BENGAL SCHOOL OF TECHNOLOGY & MANAGEMENT

Department of BCA, 6th Sem

SESSION (2017-2020)

LOAN PREDICTION SYSTEM

MAJOR PROJECT REPORT

SUBMITTED BY

SAYANDIP ADHIKARY

26601217012

SURAJ SHARMA

26601217004

TANMOY ADHIKARY

26601217002

SUBMITTED TO

Mr. SATANU MAITY

Acknowledgement

We have great pleasure in expressing our gratitude to our respected Principal Sir, Dr. Manas Roy of Bengal School of Technology and Management for his kind support in providing us an opportunity to do research in this college. And we are really grateful to Satanu Maity Sir for his valuable time and effort towards us in this project. We thank our respected class teacher Arup Mitra Sir for his concern towards us for the completion of our project earlier. We want to thank again our respected Samit Mondal Sir for encouraging us for this project and doing every possible ways like building a training environment in our collage, providing us internet services etc. And at last but not the least we are immensely grateful to our other BCA Faculty members for their helpful nature towards us.

Abstract

With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So, in this paper we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this paper is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four sections (i)Data Collection (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing Keywords - Loan, Machine Learning, Training, Testing, Prediction.

TABLE OF CONTENTS

List of figures

1.Introduction

1.1 Background

A recent development of machine learning techniques and data mining has led to an interest of implementing these techniques in various fields. The banking sector is no exclusion and the increasing requirements towards financial institutions to have robust risk management has led to an interest of developing current methods of risk estimation. Potentially, the implementation of machine learning techniques could lead to better quantification of the financial risks that banks are exposed to. Within the credit risk area, there has been a continuous development of the Basel accords, which provides frameworks for supervisory standards and risk management techniques as a guideline for banks to manage and quantify their risks. From Basel II, two approaches are presented for quantifying the minimum capital requirement such as the standardized approach and the internal ratings based approach (IRB) . There are different risk measures banks consider in order to estimate the potential loss they may carry in future. One of these measures is the expected loss (EL) a bank would carry in case of a defaulted customer. One of the components involved in EL-estimation is the probability if a certain customer will default or not. Customers in default means that they did not meet their contractual obligations and potentially might not be able to repay their loans . Thus, there is an interest of acquiring a model that can predict defaulted customers. A technique that is widely used for estimating the probability of client default is Logistic Regression . In this thesis, a set of machine learning methods will be investigated and studied in order to test if they can challenge the traditionally applied techniques.

1.2 Purpose

The objective of this thesis is to investigate which method from a chosen set of machine learning techniques performs the best default prediction. The research question is the following

- For a chosen set of machine learning techniques, which technique exhibits the best performance in default prediction with regards to a specific model evaluation metric?

2. Theory

2.1 Formulation of binary classification problem

Binary classification refers to the case when the input to a model is classified to belong to one of two chosen categories. In this project, customers belong either to the non-default category or to the default category. The categories can therefore be modeled as a binary random variable $Y \in \{0, 1\}$, where 0 is defined as non-default, while 1 corresponds to default. The random variable Yi is the target variable and will take the value of yi , where i corresponds to the ith observation in the data set. For some methods, the variable $\bar{}yi = 2yi-1$ will be used, since these methods require the response variable to take the values $\bar{}yi \in \{-1, 1\}$. The rest of the information about the customers, such as the products the customers posses, account balances and payments in arrears can be modeled as the input variables. These variables are both real numbers and categories and are often referred to as features or predictors. Let Xi $\in$ R p denote a real valued random input vector and an observed feature vector be represented by xi = [xi1, xi2, ..., xip] >, where p is the total number of features. Then the observation data set with N samples can be expressed as D = {(x1, y1),(x2, y2), ...,(xN , yN)}. With this setup, it makes it feasible to fit a supervised machine learning model that relates the response to the features, with the objective of accurately predicting the response for future observations [14]. The main characteristic of supervised machine learning is that the target variable is known and therefore an inference between the target variable and the predictors can be made. In contrast, unsupervised machine learning deals with the challenge where the predictors are measured but the target variable is unknown.

The chosen classification methods in this project are Logistic Regression, Decision Tree, Random Forest, KNN , SVM , K-SVM.

2.2 Logistic Regression

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the ‘Sigmoid function’ or also known as the ‘logistic function’ instead of a linear function. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

$$0 \leq h_\theta(x) \leq 1$$

Logistic regression hypothesis expectation.

2.3 Decision Tree Algorithm

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a treestructured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions:-It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm .A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees. Below diagram explains the general structure of a decision tree:

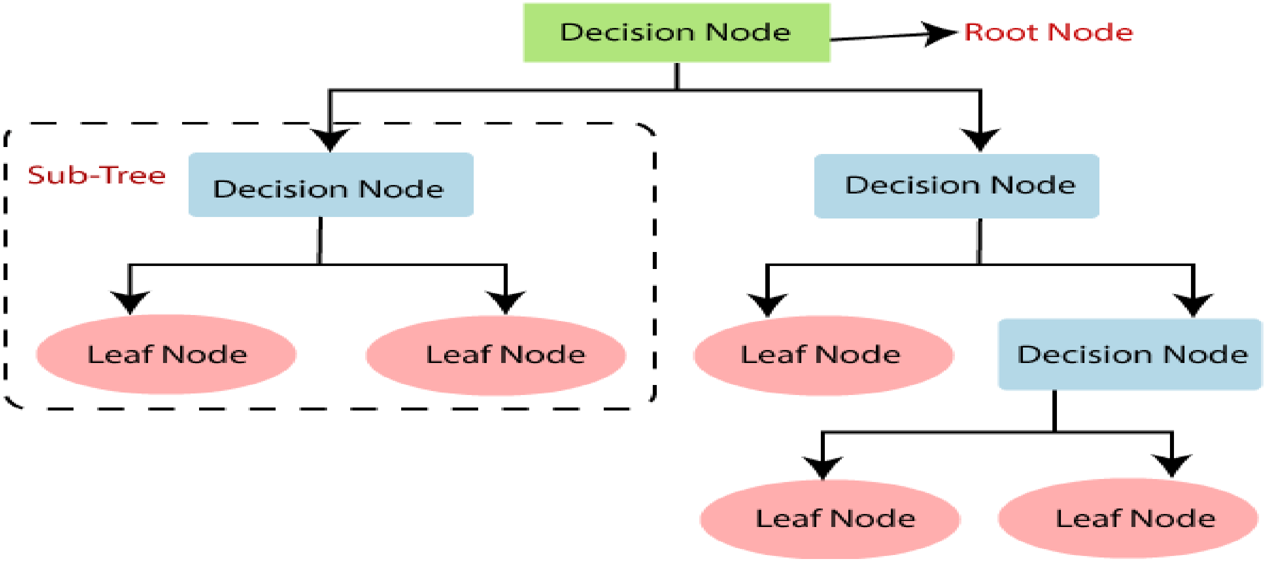


Figure 1: Decision Tree Algorithm

Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand .The logic behind the decision tree can be easily understood because it shows a tree-like structure.

