A PROJECT REPORT ON

LOAN DEFAULTER PREDICTION USING SUPERVISED MACHINE LEARNING ALGORITHMS

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN THE PARTIAL FULFILLMENT FOR THE AWARD OF THE DEGREE

OF

BACHELOR OF ENGINEERING
IN
COMPUTER ENGINEERING

BY

RAJASHREE THAKARE	**EXAMINATION SEAT NO. : B150134342**
MANASI BARGE	**EXAMINATION SEAT NO. : B150134213**
SANKET PADWAL	**EXAMINATION SEAT NO. : B150134284**
PRATIK DESALE	**EXAMINATION SEAT NO. : B150134229**

UNDER THE GUIDANCE OF
Prof. SMITA PATIL

DEPARTMENT OF COMPUTER ENGINEERING
K.K.WAGH INSTITUTE OF ENGINEERING EDUCATION AND RESEARCH, HIRABAI HARIDAS VIDYANAGARI, PANCHAVATI, NASHIK 422003

2021-22

CERTIFICATE

This is to certify that the project report entitled

LOAN DEFAULTER PREDICTION USING SUPERVISED MACHINE LEARNING ALGORITHMS

Submitted by

Rajashree Thakar	Examination Seat No. : B150134342
Manasi Barge	Examination Seat No. : B150134213
Sanket Padwal	Examination Seat No. : B150134284
Pratik Desale	Examination Seat No. : B150134229

is a bonafide work carried out by them under the supervision of Prof. Smita Patil and it is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University for the award of the Degree of Bachelor of Engineering (Computer Engineering)

This project report has not been earlier submitted to any other Institute or University for the award of any degree or diploma.

Prof. Smita Patil
Internal Guide
Dept. of Comp. Engg.

Prof. Dr. S. S. Sane
Head of Department
Dept. of Comp. Engg.

External Examiner

Date:
Place: Nashik

Abstract

It is essential for a bank to estimate the credit risk it carries and the magnitude of exposure it has in case of nonperforming customers. Estimation of this kind of risk has been done by statistical methods through decades and with respect to recent development in the field of machine learning, there has been an interest in investigatingif machine learning techniques can perform better quantification of the risk. The aim of this thesis is to examine which method from a chosen set of machine learning techniques exhibits. The best performance in default prediction with regards to chosen model evaluation parameters. The investigated techniques were KNN, Random Forest, XGBoost. An oversampling technique was implemented in order to treat the imbalance between classes for the response variable. The results showed that Random forest obtained the best result with respect to the chosen model evaluation metric.

Acknowledgment

First and foremost, we would like to thank our project guide, Prof. Smita Patil for her guidance and support. We will forever remain grateful for the constant support and guidance extended by my guide, in making this project report. Through our many discussions, she helped me to form and solidify ideas. With a deep sense of gratitude, we wish to express our sincere thanks to, Prof. Dr. S. S. Sane for his immense help in planning and executing the works in time. Our grateful thanks to the departmental staff members for their support. We would also like to thank our wonderful colleagues and friends for listening my ideas, asking questions and providing feedback and suggestions for improving our ideas.

Rajshree Thakare
Manasi Barge
Sanket Padwal
Pratik Desale
(B.E. Computer Engg.)

INDEX

1 Introduction **1**

1.1 Project Idea 2

1.2 Motivation of the Project 3

2 Literature Survey **4**

2.1 Literature Survey 5

3 Problem Definition and scope **8**

3.1 Problem Statement 9

3.1.1 Goals and objectives 9

3.1.2 Assumption and Scope 9

3.2 Methodology 10

3.3 Outcome 10

3.4 Type of Project 10

4 Project Plan **11**

4.1 Project Estimate 12

4.1.1 Reconciled Estimate 12

4.2 Project Timeline 13

4.3 Team Organization 15

4.3.1 Team structure 15

5 Software requirement specification **17**

5.1 Functional Requirements 18

5.2 Non Functional Requirements 18

5.2.1 Performance requirements....18
5.2.2 Safety requirements....18
5.2.3 Security requirements....19
5.2.4 Software quality attributes....19
5.3 Constraints....19
5.4 Hardware Requirements....20
5.5 Software Requirements....20

6 Detailed Design 21
6.1 Architectural Design(Block Diagram)....22
6.2 Data design....24
6.2.1 Data structure....24
6.2.2 Database description....24
6.3 Component design/ Data Model....24
6.3.1 Class Diagram....24
6.3.2 Component Diagram....26

7 Project Implementation 27
7.1 Overview of Project Modules....28
7.1.1 Module 1: Data Cleaning and Preprocessing....28
7.1.2 Module 2: Exploratory Data Analysis....28
7.2 Technology Used....29
7.2.1 KNN(K-Nearest Neighbor):....29
7.2.2 Random Forest:....30
7.2.3 XGBoost:....30

8 Results and Discussion 33
8.1 Experimental Setup....34
8.1.1 Data Set....34
8.1.2 Performance Parameters....34
8.1.3 Efficiency Issues....35
8.2 Software Testing....35
8.2.1 Test Cases and Test Results....36

8.2.2 GUI Testing 36
8.3 Results 36
8.3.1 Result Analysis and Discussion 36
8.3.2 Graphical Interface(if applicable) 37

9 Conclusion and Future Work **40**
9.1 Conclusion 41
9.2 Future Work 41

Annexure A Plagiarism Report **46**

List of Figures

4.1 Task Network. 14
4.2 Timeline Chart. 14

6.1 System Architecture. 22
6.2 DFD Stage 1. 23
6.3 DFD Stage 2. 23
6.4 DFD Stage 3. 23
6.5 Activity Diagram. 25
6.6 Component Diagram. 26

7.1 KNN(K-Nearest Neighbor). 29
7.2 Random Forest. 31
7.3 XGBoost. 32

8.1 Login Page 37
8.2 User 1 38
8.3 User 1 Result 38
8.4 User 2 39
8.5 User 2 Result 39

List of Tables

8.1 GUI Test Cases 36
8.2 Comparative Result 37

CHAPTER 1

INTRODUCTION

1.1 PROJECT IDEA

Nowadays clients can invest in client loans through peer-to-peer financing platforms such as Borrowing Club. Borrowing Club enables investors to browse clients' loan applications containing the applicant's credit history, loan details, employment status, and other self-reported personal information, to make determinations as to which loans to fund. Loan Borrowing has been an important part of daily lives for organizations and individuals alike. With the everincreasing competition in the financial world and due to a substantial amount of financial constraints, the activity of taking a loan has become more or less expected. Individuals around the world depend on the activity of loan Borrowing for reasons such as overcoming their financial constraints for them to achieve some personal goals. Similarly, banks and small to large firms depend on the activity of loan Borrowing for the basic purpose of managing their affairs and functioning smoothly in times where there are financial constraints. Without a doubt, financial Borrowing services hold a great amount of significance for any individual, business, or enterprise. As such services are required by an individual or a business to achieve or accomplish their goals and to compete with the giants of their fields. Financial loans are a major part of the primary source of capital not only in the emerging economies but also in the developed capital markets by both individuals and enterprises. The Borrowing growth by the financial firms and the banks is considered the key factor for the inflation level and interest rate of any country which drives its economic growth and depicts its economic condition. According to the mission statement of the study on the role of financial services, the economic growth of the real economy is the primary role of the financial firms. With such great importance and benefits of financial Borrowing comes some major issues and bottleneck problems. The most common and substantial issue in the domain of financial Borrowing is the fair and successful Borrowing of loans while keeping the ratio of loan defaulters to the least minimum value. In financial lending, the risk of loan defaulters can never be neutralized but can be minimize Loan

repayments should be monitored and whenever a customer defaults, ac- tion should be taken. Thus banks should avoid loans to risky clients , monitor loan repayments, and renegotiate loans when clients get into difficulties. Bad loans can be constrained by confirming that loans are made to only borrowers who are likely to be able to repay, and who are unlikely to become ruined. This paper proposed an onscreen solution for predicting loan defaulters using machine learning. we used a machine learning algorithm and expressive data sets attribute/factors including income, age, experience, profession, and married have been considered for the prediction and classification of performance respectively using two machine learning algorithms including Random forest, KNN and XGBoost are implemented to predict the loan defaulter. This dataset for the current study was based on clients ' behaviors.

1.2 MOTIVATION OF THE PROJECT

To build the predictive model for the individual assessment of loan application to determine whether the applicant will default or not. Create predicative model to classify each borrower as a defaulter or not using the data collected when the loan has been given. Determining the probability of user liability. Creating an interactive UI that will take users input and return an out-put. The results of these case educations give vision into techniques for precisely forecasting Loan Defaulters, and compare the accurateness of MI algorithms. have been considered for the prediction and classification of student performance respectively using two machine learning algorithms including Random forest, KNN, XGBoost are implemented to predict the loan defaulter. This dataset for the current study was based on customers behaviors. If the project gets successful then it will be a great help for faculty to boost the Finance organization.

