,000 for testing.

I used the train_test_split function from sklearn machine learning library to split the data set. The X is the independent features data frame and the y is the dependent feature or output variable. The test_size is showing the percentage of data you are separating for testing the model from the whole data set. The random_state parameter decide which data rows will be assigned to training data and testing data. Every different value in random_state will give you the different training data and testing data[20].

There are two important tool that I am going to use in evaluating the accuracy and the performance of models. These are confusion matrix and the ROC curve (Receiver operating characteristic curve). The confusion matrix gave some important measures like Accuracy, Recall, Precision etc. and also from confusion matrix we create an ROC curve another major tool in evaluating the classification models. Let’s discuss more about these tools.

Confusion Matrix

Figure 18: Confusion matrix table



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Source: https://towardsdatascience.com/demystifying-confusion-matrix-29f3037b0cfa

[20] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html, access: 14.09.2020

The above confusion matrix is used to evaluate binary classification problem. There are only two possible outputs: 0 and 1. In reference to our data set, the 0 represents the customer will “not default or pay duly” and 1 represents that they will “default” next payment.

Let’s discuss the terms TN, TP, FP, and FN in detail[21]:

True Negative (TN): The actual output was 0, and also we predicted the 0 output. In reference to our problem, The customer who will “not default or pay duly” was also predicted as “non-defaulter”.

True Positive (TP): The actual output was 1, and also we predicted the 1 output. In reference to our problem, The customer who will “default” was also predicted as “defaulter”.

False Positive (FP): The actual output was 0, but we predicted the 1. In reference to our problem, The customer who will pay predicted as defaulter. False Positive are also called Type 1 error.

False Negative (FN): The actual output was 1, but we predicted the 0. In reference to our problem, The customer who will default was predicted as payer. False Negative are also called Type 2 error.

With the help of confusion matrix we can calculate many important measures of classification models such as Accuracy, Precision, Specificity, Recall also called as sensitivity and most important ROC-AUC curve. Let’s discuss these terms in detail.

Accuracy: It tells how many output values we predicted correctly out of total output values. From confusion matrix it can calculated by the following formula,

$$\text{Accuracy} = (TN + TP) / (TN + TP + FP + FN)$$

Recall/Sensitivity: It is the proportion of True positive in a confusion matrix to the total actual positive values. This can be predicted by following formula,

$$\text{Recall} = TP / (FN + TP)$$

[21] https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/, access: 14.09.2020

Specificity: It is the proportion of True negative to the total actual negative values. This can be predicted by following formula,

Specificity = TN / (FP + TN)

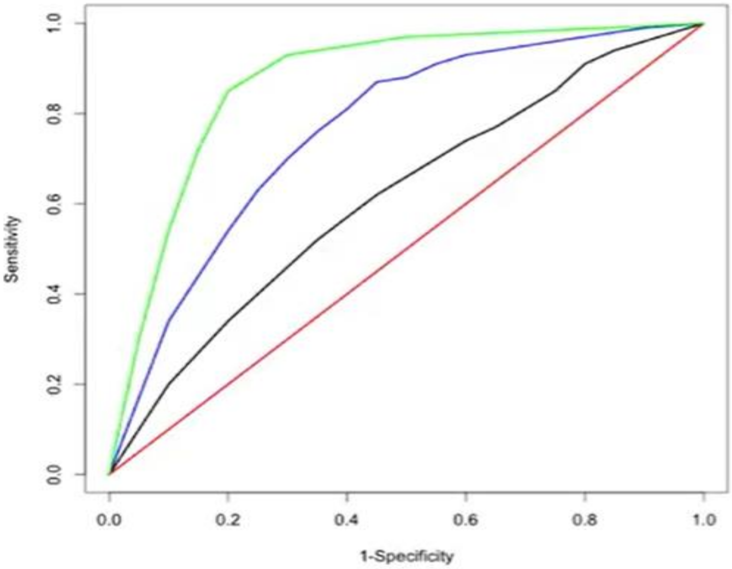
Precision: It is the proportion of True positive to the total predicted positive values. This can be predicted by following formula,

Precision = TP / (TP + FP)

ROC-AUC Curve

The ROC curve is constructed by plotting the sensitivity against (1-specificity) as shown below.