Decision Tree Terminologies:

Root Node: Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

Leaf Node: Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

Splitting: Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.

Branch/Sub Tree: A tree formed by splitting the tree.

Pruning: Pruning is the process of removing the unwanted branches from the tree.

Parent/Child node: The root node of the tree is called the parent node, and other nodes are called the child nodes.

2.4 Random Forest Algorithm

Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach .The author gives 4 links to help people who are working with decision trees for the first time to learn it, and understand it well. The decision tree is a decision support tool. It uses a tree-like graph to show the possible consequences. If you input a training dataset with targets and features into the decision tree, it will formulate some set of rules. These rules can be used to perform predictions. The author uses one example to illustrate this point: suppose you want to predict whether your daughter will like an animated movie, you should collect the past animated movies she likes, and take some features as the input. Then, through the decision tree algorithm, you can generate the rules. You can then input the features of this movie and see whether it will be liked by your daughter. The process of calculating these nodes and forming the rules is using information gain and Gini index calculations.

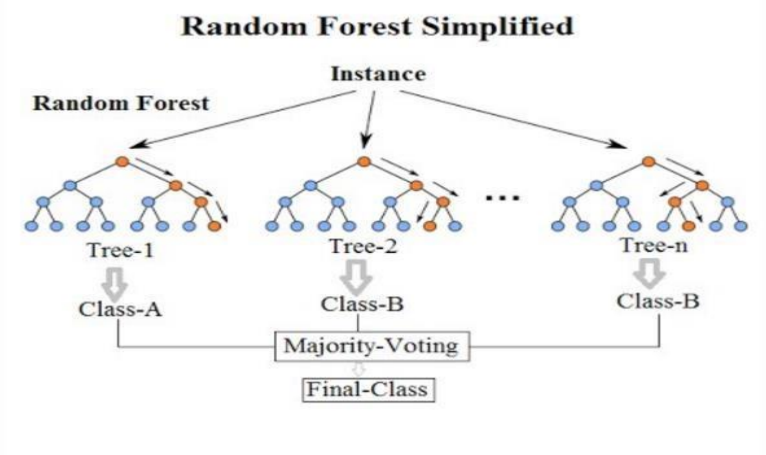
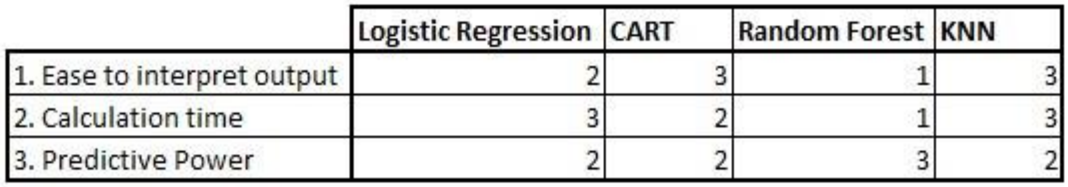


Figure 2: Random Forest Algorithm

2.5 K-Nearest Neighbors Algorithm

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:1. Ease to interpret output2. Calculation time3. Predictive Power Let us take a few examples to place KNN in the scale :



	Logistic Regression	CART	Random Forest	KNN
1. Ease to interpret output	2	3	1	3
2. Calculation time	3	2	1	3
3. Predictive Power	2	2	3	2

Figure 3: K-Nearest Neighbor Chart

KNN algorithm fairs across all parameters of considerations. It is commonly used for its easy of interpretation and low calculation time. KNN algorithm working method:Let's take a simple case to understand this algorithm. Following is a spread of red circles (RC) and green squares (GS) :

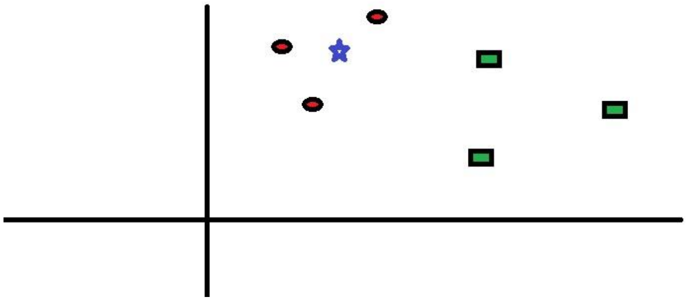


Figure 4: K-Nearest Neighbor Algorithm 1

You intend to find out the class of the blue star (BS). BS can either be RC or GS and nothing else.

The "K" is KNN algorithm is the nearest neighbor we wish to take the vote from. Let's say K = 3. Hence, we will now make a circle with BS as the center just as big as to enclose only three datapoints on the plane. Refer to the following diagram for more details:

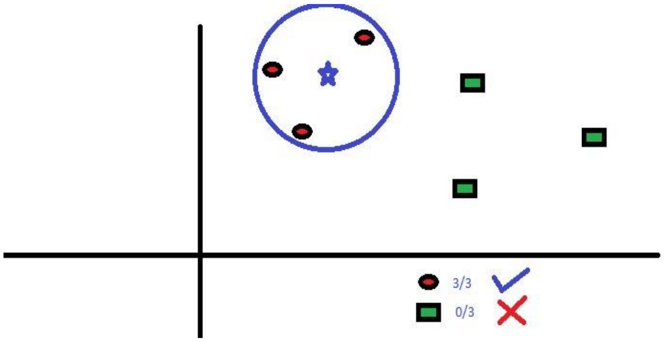


Figure 5: K-Nearest Neighbor Algorithm 2

The three closest points to BS is all RC. Hence, with a good confidence level, we can say that the BS should belong to the class RC. Here, the choice became very obvious as all three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm.

Next, we will understand what are the factors to be considered to conclude the best K.

2.6 Support Vector Machine Algorithm

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).

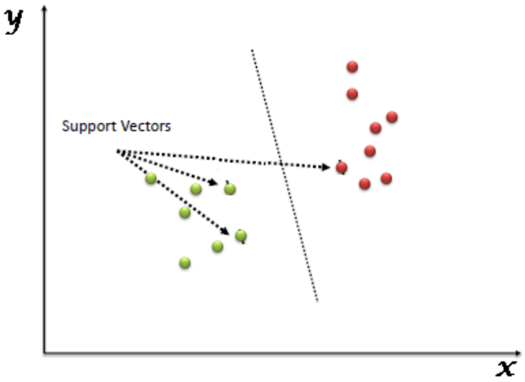


Figure 6:

SEQUENCE DIAGRAM

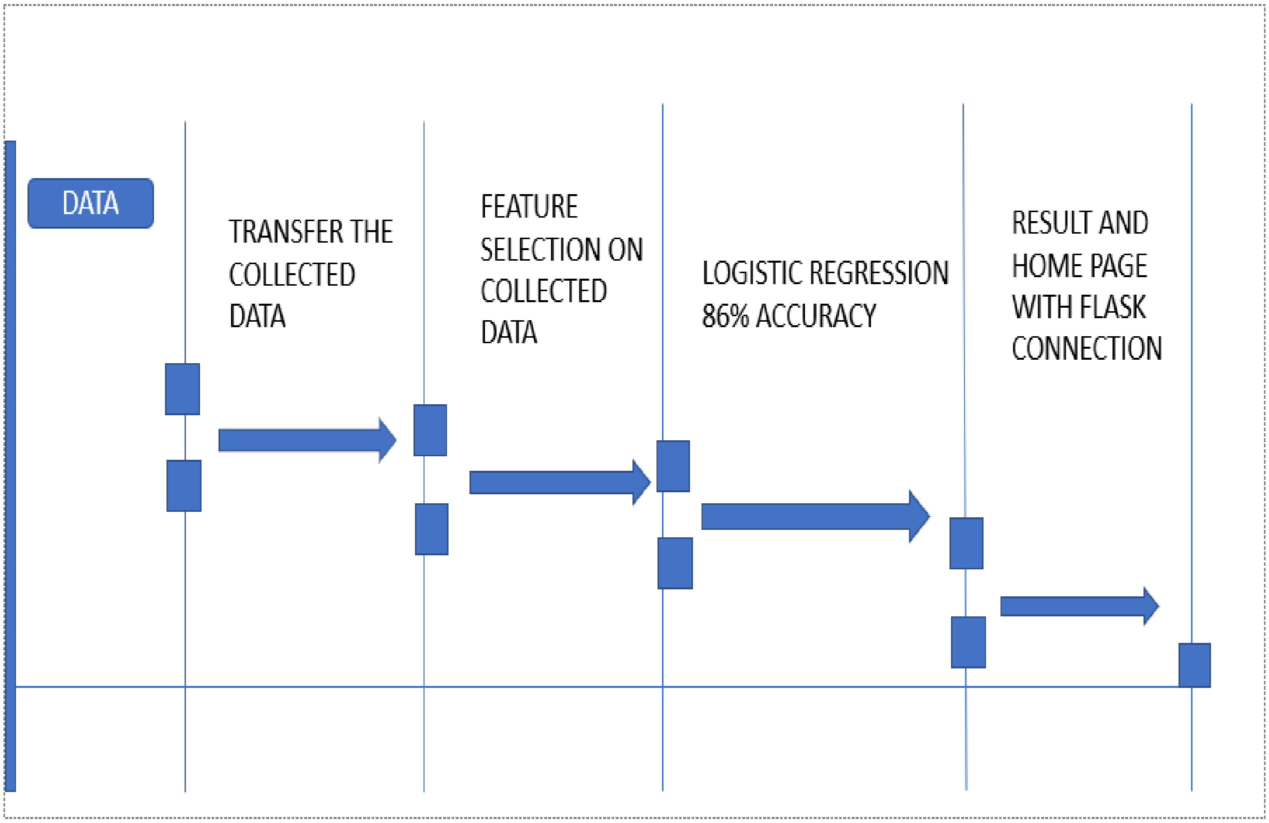

DATA
TRANSFER THE COLLECTED DATA
FEATURE SELECTION ON COLLECTED DATA
LOGISTIC REGRESSION 86% ACCURACY
RESULT AND HOME PAGE WITH FLASK CONNECTION

Figure 7

3. Methodology

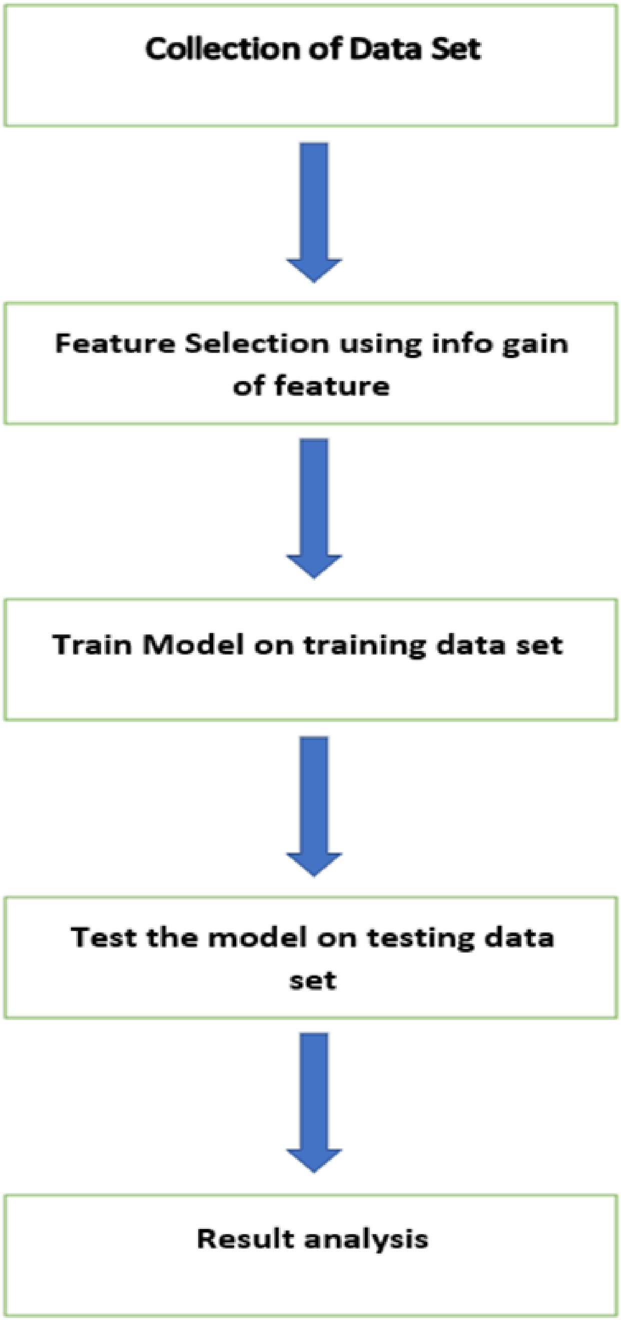

Collection of Data Set
Feature Selection using info gain of feature
Train Model on training data set
Test the model on testing data set
Result analysis