CHAPTER 2

LITERATURE SURVEY

2.1 LITERATURE SURVEY

This section discusses in brief about some of the work that has already been done on creating ML models using various algorithms to improve the loan prediction process and help the banking authorities and financial firms select an eligible candidate with very low credit risk. Serrano-Cinca et al. used loan sample data from the Lending Club to account for default factors by adopting single factor mean test and survival analysis [7] .Advanced-support vector regression (SVR) techniques are applied to predict loss given default of corporate bonds by Yao et al., the results show that versions of SVR techniques perform better than other methods[8] . Malekipirbazari and Aksakalli proposed a Random Forest (RF) based classification method to identify high-quality P2P borrowing customers, by comparing with different machine learning methods, the results indicate the RF-based method is significantly preferred than the FICO credit scores as well as LC grades in identifying the good borrowers. Emekter et al.constructed a logistic regression (LR) model to predict the default probability of the borrower in the Lending Club, and the empirical evidence suggests that credit grade, debt-to-income ratio, FICO score and revolving line utilization play an important role in loan defaults[9] .Bagherpour used KNN, SVM, Random Forest and Sand Factorization Machines (FM) algorithms for predicting loan default on a large set of data[10] . Kvamme et al. based on Convolutional Neural Networks (CNN) to predict loan default by taking into account the time series data related to customer transactions in current accounts, savings accounts and credit cards. The research showed that CNN model outperforms Random Forest classifier[11] . Kim et al. proposed a method combining label propagation and transductive support vector machine (TSVM) with Dempster-Shafer theory to accurately predict the default of social lending with unlabelled data [12] . Research on the credit risk assessment: Tang et.al proposed a model of trust spiral and applied it to the study of credit risk issues within the lending relationship between banks and small businesses [13]. Moradi and Mokhatab Rafiei train an adaptive network-based fuzzy in-

ference system (ANFIS) using monthly data from a customer profile dataset and then using the newly defined factors and their underlying rules, a second round of assessment begins in a fuzzy inference system. Thus, they produce a table of bad customers on a monthly basis and creating a dynamic model based on the table for assessing the credit risk of the customers [14]. Brown, I. et al. obtained that Random Forest and Gradient Boosting classifiers perform outstanding in a credit scoring context and are able to cope preferably with pronounced class imbalances in these data sets by empirical study [15] . Li qualitatively analysed the possibility of loan defaults of borrowers who loan in the lending club, based on the borrowers' loan purpose, income level, residential address and work seniority, then via logistic regression model for predicting the default probability of borrowers so as to calculate the credit score of borrowers[16] . Zhang et al. adopted Multiple Instance Learning (MIL) to build a novel credit scoring model by using the socio-demographic and loan application data as well as the transaction history data of the applicant [17] . Djeundje B. V. and Crook used GAMs with cubic B-splines to estimate credit card survival models , the results show that GAMs outperform in improving the accuracy of predictions and so on[18] . Masmoudi K et al. adopted a discrete Bayesian network with a latent variable to model the loans subscribers who have default payment behaviour. The model is constructed to evaluate credit risk and cluster loans subscribers[19] . Papouskova and Hajek used heterogeneous ensemble learning to build a two-stage consumer credit risk model which adopts class-imbalanced ensemble learning and regression ensemble for predicting credit scoring and exposure at default respectively [17] . LightGBM, XGBoost, Logistic Regression and Random Forest are used by Ma et al. [15] and Coser et al. [16] to establish a series of prediction models for evaluating the probability of a customer's loan default. Cho et al. proposed an investment decision model in P2P lending market based on the instance-based entropy fuzzy support vector machine (IEFSVM) classification [2 1) Many researches have been conducted based on data mining and data analysing in the field of financial and banking sector. This section presents briefly some

of these techniques which are used in loans management and their finding Sudhakar et al focused on specifying the data mining applications usefulness, these applications are using several machine learning algorithms such as decision trees and Radial Basis Neural Networks. This study came with in which way to apply these applications in a loan approval assessment field. McLeod presents Neural networks properties and their fitness for the credit granting process. [1] 2) In another study the authors have proposed the pre-diction of the loan defaulters by including the relation ship of borrower mobile phone usage with the other loan default variables. Three different variables from the phone data were selected based on the high significance of the loan defaulter. The variables names are mobility patterns, telecommunication patterns, and App usage patterns. These variables are extracted carefully by using the recursive feature elimination (RFE) on the real data set. To keepthe privacy of the individuals the contact details such as the name, ID, and phone numbers were encrypted. With on the selected variables AdaBoost algorithm was applied as a decision tree classifier [2]. 3) Paper [3] reviewed many methods available like logistic regression, k- nearest neighbours, random forest, neural networks, support vector machines, stochastic gradient boosting, Naive Bayes, etc. and concluded that it is nearly impossible to declare one best method of all. 4) Pidikiti Supriya et al. [4] used Decision Trees as a machine learning tool to implement their model. They started their analysis with data cleaning pre-processing, missing value imputation, then exploratory data analysis, and finally model building and evaluation. The authors on a public test set managed to achieve the best accuracy of 815) The conducted tests using the C4.5 algorithm in decision trees in paper [5] showed that the maximum precision value achieved was 78.086) The authors in paper [6] did an exploratory data analysis. The paper's main purpose was to classify and examine the nature of loan applicants. Seven different graphs were plotted and visualized and using these graphs the authors concluded that most loan applicants preferred short-term loans.

CHAPTER 3

PROBLEM DEFINITION AND SCOPE

3.1 PROBLEM STATEMENT

To develop an application to Prediction of Loan Defaulter. The loan is one of the greatest significant products of finance. All the banks are trying to character out active occupational policies to encourage customers to apply for their loans. Though, there are some clients who perform undesirably later their applications are accepted. To avoid these circumstances, banks have to find some approaches to predict customers' performances. Machine learning algorithms have an attractive upright performance for this purpose and are extensively used by banking.

3.1.1 Goals and objectives

Goal and Objectives:

The objectives are as follows:

• Minimalize the risk of defaulters' nonpayment of the loans using the created model.

• Create a predictive model to classify each borrower as a defaulter or not using the data collected when the loan has been given.

• Determining the probability of user liability

• Creating an interactive UI that will take users input and return an output and return an output

3.1.2 Assumption and Scope

- The goal of this study is to build and evaluate the effects of several supervised binary classification techniques on defaulter prediction. The assessment of the model. This project's methodologies are limited to accuracy, sensitivity, F-score, and AUC value.

3.2 METHODOLOGY

The database is made up of 59400 rows and 64 columns. The columns indicate several pieces of information acquired as part of the Borrowing Club's initial investigation. According to the database structure file included with the dataset, columns contain details on the applicant and the status of their repaying. Since this loans were very certainly returned or failed on by now, data from 2014-2018 was considered. The first challenge was determining whether the columns included important data or were primarily vacant. Information gathering revealed numerous blank or almost null columns, which were eliminated from the database since it would be impossible to go back and respond for every dataset that did not appear to be essential at the moment of the loan process. Columns that linked to the user's profile and a descriptive of the request (provided by the client) were eliminated since they were usually filled with textual information..

3.3 OUTCOME

To conduct experiments on larger data sets or try to tune the model so as to achieve the state-of-art performance of the model.

3.4 TYPE OF PROJECT

- Research oriented
- And Using KNN, XGBOOST and Random forest

CHAPTER 4

PROJECT PLAN

4.1 PROJECT ESTIMATE

4.1.1 Reconciled Estimate

The model followed is the Constructive Cost Model (COCOMO) for estimating the efforts required in the completion of the porject. Like all estimation models, the COCOMO model requires sizing information. This information can be specified in the form of:

- Object Point
- Function Point(FP)
- Lines of Source Code(KLOC)

For our project, sizing information in the form of Lines of Source Code is used. The total lines of code,

Lines of Code (LOC) = 1250

The initial effort(Ei) in man-months is calculated using equations:

$$E = ax(KLOC)^b$$

where, a = 3.0, b = 1.12, for a semidetached project
E = Efforts in person-hours
E = 3.85 PM => 4 Months Approx.

$$TIME = c * (E)^d$$

Where, c = 2.5, d = 0.35 (for a semidetached project)
Time = Duration of Project in months => 4 Months

4.1.1.1 Time Estimates

$$C = D \times Cp \times hrs$$

Where,

C = Cost of project,

D = Duration in Hours,

Cp = Cost incurred per person-hour = 20 Rs ,

hrs = hours

Total of 4 person-months are required to complete the project successfully.

Total Duration of Project D (Time) = 6 Months

The approximate duration of the project is 4

There are 4 people required approximately to do the project

C = 4 *20 * 120 = Rs. 9,600/-

4.2 PROJECT TIMELINE

• Task 1: Project Topic Selection

• Task 2: Paper Research and Topic Finalization

• Task 3: Literature Survey

• Task 4: Studying Existing System

• Task 5: Navigating through various algorithm techniques

• Task 6: Process Flow and Block Diagram

• Task 7: Planning and Design

• Task 8: Dataset Collection

• Task 9: Implementation

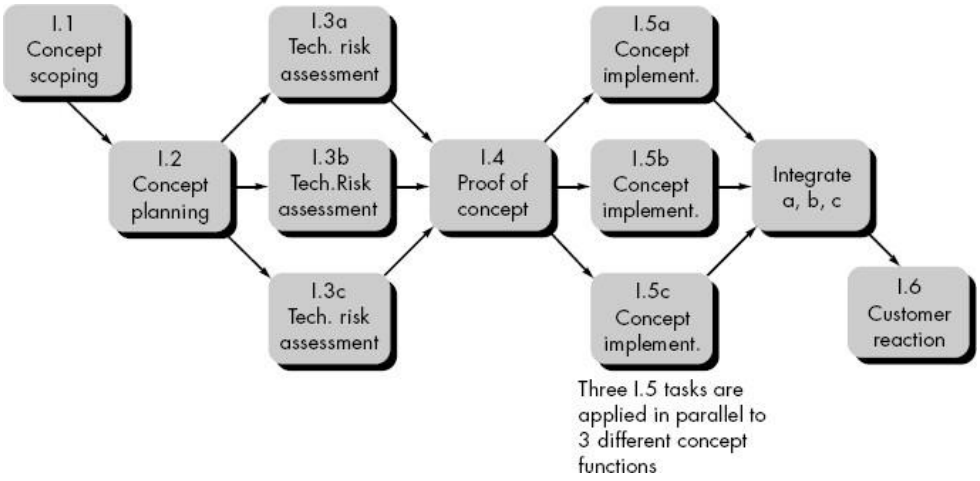


Figure 4.1: Task Network.