Figure 19: ROC curve



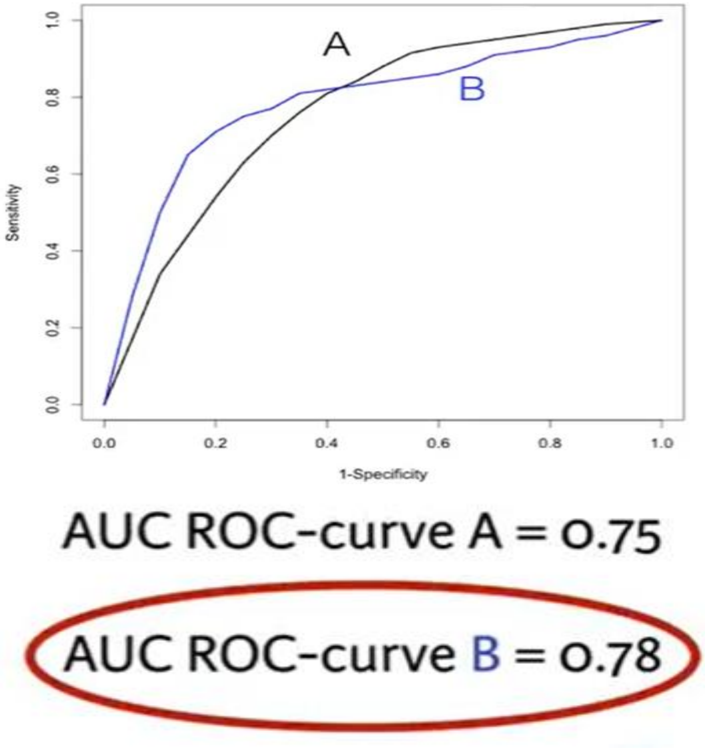
Source: https://www.youtube.com/watch?v=MUCo7NvB9SI&t=1s

The curve always starts in the low left corner where sensitivity is zero and specificity is one corresponding to a cut-off of 1. The other end of the curve, in the top right corner corresponds to cut-off equal to 0 where sensitivity is equal to one and specificity is equal to zero.

The curve which is closer to the left top corner is the best curve among all the other curves that are displayed on graph. So, in the above figure the green curve model will be considered as best[22].

When comparing classification models you will often see that the differences between several ROC curves are not as clear as shown in figure below. Sometimes, it hard to tell that which model is better.

Figure 20: Overlapping in ROC curves between two different models



Source: https://www.youtube.com/watch?v=MUCo7NvB9SI&t=1s

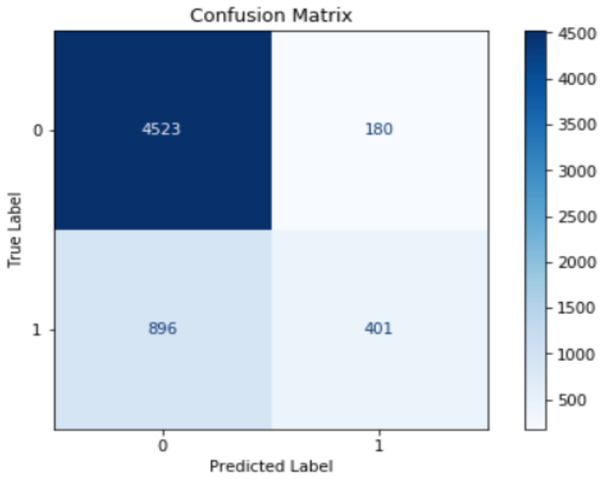
A black ROC curve is from model A and the blue ROC curve from model B. The measure that often use when comparing ROC curve in this situation is the so called Area under the curve (AUC). The AUC of a model is between “0.5” which corresponds to the red line that I showed in previous page and “1” which corresponds to a perfect model. Computing the AUC here, we see that the model B leads to a higher AUC and should be preferred over Model A.

[22] https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5, access: 14.09.2020

3.5.1 Logistic Regression

Logistic regression was easiest to implement in our classification models. As seen in the confusion matrix below, logistic regression successfully managed to predict 4924 correct predictions out of total 6000 test data set with a classification accuracy of 0.8207 or 82.07% which is quite good. Out of 4924 correct predictions, 4523 are True Negative and 401 as True positive. The Recall (also known as Sensitivity) for the test data set is 30.92% and Precision is 69.02%.