Figure 8

Activity diagram

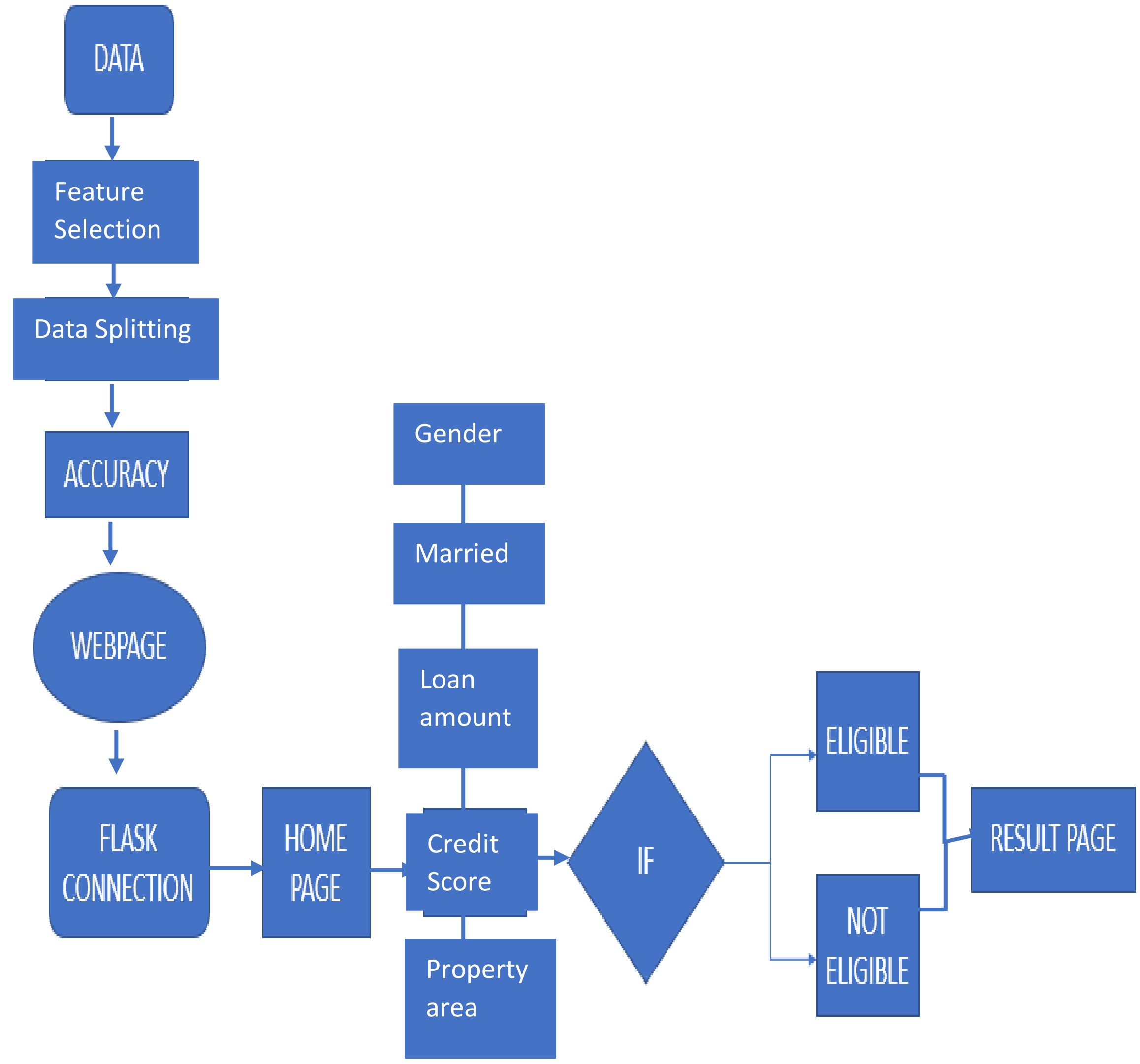
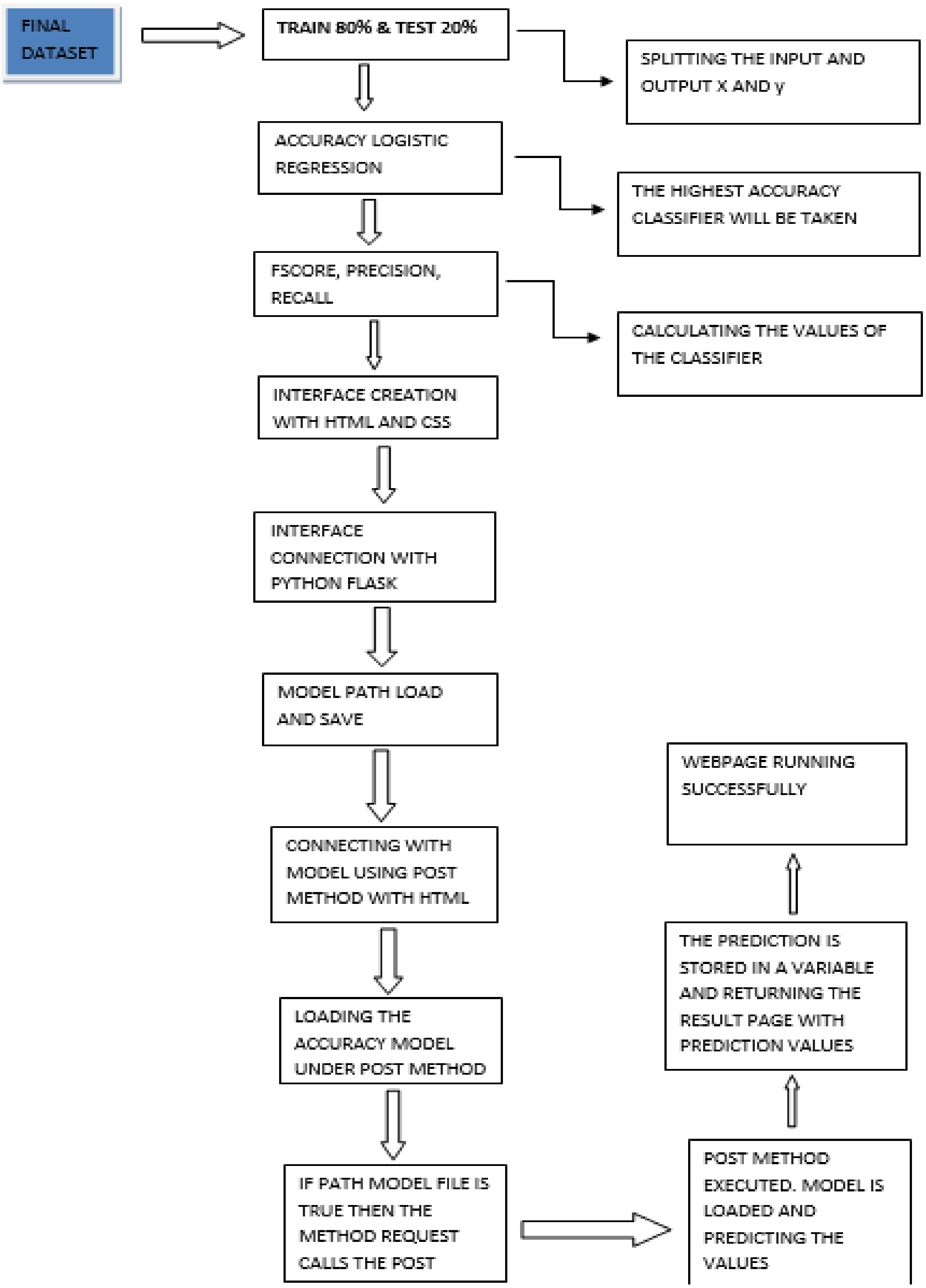