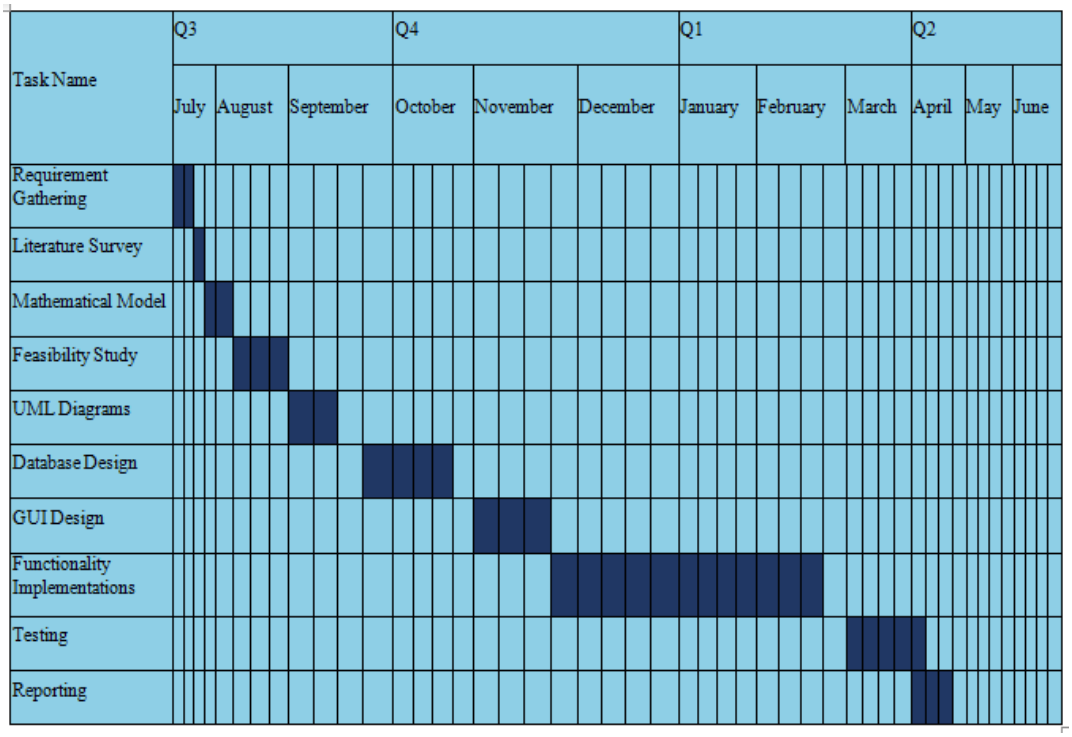


Figure 4.2: Timeline Chart.

• Task 10: Testing

• Task 11: Documentation

• Task 12: Report Generation

4.3 TEAM ORGANIZATION

• Prof. Smita Patil (Project Guide)

– Multiple Meetings conducted over the course of entire semester in the alloted span as well as whenever we required guidance in both online and in-person format.

– Communication over written medium as well as in person meets

• Prof. Priyadarshini I. (Project Co-ordinator)

– Guidance regarding scheduling of Project Review as well sorting out our technical difficulties

4.3.1 Team structure

The team structure for the project is identified. There are total 4 members in our team and roles are depend. All members are contributing in all the phases of project. Management reporting and communication: Well planning mechanisms are used for progress reporting and inter/intra team communication are identified as per requirements of the project.

Student Name	Responsibility of Student
Rajshree Thakare	Data Collection,Documentation
Manasi Berge	Literature Survey
Sanket Padwal	Coding, Documentation
Pratik Desale	Data Collection,Presentation

4.3.1.1 **Management Reporting And Communication**

• Feedback: Affords response, which confirms to the user that the association (which develops the software) recognizes the problems or difficulties to be answered and the software performance required to report those difficulties.

• Decompose the problem into components: Organizes the information and divides the problem into its component parts in an orderly manner.

• Validation: Uses authentication tactics practical to the supplies to admit that requirements are specified appropriately.

• Input to design: Contains the sauciest detail in the functional system requirements to devise a design solution.

• Basis for arrangement among the user and the organization: Offers a comprehensive explanation of the functions to be performed by the system. In addition, it helps the users to determine whether the spiced requirements are accomplished.

CHAPTER 5

SOFTWARE REQUIREMENT SPECIFICATION

5.1 FUNCTIONAL REQUIREMENTS

• The system shall be able to build Users profile.

• The system should be able maintain the user's record.

• The system will predict a users performance on the basis of the previous present record.

• The system should be able to display the users previous performance.

• The system should be able to predict the users next loan performance.

• The system should be able to tell about the users performance.

• On the basis of previous record the system should be able to notify about the users, that user is good or bad in that particular Loan Facility.

• Determining probability of user liability .

5.2 NON FUNCTIONAL REQUIREMENTS

5.2.1 Performance requirements

• The system gives advice or alerts user immediately.

• The System gives accurate results.

• Interactive, minimal delays, safe info transmission

5.2.2 Safety requirements

• Nobody will be harm while developing the system.

• Easy to use.

• System embedded with management procedures and validation procedures

5.2.3 Security requirements

• The system keeps all students' information's with high security.

• Identify all user, authenticate/verify user credentials, authorize user/third party, audit for user's usability, backup, server clustering, system policies

5.2.4 Software quality attributes

• Predictability

• Accuracy

• Maintainability

• Usability

• Modifiability

• Interoperability

• Efficiency

5.3 CONSTRAINTS

•Operational Constraints

– Dataset should be large enough to train ML model.

• Hardware Constraints

– The system meets the minimum requirement specifications

• Software Constraints

– All the modules required are updated to the minimum required version

• Assumptions

– we were able to identify significant factors that influence the performance of Users from the given dataset.

– After predicting the user performance, the system will also compare the results generated by two classification algorithms and there after determine which of them is more accurate and efficient.

5.4 HARDWARE REQUIREMENTS

- Processor – i3
- Hard Disk – 5 GB
- Memory – 1GB RAM

5.5 SOFTWARE REQUIREMENTS

- Operating System: Windows
- IDE: Jupyter Notebook
- Programming Languages: python 3
- Libraries and Frameworks: Numpy, OpenCV, Matplotlib, Spectral python(Spy).

CHAPTER 6

DETAILED DESIGN

6.1 ARCHITECTURAL DESIGN(BLOCK DIAGRAM)

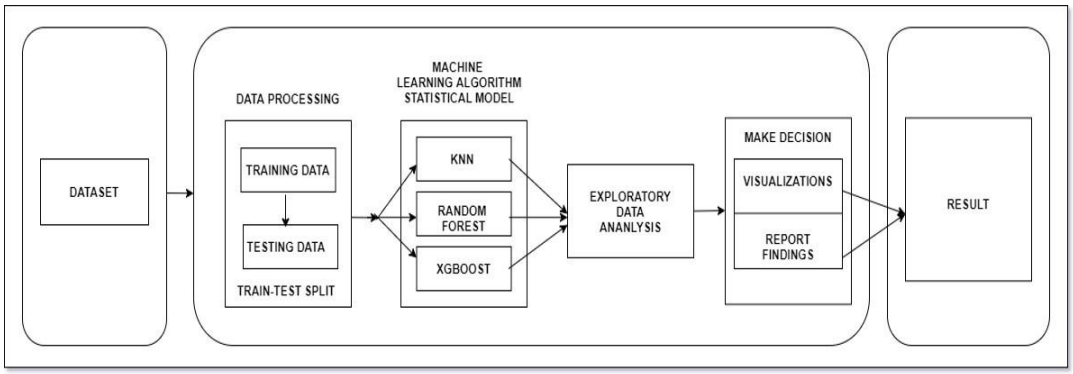


Figure 6.1: System Architecture.