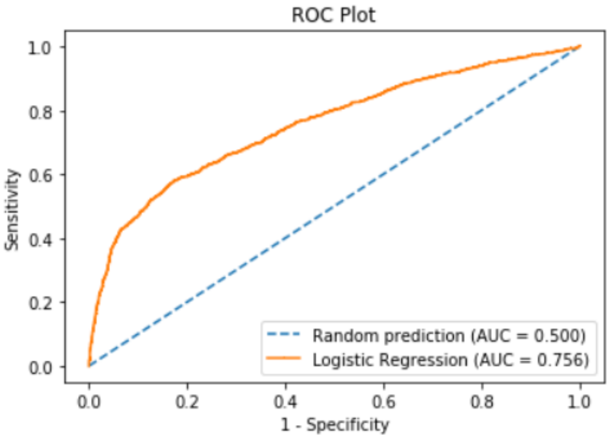
Figure 21: Logistic regression confusion matrix



Source: own figure based on research

The ROC curve plot of the logistic Regression model is given below. As we can see we got an Area under curve (AUC) value 0.756. Higher the value of AUC, better the model.

Figure 22: ROC curve plot of logistic regression

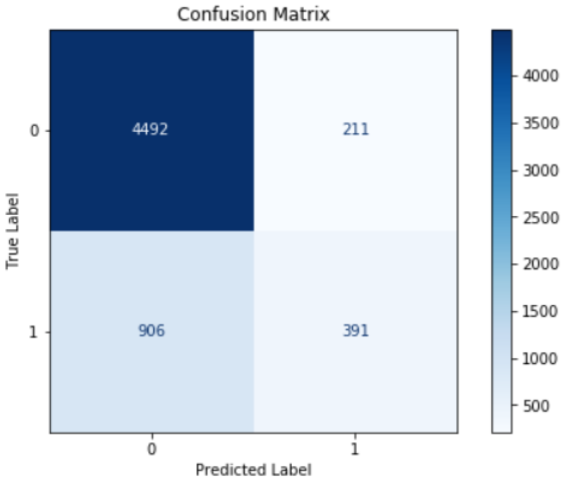


Source: own figure based on research

3.5.2 K-Nearest Neighbor classifier

As we discussed in KNN algorithm methodology part, we have to select K number of neighbor to implement this model. After carefully testing the model on many different K values, I found K = 24 gave me optimal classification accuracy with 81.38% which is little less to the accuracy of logistic regression. The KNN classifier successfully managed to predict 4883 observations out of 6000 test data set as shown in confusion matrix below. Out of 4883 correct predictions, 4492 predicted as True Negative and 391 as True positive. The Recall value is calculated as 30.15% and Precision is 64.95%.

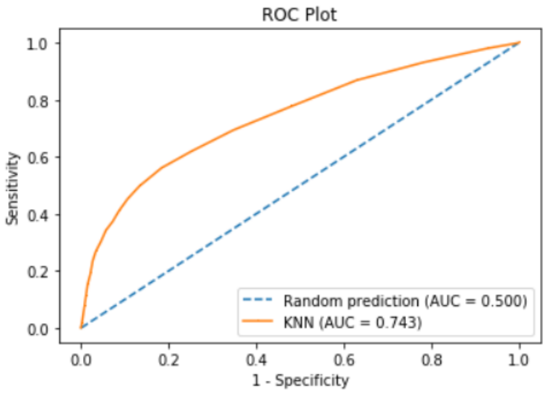
Figure 23: K-Nearest Neighbor classifier confusion matrix



Source: own figure based on research

The ROC curve of KNN classifier is given below. As we can see, we got an AUC value 0.743, a little less than the value we got in logistic regression. So till now, the logistic regression model performed better than KNN classifier.