Figure 9

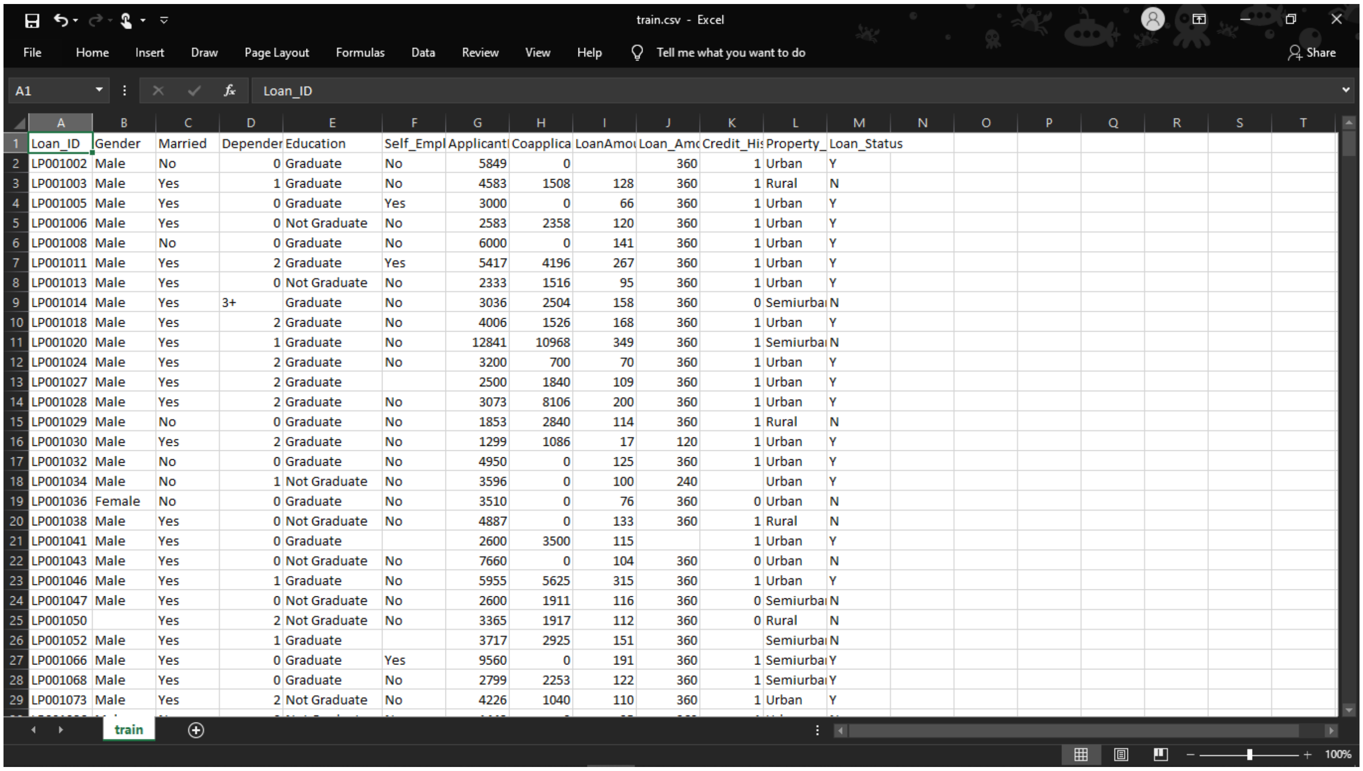
Methodology Figure 10


FINAL DATASET
TRAIN 80% & TEST 20%
SPLITTING THE INPUT AND OUTPUT X AND y
ACCURACY LOGISTIC REGRESSION
THE HIGHEST ACCURACY CLASSIFIER WILL BE TAKEN
FSCORE, PRECISION, RECALL
CALCULATING THE VALUES OF THE CLASSIFIER
INTERFACE CREATION WITH HTML AND CSS
INTERFACE CONNECTION WITH PYTHON FLASK
MODEL PATH LOAD AND SAVE
CONNECTING WITH MODEL USING POST METHOD WITH HTML
LOADING THE ACCURACY MODEL UNDER POST METHOD
IF PATH MODEL FILE IS TRUE THEN THE METHOD REQUEST CALLS THE POST
POST METHOD EXECUTED. MODEL IS LOADED AND PREDICTING THE VALUES
THE PREDICTION IS STORED IN A VARIABLE AND RETURNING THE RESULT PAGE WITH PREDICTION VALUES
WEBPAGE RUNNING SUCCESSFULLY

3.1. DATA COLLECTION AND DATA PREPROCESSING

DATASET COLLECTION:

The Loan dataset is collected from kaggle site. The required dataset is represented in the excel sheet with thirteen columns named loan_id, gender, married, dependent, education, self employed, applicant income, coapplicant income, loan amount, loan amount term, credit history, property area and loan status. The result stating true or false (0 or 1)is placed in the loan status column. The dataset collected for predicting loan default customers is predicted into Training set and testing set. Generally 80:20 ratio is applied to split the training set and testing set. The data model which was created using Decision tree is applied on the training set and based on the test result accuracy, Test set prediction is done. Following are the attributes.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Loan_ID	Gender	Married	Depender	Education	Self_Empl	Applicant	Coapplica	LoanAmou	Loan_Amo	Credit_His	Property_	Loan_Status
2	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurba	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurba	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
16	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
17	LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
18	LP001034	Male	No	1	Not Graduate	No	3596	0	100	240		Urban	Y
19	LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
20	LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	N
21	LP001041	Male	Yes	0	Graduate		2600	3500	115		1	Urban	Y
22	LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	N
23	LP001046	Male	Yes	1	Graduate	No	5955	5625	315	360	1	Urban	Y
24	LP001047	Male	Yes	0	Not Graduate	No	2600	1911	116	360	0	Semiurba	N
25	LP001050		Yes	2	Not Graduate	No	3365	1917	112	360	0	Rural	N
26	LP001052	Male	Yes	1	Graduate		3717	2925	151	360		Semiurba	N
27	LP001066	Male	Yes	0	Graduate	Yes	9560	0	191	360	1	Semiurba	Y
28	LP001068	Male	Yes	0	Graduate	No	2799	2253	122	360	1	Semiurba	Y
29	LP001073	Male	Yes	2	Not Graduate	No	4226	1040	110	360	1	Urban	Y

Figure 11

Pre processing

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency o the algorithm .The outliers have to be removed and also variable conversion need to be done. In order to overcoming these issues we use map function.

3.2 Correlating attributes

Based on the correlation among attributes it was observed more likely to pay back their loans. The attributes that are individual and significant can include Property area, education, loan amount, and lastly credit History, which is since by intuition it is considered as important. The correlation among attributes can be identified using corplot and boxplot in Python platform.