Data objects characterized by labeled arrows and revolution are represented by circles also called foams. DFD has presented hierarchically i.e. the first data flow classical signifies the system as an entire. Subsequent DFD refines the finest diagram (level 0 DFD), providing increasing details with each subsequent level. The DFD permits the software engineer to progress models of the evidence field practical field at the same time. As the DFD is advanced into superior levels of detail, the specialist performs an implied functional breakdown of the system. At the same time, the DFD modification results in a conforming modification of the data as it moves through the process that symbolizes the applications. In context-level level DFD for the system, the prime peripheral objects produce information for use by the system besides consuming information produced by the system. The categorized arrow signifies data items or object pyramid. The framework diagram is the most intellectual data flow picture of a system. It signifies the complete system as an only bubble. The various external entities with which the system interacts and the data rows occurring between the system and the external entities are also represented. The name framework diagram is well accepted because it signifies the framework in which the system is to exist i.e. The peripheral objects (users) that would interrelate with the system and specific data items they would be getting from the system

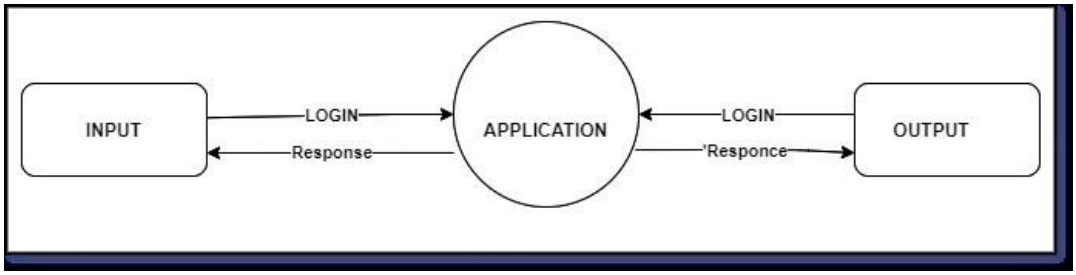


Figure 6.2: DFD Stage 1.

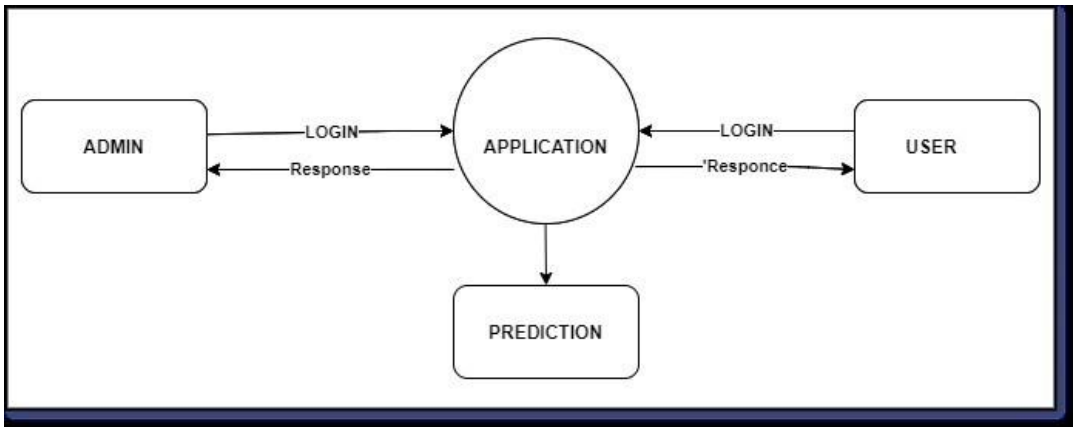


Figure 6.3: DFD Stage 2.

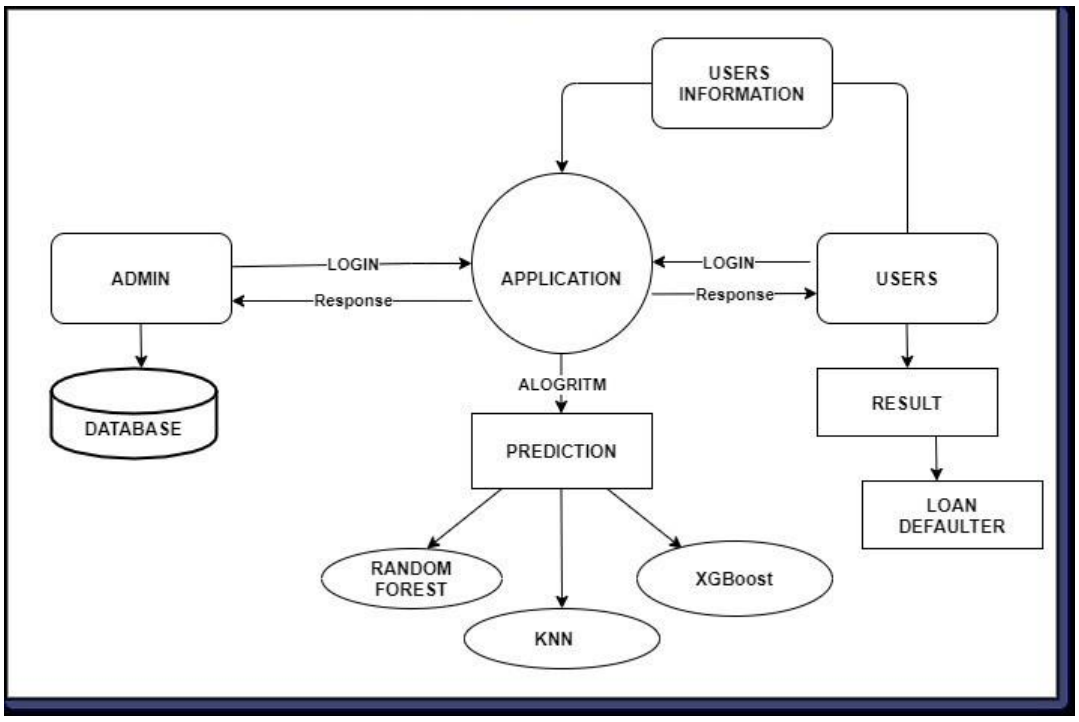


Figure 6.4: DFD Stage 3.

6.2 DATA DESIGN

A description of all data structures including internal, global, and temporary data structures, database design (tables), file formats.

6.2.1 Data structure

Array, Martrix

6.2.2 Database description

DataSet Attribute	Description	Type
Income	Income of the user	Int
Age	Age of the user	int
Experience	Professional experience of the user in years	Int
Profession	Profession	String
Married	Whether married or single	String
House ownership	Owned or rented or neither	string
Car ownership	Does the person own a car	String
Risk flag	Defaulted on a loan	String
Current job years	Years of experience in the current job	Int
City	City of residence	String
State	State of residence	String

6.3 COMPONENT DESIGN/ DATA MODEL

6.3.1 Class Diagram

The class diagram is a static diagram. It signifies the stationary opinion of an application. The class diagram is not only used for picturing, telling, and authenticating different characteristics of a system but also for creating executable code of the software application. A class diagram describes the features and actions of a class and also the limits forced on the system. The class diagrams are extensively used in the demonstration of object-oriented systems for they are the only UML diagrams, which can be planned straight with object-oriented languages.

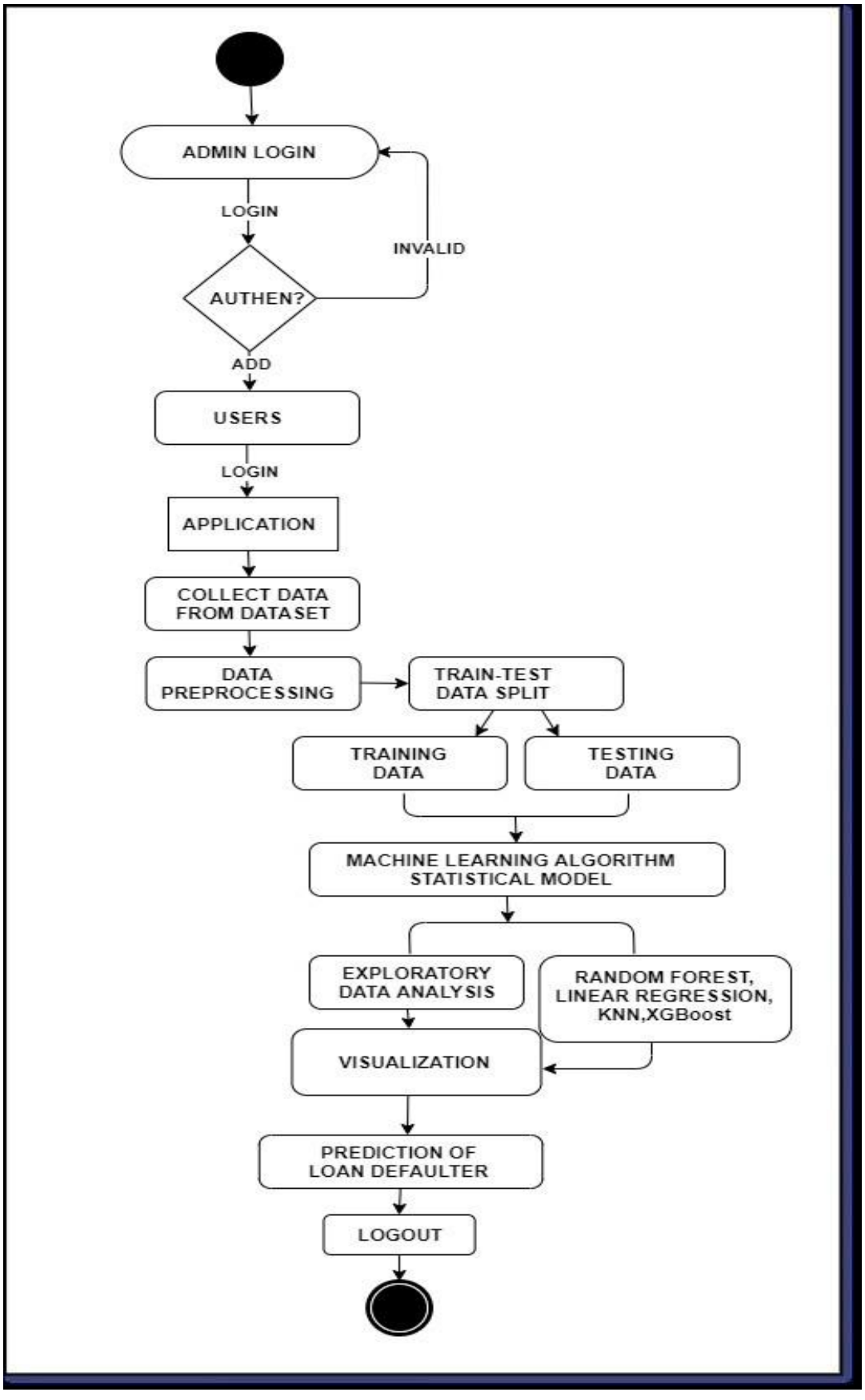


Figure 6.5: Activity Diagram.