Figure 24: ROC curve plot of K-Nearest Neighbor classifier

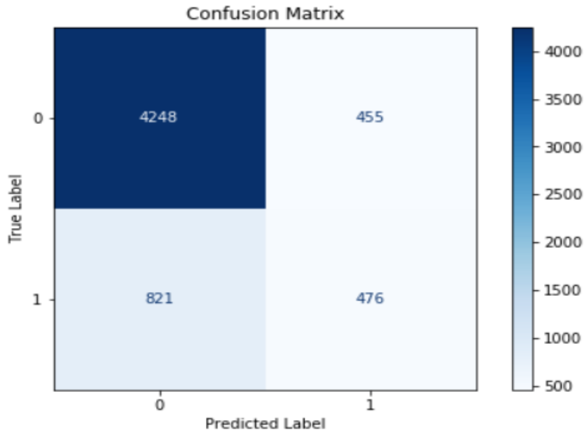


Source: own figure based on research

3.5.3 Decision Tree classifier

The third model that we are using to check the accuracy is decision tree classifier. The confusion matrix of decision tree is shown below. Out of total 6000 test observations, decision tree model predicted 4724 correctly with accuracy of 78.73% a less than from what we got in KNN and in logistic regression. Out of 4248 correct predictions, 4248 predicted as True Negative and 476 as True positive. The Recall value is calculated as 36.70% and Precision is 51.13%.

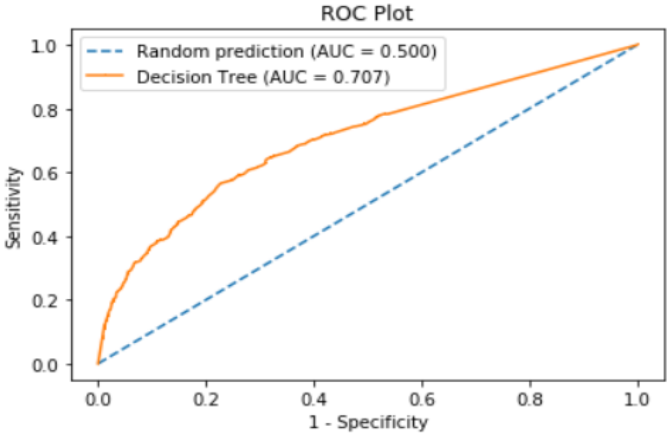
Figure 25: Decision tree classifier confusion matrix



Source: own figure based on research

The ROC curve of the decision tree model is shown below. The AUC is calculated as 0.707 less than in comparison to KNN and logistic regression.

Figure 26: ROC curve plot of decision tree classifier

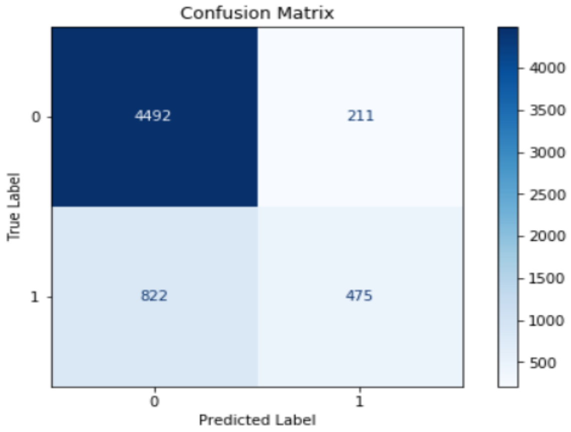


Source: own figure based on research

3.5.4 Random Forest classifier

The random forest classifier is one of the most powerful classification algorithm. Instead of one decision tree algorithm to predict output, we can take several decision tree to make it random forest, where each tree is predicting their own output on the test data set. After that the new data point is assigned to output class (1 or 0) who win the majority of vote. The confusion matrix of random matrix is shown below. This model successfully managed to predict 4967 observations out of 6000 data set with a classification accuracy of 82.78%. Out of 4927 correct predictions, 4492 predicted as True Negative and 475 as True positive. The Recall value is calculated as 36.62% and Precision is 69.24%.