3.3 Classifiers which is used in this Problem:

Decision tree Classifier:

Definition:

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed .A decision node has two or more branches. Leaf node represents a classification or decision.

Usage:

The key idea is to use a decision tree to partition the data space into cluster (or dense) regions and empty (or sparse) regions. In Decision Tree Classification a new example is classified by submitting it to a series of tests that determine the class label of the example. Library that is imported in spyder: from sklearn.tree import DecisionTreeClassifier

Definition: In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

Random forest Classifier:

Definition: A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size.A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction.There needs to be some actual signal in our features so that models built using those features do better than random guessing.The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

Usage:

Random forest algorithm can be used for both classifications and regression task. It provides higher accuracy. Random forest classifier will handle the missing values and maintain the accuracy of a large proportion of data.

Library that is imported in spyder: from sklearn.ensemble import RandomForestClassifier

Logistic Regression:

Definition: The logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression[1] (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the logodds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a *logit*, from *logistic unit*, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio. **Usage:**

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. ... The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.

<u>Library that is imported in spyder:</u> from sklearn.linear_model import LogisticRegression

KNeighbors Classifier:

Definition: k-NN (Image credit) k-Nearest-Neighbors (k-NN) is a supervised machine learning model. Supervised learning is when a model learns from data that is already labeled. A supervised learning model takes in a set of input objects and output values.

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors.

Usage:

KNN makes predictions just-in-time by calculating the similarity between an input sample and each training instance. There are many distance measures to choose from to match the structure of your input data. That it is a good idea to rescale your data, such as using normalization, when using

KNN.

<u>Library that is imported in spyder:</u> from sklearn.neighbors import KNeighborsClassifier

Support Vector Machine:

Definition: SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. Simply put, it does some extremely complex data transformations, then figures out how to seperate your data based on the labels or outputs you've defined.

Usage:

SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

Library that is imported in spyder: sklearn.svm.LinearSVC

3.4 Building the classification model using Logistic Regression algorithm

For predicting the loan defaulter's and non defaulter's problem Logistic regression algorithm is used. It is effective because it provides better results in classification problem. It is extremely intuitive, easy to implement and provide interpretable predictions. It produces out of bag estimated error which was proven to be unbiased in many tests. It is relatively easy to tune with. It gives highest accuracy result for the problem.

Python Flask

Flask is a micro web framework written in Python. It can create a REST API that allows you to send data, and receive a prediction as a response. By Tim Elfrink, Data Scientist at Vantage AI. As a data scientist consultant, I want to make impact with my machine learning models. However, this is easier said than done.

To import flask method the library which is used: from flask

import Flask,render_template,request

WEBPAGE CREATION

We use html and css to develop interface. Some descriptions of html and css are below: Hypertext Markup Language (HTML) is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets (CSS) and scripting languages such as JavaScript. Web browsers receive HTML documents from a web server or from local storage and render the documents into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document.Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language like HTML. CSS is a cornerstone technology of the World Wide Web, alongside

HTML and JavaScript. CSS is designed to enable the separation of presentation and content, including layout, colors, and fonts. This separation can improve content accessibility, provide more flexibility and control in the specification of presentation characteristics, enable multiple web pages to share formatting by specifying the relevant CSS in a separate .css file, and reduce complexity and repetition in the structural content.

4.RESULT & ANALYSIS

4.1. Measurement Matrices Accuracy:

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition: Accuracy = Number of correct predictions Total number of predictions. In my problem, I used 10 classifiers to get the highest accuracy of all. And in my problem ,Logistic regression classifier has the highest accuracy.

Precision: precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, It is the number of correct positive by the number of positives.

$$Precision = \frac{TP}{TP+FP}$$

Recall: recall (also known as sensitivity) is the fraction of the total amount of relevant instances It is the number of correct positive by the number of relevant samples (all samples that should have been classified positive)

$$Recall = \frac{TP}{TP+FN}$$

F Score: The F score, also called the F1 score or F measure, is a measure of a test's accuracy. The F score is defined as the weighted harmonic mean of the test's precision and recall. This score is calculated according to:

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precisior}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\mathrm{precisior} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

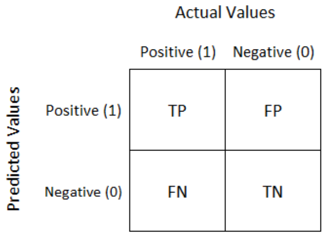
with the precision and recall of a test taken into account. Precision, also called the positive predictive value, is the proportion of positive results that truly are positive. Recall, also called sensitivity, is the ability of a test to correctly identify positive results to get the true positive rate. The F score reaches the best value, meaning perfect precision and recall, at a value of 1. The worst F score, which means lowest precision and lowest recall, would be a value of 0. The F score is used to measure a test's accuracy, and it balances the use of precision and recall to do it. The F score can provide a more realistic measure of a test's performance by using both precision and recall. The F score is often used in information retrieval for measuring search, document classification, and query classification performance.

$$F1 = 2 * \frac{precision\ * recall}{precision\ + recall}$$

Where TP represents True Positives, FP represents False Positives and FN represents False Negatives.

Confusion Matrix

One common way to evaluate the performance of a model with binary responses is to use a confusion matrix. The observed cases of default are defined as positives and non-default as negatives [10]. The possible outcomes are then true positives (TP) if defaulted customers have been predicted to be defaulted by the model. True negatives (TN) if non-default customers have been predicted to be non-default. False positives (FP) if non-default customers have been predicted to be defaulted, and false negatives (FN) if defaulted customers have been predicted to be nondefault. A confusion matrix can be presented as in the Figure below.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

From a confusion matrix there are certain metrics that can be taken into consideration. The most common metric is accuracy which is defined as the fraction of the total number of correct classifications and the total number of observations. It is mathematically defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

The issue with using accuracy as a metric is when applying it for imbalanced data. If the data set contains 99% of one class it is possible to get an accuracy of 99%, if all of the predictions are made for the majority class. A metric that is more relevant in the context of this project is specificity. It is defined as

$$\text{Specificity} = \frac{TN}{FP + TN}$$

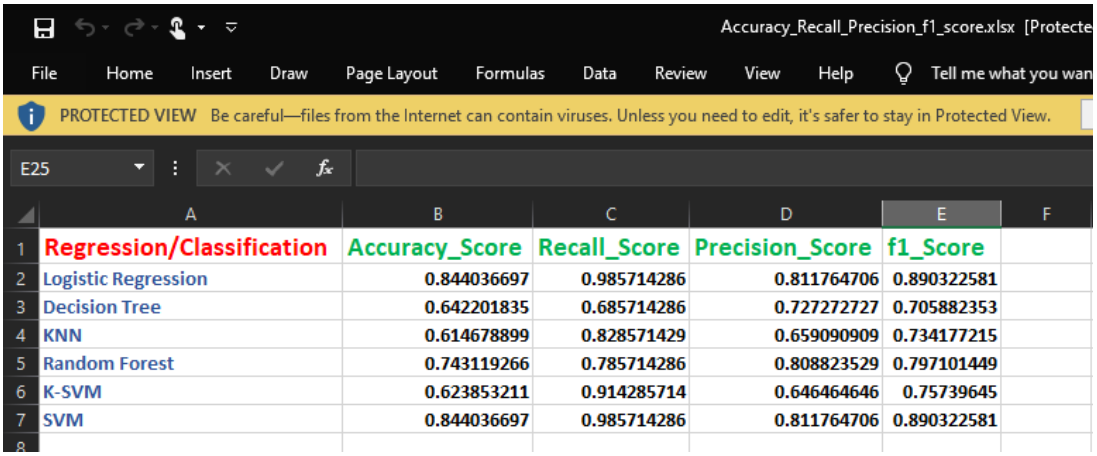
4.2 Problem Result:

Here is the result of accuracy,fscore,precision and recall of 7 classifiers.