6.3.2 Component Diagram

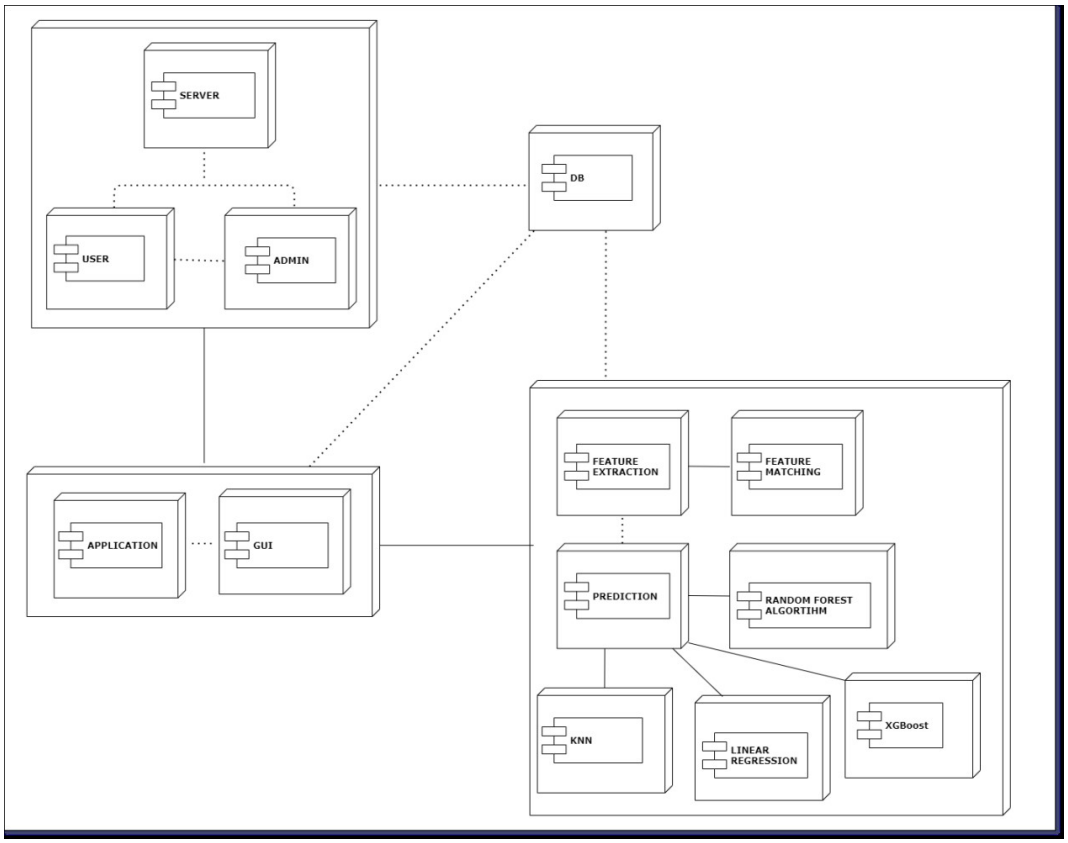


Figure 6.6: Component Diagram.

Component diagrams are dissimilar in terms of nature and performance. Componext diagrams are used to model the physical aspects of a system. Now the query is, what are these physical characteristics? Physical characteristics are the elements such as executables, reference libraries, files, papers, etc. which reside in a node. Component diagrams are used to picture the association and contacts among components in a system. These diagrams are also cast-off to make executable systems.

CHAPTER 7

PROJECT IMPLEMENTATION

7.1 OVERVIEW OF PROJECT MODULES

7.1.1 Module 1: Data Cleaning and Preprocessing

This paper proposed an automated solution for the predicting loan defaulter using machine learning. we used a machine learning algorithm and descriptive data sets attribute/factor include income, age , experience, profession n, married, house ownership, car ownership, risk flag, current job year, current house year, city, state. have been considered for the prediction and classification of student performance respectively using two machine learning algorithms including Random forest, KNN , XG-Boost are implemented to predict the loan defaulter. This dataset for the current study was based on customers' behaviours. It consists of 12 independent Attribute or Factors.

Columns, where there were more than was performed to clear up capacity and speed up operations And it might have been a model enhancement, the functionality would have been difficult to judge. Some type of trend analysis may be necessary, and specific measurements must offer an accurate estimation of an article's value and worth in the borrowing environment. The data processing procedure cleaning was executed in the following manner:

- Step 1: Determined the model's target
- Step 2: Removed features with just one distinct value
- Step 3: Removed features with less than 5
- Step 4: Removed elements that were unnecessary to the aim
- Step 5: Grouped characteristics with the same meaning.
- Step 6: Removed columns that had more than 30
- Step 7: Removed characteristics with the highest number of null values.
- Step 8: Removed any rows with null values.
- Step 9: Double-checked the characteristics

7.1.2 Module 2: Exploratory Data Analysis

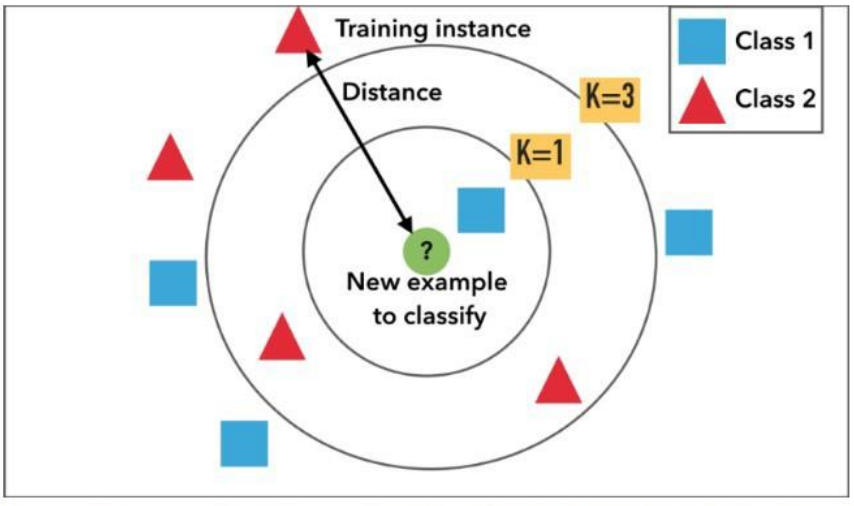
The initial stage was to tally the number of mortgage category kinds. The bulk of mortgages were classified as "active." Predictions were made for the "Completely

Funded" and "Billed Off" subcategories. The objective of the mortgages, as well as the loan amount, were then examined. The loan amount for credit counseling was found to be the highest, closely by credit or debit card. The correlation of house ownership and mortgage amount with user type was noted. It was discovered that collaborative product candidates leased, bought, or had their properties refinanced.. A majority of individuals who used a joint application had their properties borrowed money. With the user type, the objective allocation was noticed against the amount borrowed. There were several observations. Local company as the aim of the borrowing was often seen for collaborative types of applications above personal. Collaborative product types were infrequently used for purposes such as "travel," "break," "home," "learning," or "sustainable sources." different data type.

7.2 TECHNOLOGY USED

7.2.1 KNN(K-Nearest Neighbor):

KNN Algorithm is based on feature similarity: How closely out-of-sample features resemble our training set determines how we classify a given data point:



Example of k-NN classification. The test sample (inside circle) should be classified either to the first class of blue squares or to the second class of red triangles. If k = 3 (outside circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If, for example k = 5 it is assigned to the first class (3 squares vs. 2 triangles outside the outer circle).

Figure 7.1: KNN(K-Nearest Neighbor).

KNN can be used for classification — the output is a class membership (predicts a class — a discrete value). An object is classified by a majority vote of its

neighbors, with the object being assigned to the class most common among its k nearest neighbors. It can also be used for regression — output is the value for the object (predicts continuous values). This value is the average (or median) of the values of its k nearest neighbors.

7.2.2 Random Forest:

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.We use Random Forest to predict whether the customer is going to cancel his subscription. Random Forest uses Decision trees for classifying whether the customer is going to cancel his subscription. The random forest consists of a large number of decision trees. A decision tree points to a specific class. A class with more number of votes will be the classifier for a particular customer. Decision trees are sensitive to the data they are trained in. To avoid this, we use Bagging. Bagging is a kind of process where we take a random sample from the dataset for training decision trees.