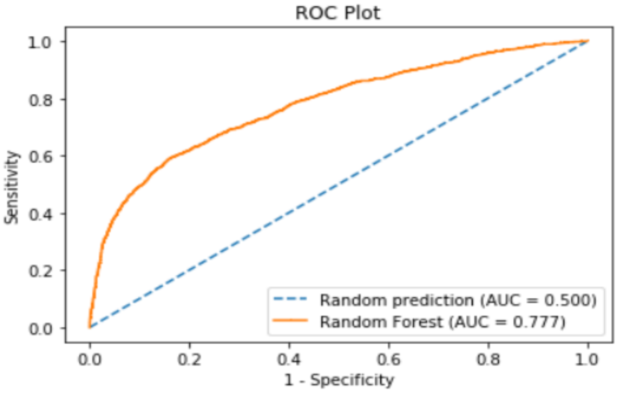
Figure 27: Random forest classifier confusion matrix



Source: own figure based on research

The ROC curve of random forest model is shown below. The AUC is calculated as 0.777, Higher than all the models that we analyzed till now.

Figure 28: ROC curve plot of random forest classifier

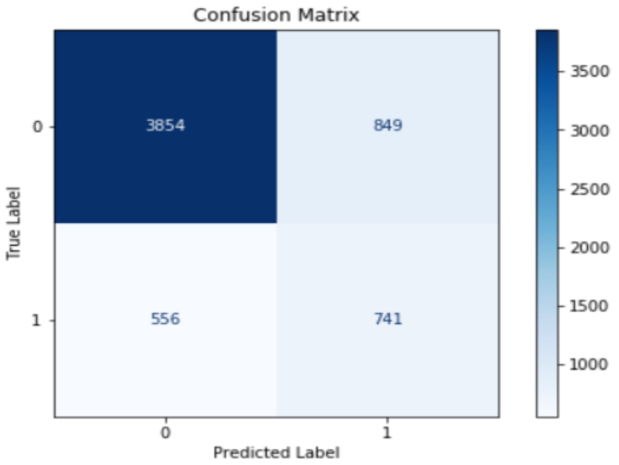


Source: own figure based on research

3.5.5 Naive Bayes classifier

Next model that I implemented is the naive bayes classifier. The confusion matrix of Naive Bayes is shown below. Naive Bayes predicted 4595 correct output out of 6000 test data set. The classification accuracy we got is 76.58% less than all the models that we implemented till now. Out of 4595 correct predictions, 3854 predicted as True Negative and 741 as True positive. The Recall value is calculated as 57.13% and Precision is 46.60%.

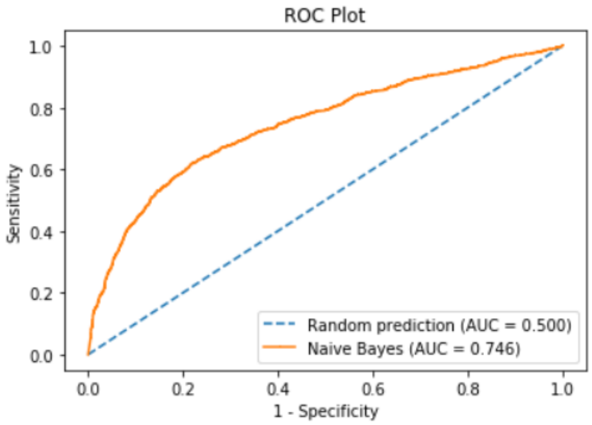
Figure 29: Naive Bayes classifier confusion matrix



Source: own figure based on research

The ROC curve of Naive Bayes model is shown below. The AUC is calculated as 0.746.

Figure 30: ROC curve plot of Naive Bayes classifier

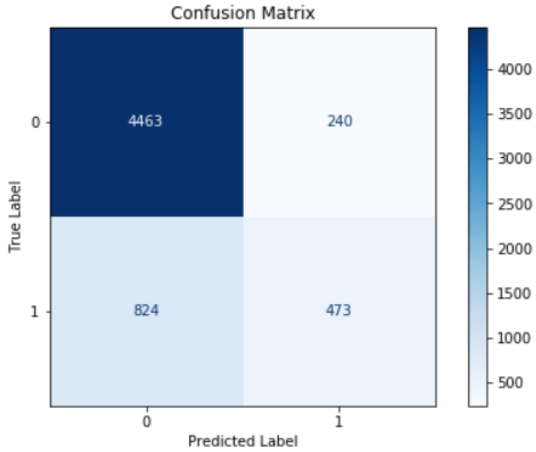


Source: own figure based on research

3.5.6 Neural Network classifier

The last model that I implemented is Neural Network classifier. I used stochastic gradient descent method that I discussed in methodology part to train neural network. The Neural network manage to predict 4936 correct output out of 6000 test data set with a classification accuracy of 82.27% which is second best after Random forest model. Out of 4936 correct predictions, 4463 predicted as True Negative and 473 as True positive. The Recall value is calculated as 36.47% and Precision is 66.34%.