The library which is imported to Spyder

from sklearn.metrics import accuracy_score

from sklearn.metrics import precision_recall_fscore_support The scores the represented in excel sheet is as below:



Accuracy_Recall_Precision_f1_score.xlsx [Protecte
File
Home
Insert
Draw
Page Layout
Formulas
Data
Review
View
Help
Tell me what you wan
PROTECTED VIEW Be careful—files from the Internet can contain viruses. Unless you need to edit, it's safer to stay in Protected View.
E25

	A	B	C	D	E	F
1	Regression/Classification	Accuracy_Score	Recall_Score	Precision_Score	f1_Score	
2	Logistic Regression	0.844036697	0.985714286	0.811764706	0.890322581	
3	Decision Tree	0.642201835	0.685714286	0.727272727	0.705882353	
4	KNN	0.614678899	0.828571429	0.659090909	0.734177215	
5	Random Forest	0.743119266	0.785714286	0.808823529	0.797101449	
6	K-SVM	0.623853211	0.914285714	0.646464646	0.75739645	
7	SVM	0.844036697	0.985714286	0.811764706	0.890322581	

Figure 12

Bar plotting representation of accuracy score, fscore, precision and recall:

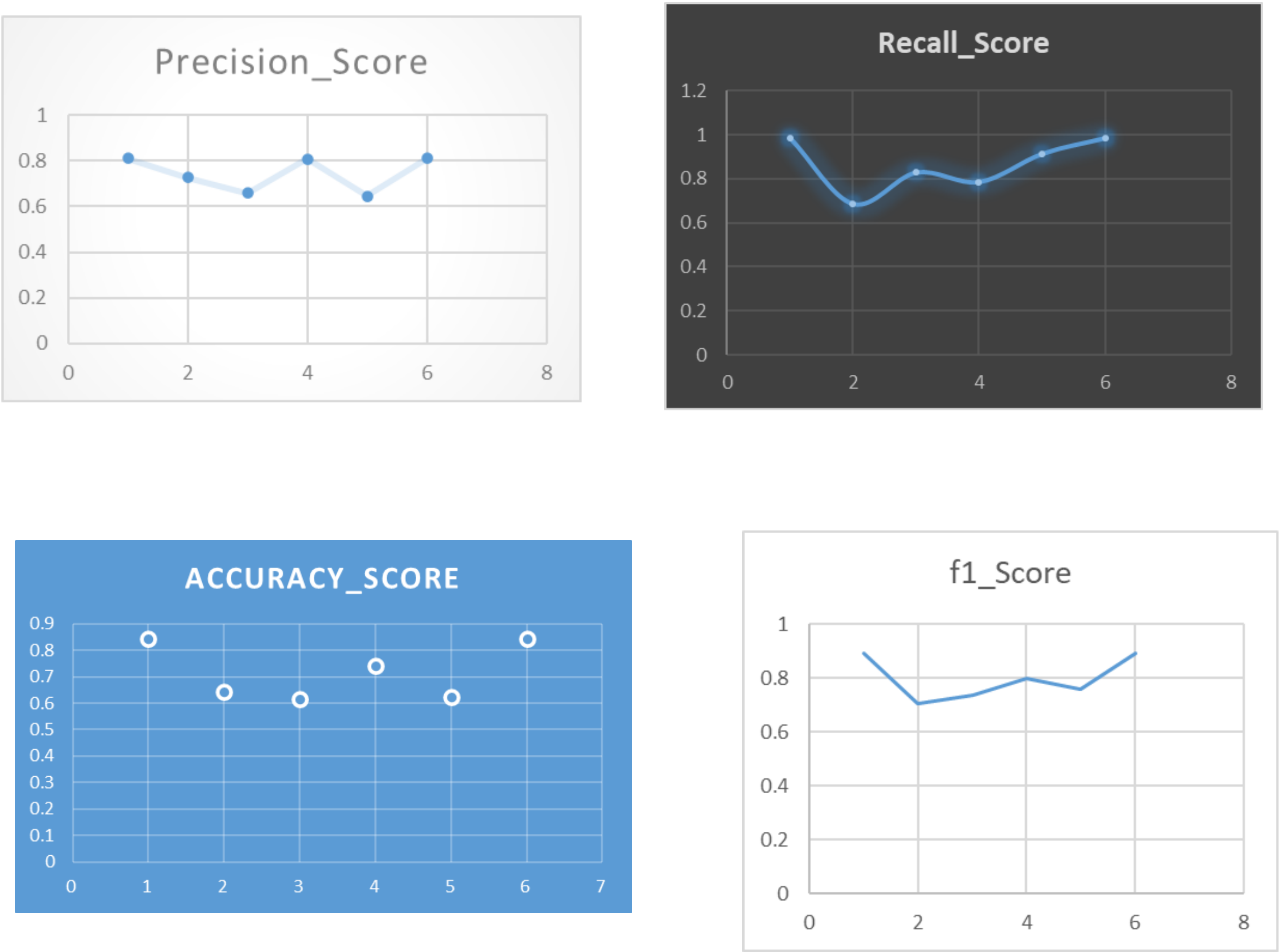
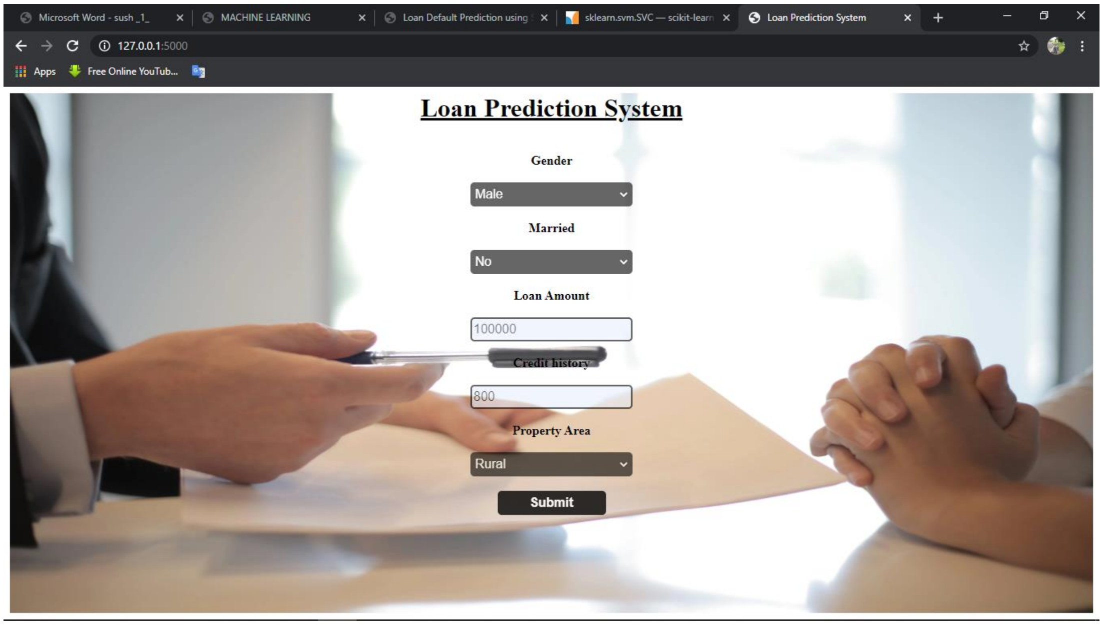