7.2.3 XGBoost:

XGBoost has become a widely used and really popular tool among Kaggle competitors and Data Scientists in industry, as it has been battle tested for production on large-scale problems. It is a highly flexible and versatile tool that can work through most regression, classification and ranking problems as well as user-built objective functions. As an open-source software, it is easily accessible and it may be used through different platforms and interface.XGBoost is the abbreviation for eXtreme Gradient Boosting. The primary purpose of using XGBoost is due to its execution speed, and it’s model performance. XGBoost uses ensemble learning methods; i.e., it uses a combination of different algorithms and produces output as a single model. XGBoost supports parallel and distributed computing while offering efficient memory usage.

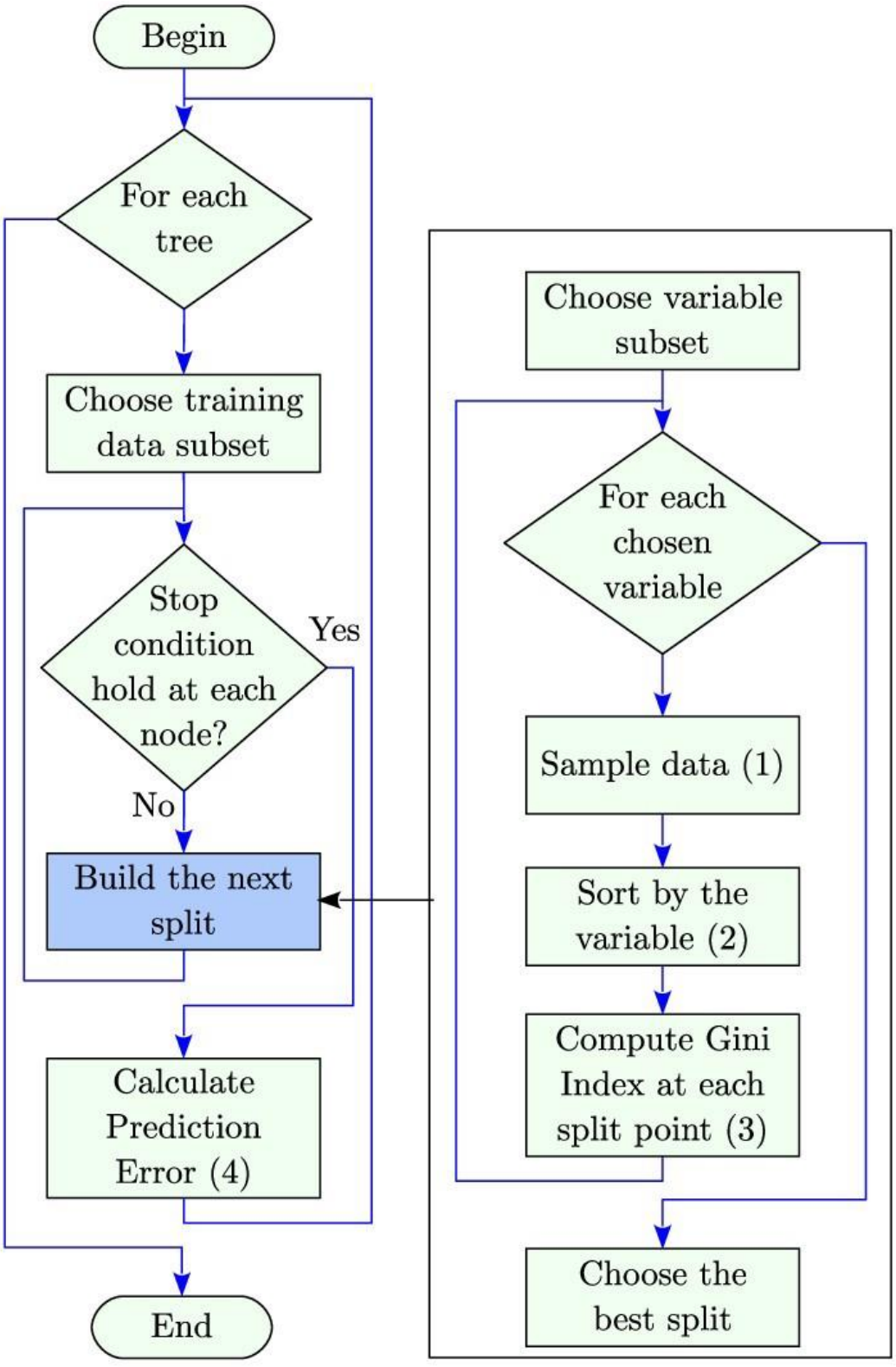


Figure 7.2: Random Forest.

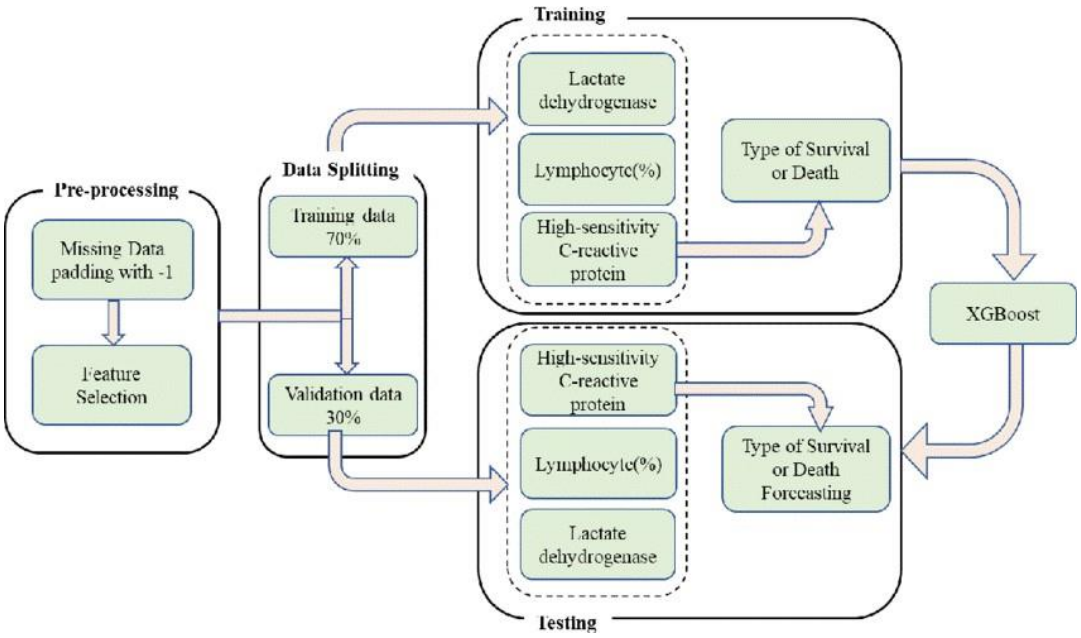


Figure 7.3: XGBoost.

CHAPTER 8

RESULTS AND DISCUSSION

8.1 EXPERIMENTAL SETUP

8.1.1 Data Set

This paper proposed an automated solution for the predicting loan defaulter using machine learning. we used a machine learning algorithm and descriptivedatasetsattribute/factorincludeincome,age,experience,profession,married,house ownership,car ownership,risk flag,current job year,current house year,city,state.have been considered for the prediction and classification of student performance respectively using two machine learning algorithms including Random forest, KNN , XGBoost are implemented to predict the loan defaulter. This dataset for the current study was based on customers behaviours.It consists of 12 independent Attribute or Factors.

8.1.2 Performance Parameters

The major emphasis of this research is on the four performance evaluation criteria for classification comparison of accuracy, AUC, Recall and ROC.

- **Accuracy:** The determined by the number of examples properly identified by the classification to the entire sample for a particular test data set is defined as accuracy.
- **F1-score:** The F1-score, also known as a balancing F Score, is the balanced average of Precision and Recall.
- **Recall:** Recall is the percentage of positive (default) cases identified by the classification. It is from time to time mentioned to as the True positive rate.
- **ROC(receiver operating characteristic curve):**An operating characteristic (roc characteristic (ROC) in statistical is a two-dimensional pictorial plot that depicts the efficiency of a binary classification system. The curve is constructed by plotting a graph (TPR) vs the false positive rate (FPR) at different criteria. The ROC curve may intuitively describe classification result.

8.1.3 Efficiency Issues

Through using Recursive Feature Elimination approach, 42 characteristics with the greatest association with the attribute value were selected and removed one by one to accomplish the initial dimension reduction. The goal feature 'risk flag' contains a high number of normal and default classifications, which will make model learning difficult.

8.2 SOFTWARE TESTING

Following testing types are considered for testing of the project:

1. **Module Testing**: This sort of testing focuses on specific modules, their functionalities, input and output. Because the project was separated into modules, each module was tested for appropriate operation during development. The module was tested by providing an input in the appropriate format and then compared the actual and expected output to determine the module's accuracy.

2. **Integration Testing**: It is concerned with testing the interfaces of multiple modules, as well as their incorporation and interconnection. Every module was tested to ensure. They were tested following their incorporation in this form of assessment. Each module uses the information from the previous module as input for the current domain, making it critical for testing the interaction. The components' interaction is also tested.

3. **Black Box Testing**: Testing only the functionality of a system without considering the code and complexity of implementation comes in Black Box Testing.

4. **White Box testing**: Implemented code is checked for logical errors, countless loops, complication, error handling and numerous conditions in White Box Testing.