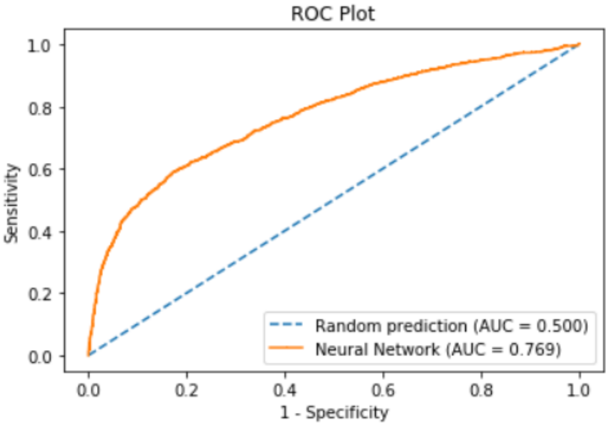
Figure 31: Neural network classifier confusion matrix



Source: own figure based on research

The ROC curve of the Neural network model is shown below. The AUC is calculated as 0.769 again second best after Random forest.

Figure 32: ROC curve plot of Neural network classifier

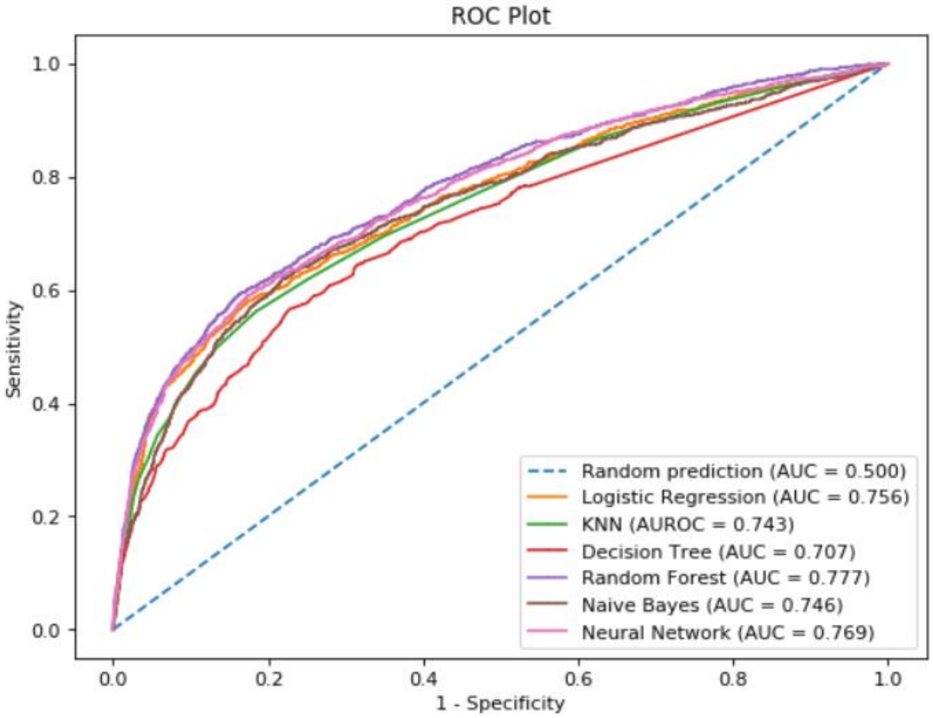


Source: own figure based on research

3.6 Models Comparison

Let's display all the ROC curve of different models that we discussed on the same graph as shown below. As we discussed previously in ROC curve section, the closer the curve to the top left corner, the better model curve it is. If there is overlapping happen at some points between models than in this case the value of AUC come in to rescue. Higher the AUC value of model, the better it is.