Figure 13

Loan Prediction System

Distribution of the loans is the core business part of almost every banks. The main portion the bank's assets is directly came from the profit earned from the loans distributed by the banks. The prime objective in banking environment is to invest their assets in safe hands where it is. Today many banks/financial companies approves loan after a regress process of verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. Through this system we can predict whether that particular applicant is safe or not and the whole process of validation of features is automated by machine learning technique. The disadvantage of this model is that it emphasize different weights to each factor but in real life sometime loan can be approved on the basis of single strong factor only, which is not possible through this system. Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this Paper is to provide quick, immediate and easy way to choose the deserving applicants. It can provide special advantages to the bank. The Loan Prediction System can can automatically calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight .A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not. Loan Prediction System allows jumping to specific application so that it can be check on priority basis. This Paper is exclusively for the managing authority of Bank/finance company, whole process of prediction is done privately no stakeholders would be able to alter the processing. Result against particular Loan Id can be send to various department of banks so that they can take appropriate action on application. This helps all others department to carried out other formalities. So basically our websites contains two Pages one is for taking the input from the user and second one is for predicting the result that whether the user is eligible for the loan or not. There are four input fields which the user has to fill in order to know the result. The four input fields are gender, marital status which is yes or no then loan amount , credit score and the property area. Here I am sharing the snapshots :


Microsoft Word - sush _1_
MACHINE LEARNING
Loan Default Prediction using
sklearn.svm.SVC — scikit-learn
Loan Prediction System
127.0.0.1:5000
Apps
Free Online YouTub...
Submit

Loan Prediction System

Gender

Male

Married

No

Loan Amount

100000

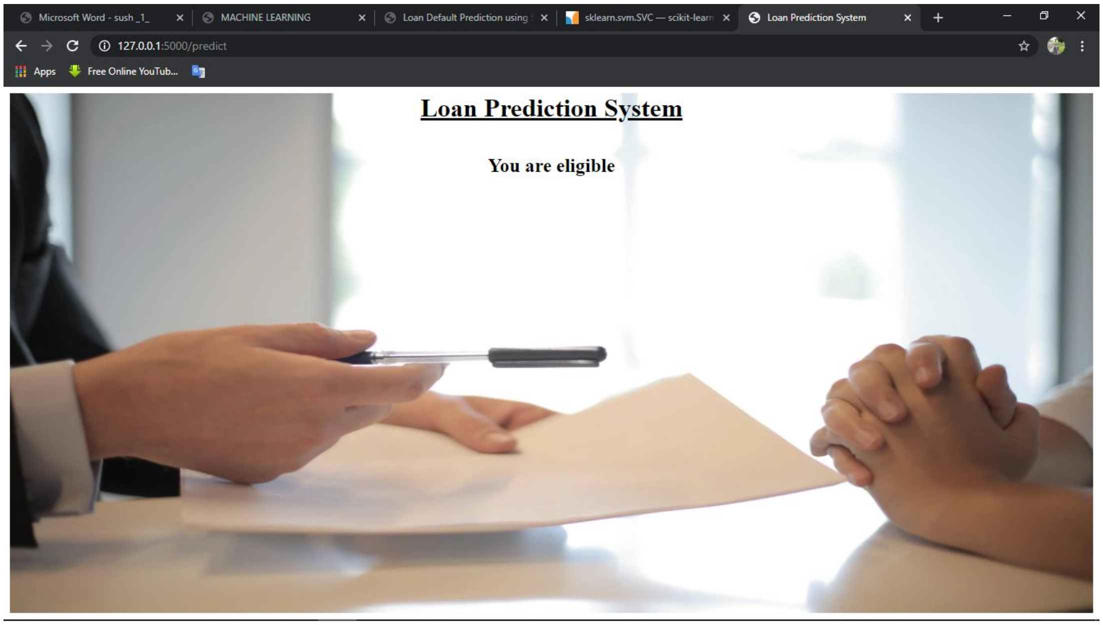
Credit history

800

Property Area

Rural

Figure 14: Snapshot of homepage



Loan Prediction System

You are eligible

Figure 15: Snapshot of result

Conclusion

The analytical process started from data cleaning and processing, Missing value imputation with mice package, then exploratory analysis and finally model building and evaluation. The best accuracy on public test set is 0.844. This brings some of the following insights about approval. Applicants with Credit history not passing fails to get approved, Probably because that they have a probability of a not paying back. Most of the Time, Applicants with high income sanctioning low amount is to more likely get approved which make sense, more likely to pay back their loans. Some basic characteristic gender and marital status seems not to be taken into consideration by the company.

Future work

The potential future work for this project will be a further development of the model by deepening analysis on variables used in the models as well as creating new variables in order to make better predictions. Data available for the scope of this thesis has constraints in terms of many years are covered by the data presented as well as geographical breadth of the Nordea's clients. The majority of customers at Nordea's clients are from the Nordic countries, thus, it should be considered that the behaviour of Nordic customers influence the results of this research. It means that the behaviour of clients outside Nordics may or may not follow the same pattern and therefore one should make additional analysis and obtain a geographically-broader data set if the objective is to have a model unbiased of the geographical location. An assumption can also be made that if there is data available for longer time span as well as broader geography of clients, there is an interest to implement marco economical variables, which in turn might open some new insights about factors impacting default of a customer as well as what machine learning methods are more suitable for this type of a problem. Further, a large part of this project was to make a grounded feature selection such that variables included in the models were valuable for prediction.

References

- https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset
- https://towardsdatascience.com/how-to-easily-deploy-machine-learning-models-using-flask-b95af8fe34d4
- https://www.geeksforgeeks.org/deploy-machine-learning-model-using-flask/