NO .	Test	Description	Result
1	Menu Navigation	User clicks on a section name on the menu bar	The respective section opens
2	Fill the given form	User has to filled all personal details	On Bases of information result will be predicated
3	Result display Position	Where the result from the predictor is displayed on the page	Result is displayed at the center of the page
4	Return to Home Button	User clicks on the **Return Home** Button	Return to Home Page

Table 8.1: GUI Test Cases

8.2.1 Test Cases and Test Results

8.2.2 GUI Testing

8.3 RESULTS

8.3.1 Result Analysis and Discussion

Result of the system:

Input user information will be used to estimate whether the user would default on a loan or not, and the system will provide suitable output based on the algorithm

Discussion:

The findings were acquired using the kaggle data set. Four separate data sets were investigated, one of which was created by doing variable selection with correlation analysis using Kendall's Tau and had 15 variables in total. RFE created the remaining three data sets, which had , 6, 18, and 26 features, respectively. Imbalance was used to oversample the minority class in half of the models. The minority class was reproduced in these circumstances until its size matched to 24.

Comparative Result

Sr no.	Techniques Used	Accuracy
1	XG-Boost	63%
2	KNN	90%
3	Random Forest	94%

Table 8.2: Comparative Result

8.3.2 Graphical Interface(if applicable)

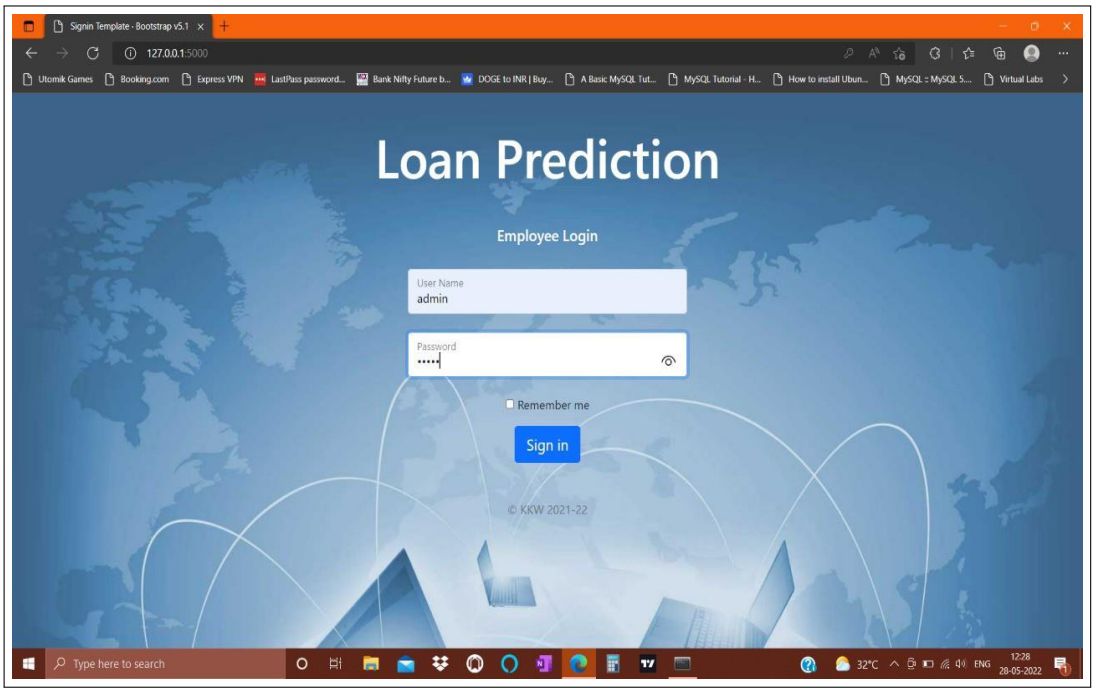


Figure 8.1: Login Page

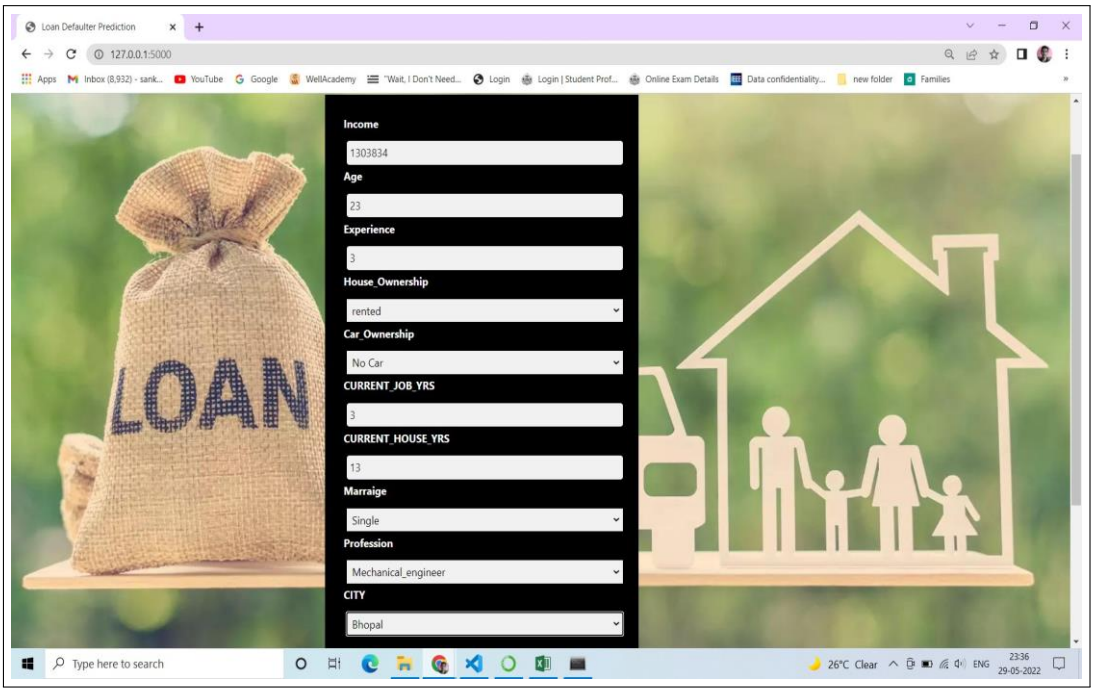


Figure 8.2: User 1

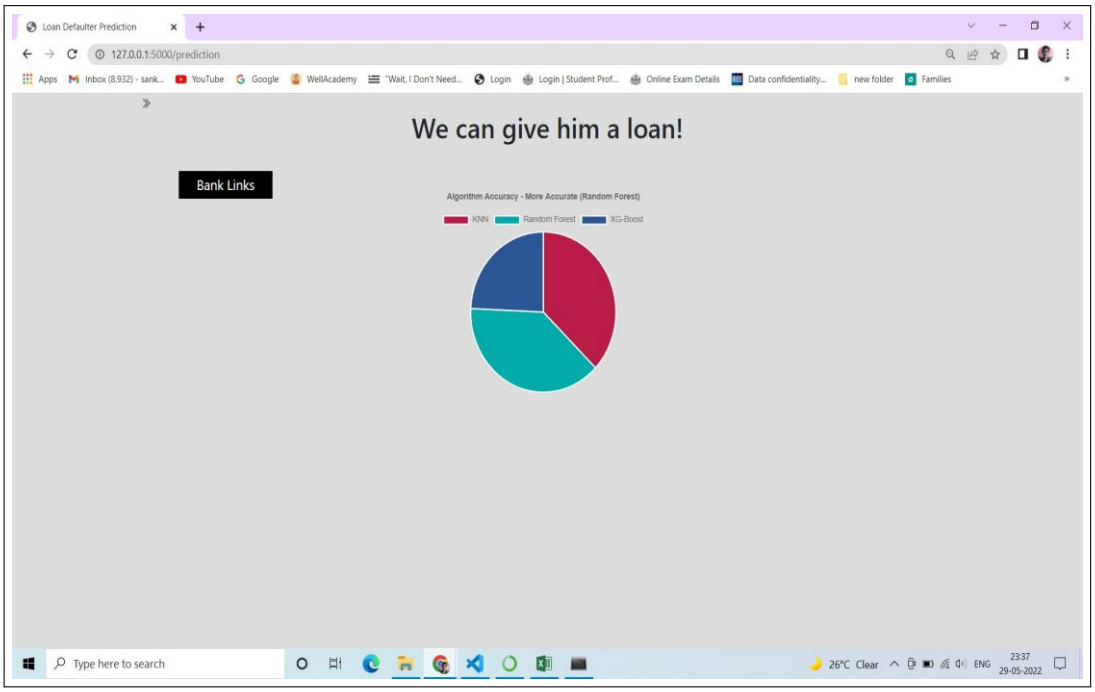


Figure 8.3: User 1 Result

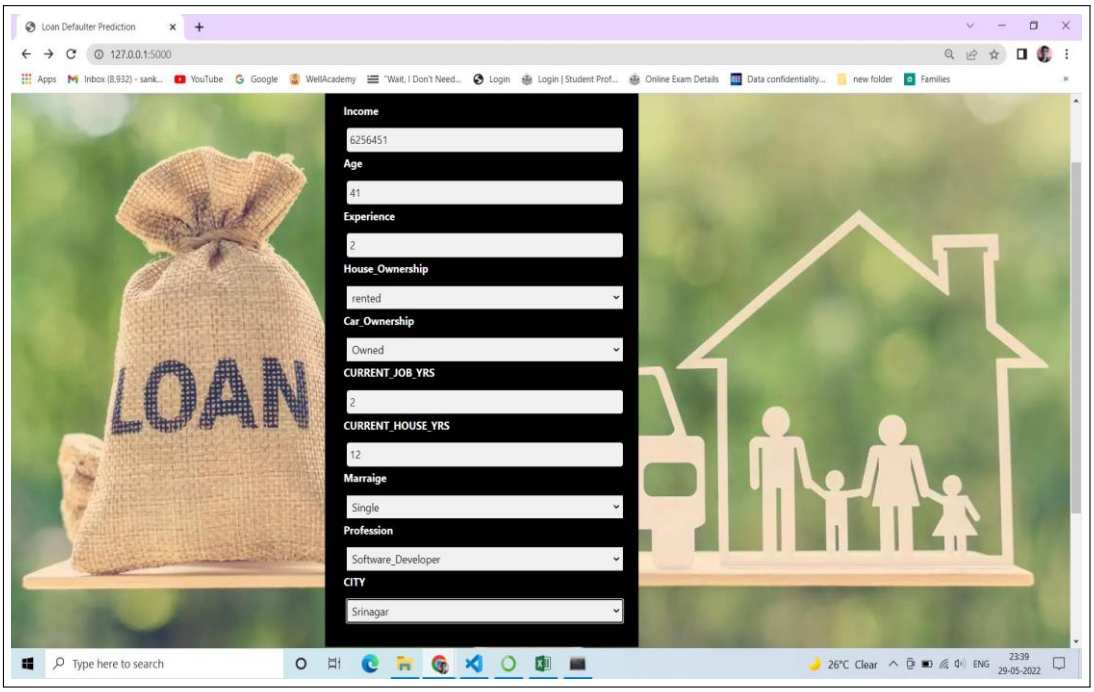


Figure 8.4: User 2

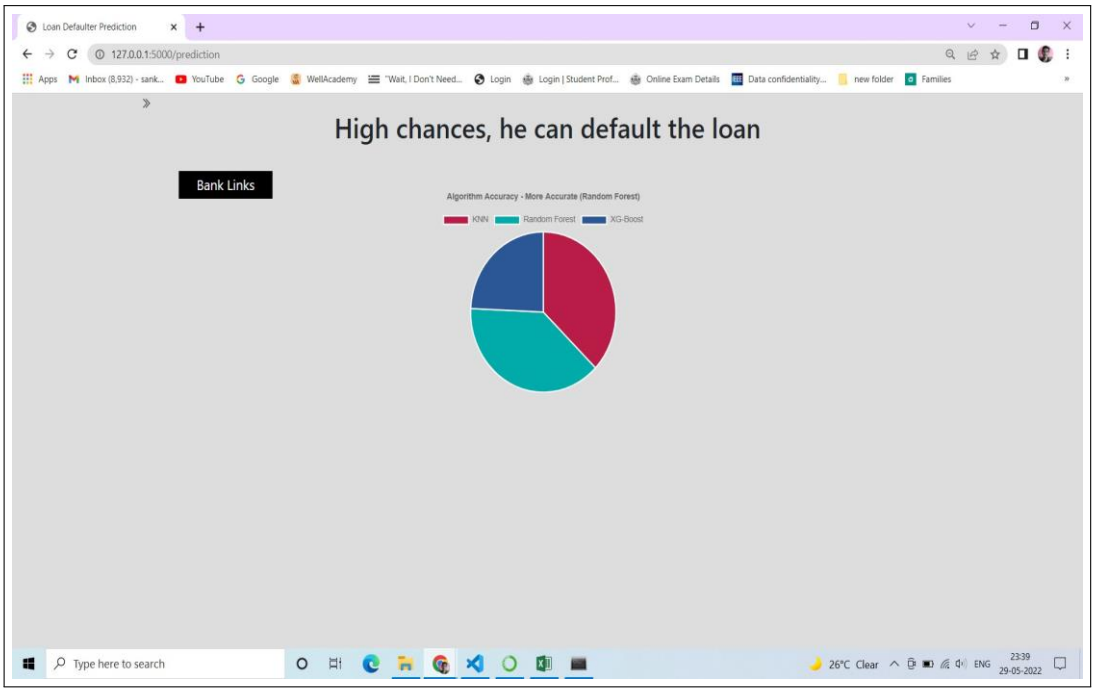


Figure 8.5: User 2 Result

CHAPTER 9

CONCLUSION AND FUTURE WORK

9.1 CONCLUSION

The random forest methodology is used in this research to develop a model for forecasting nonperforming loans in the mortgage lenders, and the performance is compared to other techniques such as logistic regression and XGBOOST. The test confirms that the random forest method outperforms the other two algorithm on the basis of loan defaulter prediction and has a great capacity to generalise. In future research, we will aim to run trials on bigger data sets or adjust the model to obtain state-of-the-art efficiency. The objective of this study was to investigate, evaluate, and develop a classification algorithm that would properly states that a person, given certain characteristics, has a high likelihood of defaulting on a loan. Like the problem of a high proportion of nonperforming debt is crucial in the finance sector, particularly in microfinance institutions in many developing and developed nations. Despite the fact that loan borrowing has been shown to be quite important in the security of any nation's economy in this century, such a large number of nonperforming loans is also quite crucial.

9.2 FUTURE WORK

The prospective work for this research will be a deeper examination of factors utilized in the models as well as the creation of additional variables in order to create good forecasts. The data accessible for the scope of the study has limitations in relation to the number of years represented by the data given as well as the regional scope of the Kaggle datasets. Because the bulk of Northern clients is from the Northern nations, it is reasonable to assume that the behaviors of clients influence the findings of this study. It indicates that clients' behavior beyond the Northern part would not contain traces, and hence one. Whereas if goal is to build a model that is independent of regional area, extra research and a substantially bigger data set should be performed. An expectation can also be created that even if there is evidence exist for a longer time horizon and also a widening geographic region of clients, there will be a desire to enforce macroeconomic factors, which may lead to new perspectives about factors influencing client default as well as which machine learning techniques