Figure 33: Comparison of ROC curve and AUC value of six different classification models



Source: own figure based on research

Table 4. Matrices Table

	Logistic Regression	K-Nearest Neighbor	Decision Tree	Random Forest	Naive Bayes	Neural Network
AUC	0.756	0.743	0.707	**0.777**	0.746	**0.769**
Accuracy	82.07%	81.38%	78.73%	**82.78%**	76.58%	**82.27%**
Recall	30.92%	30.15%	36.70%	36.62%	57.13%	36.47%
Precision	69.02%	64.95%	51.13%	69.24%	46.60%	66.34%

Source: own calculations based on research

As we can see the decision tree model performs worst among all the models with AUC value only 0.707. The Logistic regression, KNN and Naive Bayes are overlapping at many point and we can't see clearly which model performs better. If you see the values of AUC, among these 3 models, the logistic regression performed best with AUC value 0.756 followed by Naive Bayes (0.746) and KNN (0.743).

Now we see the two best performing models which are Random Forest and Neural Network. As seen in the plot, we can't see it clearly that which model perform better. If we talked about the AUC value, Random Forest with AUC value 0.777 performs slightly better than the Neural network with AUC value 0.769.

CONCLUSION

In this research paper we analyzed and compared six most frequently used machine learning classification algorithms, namely Logistic regression, K-Nearest neighbor, Decision tree, Random forest, Naive Bayes and Neural network classifier on the credit card customers data set. The research problem is to build most accurate and robust model that can able to predict, weather a customer will do a default payment or not in next month based on the demographic variable of the customers and their past payment behavior of six months.

First, I performed Exploratory analysis on data to get a broad understanding of the data set. This step is crucial before we trained our data in algorithms and also I found some irregularity in the data set. Even though, there was no missing value in the data set, I found some values in columns which are not defined. These irregularities, if not treated properly may seriously affect the accuracy of the models later. Next, I found one interesting visualization, which was clearly showing the relation between the defaulters and amount of credit given to them. The customers with low credit limit are doing more default payment than in comparison to the customers with higher credit limit.

Next, I converted and replaced all the categorical attributes in the data set in to their respective dummy variables. Then, I scaled all the attributes and make them follow standard normal distribution using feature scaling data standardization technique.

Next, I divided the whole data set in training data and testing data. 80% of the data will be used in training the model and the rest 20% to test the validity of model.

As we saw in Models comparison section, Random forest classifier has highest area under the curve (AUC) value among all the models followed by Neural network classifier. The ROC-AUC value is a robust method in evaluating the classification models. Higher the ROC-AUC value, better the model is. Even the classification accuracy of Random forest is highest among all the models. The random forest classifier successfully managed to predict the sample test data of default or non-default payment of credit card customers in next month at an accuracy of 82.78%. The next model that manged to predict accurately is Neural Network classifier. The difference between the Random Forest classifier and Neural Network classifier in term AUC and accuracy value is really small.

Among all the model, the decision tree performed worst with AUC value only 0.707, far less than the value we achieved in our best model. The logistic regression model performed satisfactorily with an accuracy of more than 82% and an AUC (0.753).

So among all the models we can conclude that the Random Forest model performs best in predicting the default or non-default of a credit card customers in next month. The lending institutes and banks all around the world can use this model in risk predictive modeling and decrease the delinquency rate of credit card loan.

REFERENCES

1. Python for Data Analysis: Data Wrangling 2nd Edition by Wes McKinney
2. Supervised machine learning algorithms: classification and comparison by Osisanwo F.Y., (IJCTT) – Volume 48 Number 3 June 2017
3. Storytelling with data by cole nussbaumer knaflic published by wiley
4. Python for Finance: Mastering Data-Driven Finance 2nd Edition by Yves Hilpisch

ELECTRONIC SOURCES

5. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
6. https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
7. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
8. https://www.statista.com/statistics/935115/credit-card-loan-delinquency-rates-usa/
9. https://www.udemy.com/course/machinelearning/
10. https://www.coursera.org/learn/machine-learning-with-python/home/welcome
11. https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/
12. https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/
13. https://stattrek.com/multiple-regression/dummy-variables.aspx
14. https://www.geeksforgeeks.org/ml-feature-scaling-part-2/?ref=lbp

LIST OF TABLES

LIST OF FIGURES