are best suited for this category. Furthermore, a big portion of this effort was to construct a grounded classification algorithm such that characteristics in the modeling techniques were efficient and accurate.

REFERENCES

[1]Ogawa, Ms Sumiko, et al. Financial Interconnectedness and Financial Sector Reforms in the Caribbean. No. 13-175. International Monetary Fund, 2013. 10. Ma, L., Zhao, X., Zhou, Z. and Liu, Y., 2018. A new aspect on P2P.

[2]Ma, L., Zhao, X., Zhou, Z. and Liu, Y., 2018. A new aspect on P2P online lending default prediction using meta-level phone usage data in China. Decision Support Systems, 111, pp.60–71

[3] Breeden J L 2020 A survey of machine learning in credit risk .

[4]Supriya P, Pavani M, Saisushma N, Kumari N V and Vikas K 2019 Loan prediction by using machine learning models Int. Journal of Engineering and Techniques 5 pp144–8

[5]Amin R K, Indwiarti and Sibaroni Y 2015 Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor (case study: bank pasar of yogyakarta special region) The 3rd Int. Conf. on Information and Communication Technol. (ICoICT) pp 75–80

[6]Jency X F, Sumathi V P and Sri J S 2018 An exploratory data analysis for loan prediction based on nature of the clients Int. Journal of Recent Technol. and Engineering (IJRTE) 7 pp 176–9

[7]Shoumo S Z H, Dhruba M I M, Hossain S, Ghani N H, Arif H and Islam S 2019 Application of machine learning in credit risk assessment: a prelude to smart banking TENCON 2019 – 2019 IEEE Region 10 Conf. (TENCON) pp 2023–8

[8]M.Ramaswami and R.Bhaskaran, “A CHAID Based Performance Prediction Model in Educational Data Mining”, International Journal of Computer Science Issues Vol. 7, Issue 1, No. 1, January 2010.

[9]Malekipirbazari M , Aksakalli V . Risk assessment in social lending via random forests[J]. Expert Systems with Applications, 2015, 42(10):4621-4631.

[10] Selvamuthu, D.Kumar, V.Mishra, A. Financ Innov(2019)5:16 https://doi.org/10.1186/s40854-019-0131-7.

[11] Ye, Xin, Dong, Lu-an, Ma, Da. Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score[J]. Electronic Commerce Research and Applications

[12] Serrano-Cinca, C., Gutiérrez-Nieto, B., López-Palacios, L., 2015. Determinants of default in P2P lending. PLoS One

[13] Yao X , Crook J , Andreeva G . Support vector regression for loss given default modelling[J]. European Journal of Operational Research, 2015

[14] Emekter, R., Tu, Y., Jirasakuldech, B., Lu, M., 2015. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Appl

[15] Bagherpour, A. (2017),Predicting Mortgage Loan Default with Machine Learning Methods; University of California, Riverside

[16] Masmoudi K , Abid L , Masmoudi A . Credit risk modeling using Bayesian network with a latent variable[J]. Expert Systems with Applications, 2019

[17] Viani Biatat Djeundje, Jonathan Crook, Identifying hidden patterns in credit risk survival data using Generalised Additive Models., European Journal of Operational Research (2019)

[18] Zhang, T. et al. (2018), Multiple Instance Learning for Credit Risk Assessment with Transaction Data;Knowledge-Based Systems

[19] Zhang, T. et al. (2018), Multiple Instance Learning for Credit Risk Assessment with Transaction Data;Knowledge-Based Systems

ANNEXURE A

PLAGIARISM REPORT