A MINI PROJECT REPORT

ON

"LOAN PREDICTION SYSTEM"

Submitted in the partial fulfillment of the requirements for

The degree of

BACHELOR OF ENGINEERING IN COMPUTER ENGINEERING

By

1. ZEESHAN ANSARI
2. KANCHAN MENGUNE
3. BINITDEV PANDEY
4. ADNAN SHAIKH

UNDER THE GUIDANCE OF

Prof. Shatabdi Bhalerao

Department of Computer Engineering
Saraswati College of Engineering, Kharghar, Navi Mumbai
University of Mumbai

2021-22

Saraswati College of Engineering, Kharghar

Vision:

To be universally accepted as autonomous center of learning in Engineering Education and Research.

Mission:

- To educate students to become responsible and quality technocrats to fulfil society and industry needs.
- To nurture student's creativity and skills for taking up challenges in all facets of life.

Department of Computer Engineering

Vision:

To be among renowned institution in Computer Engineering Education and Research by developing globally competent graduates.

Mission:

- To produce quality Engineering graduates by imparting quality training, hands on experience and value education.
- To pursue research and new technologies in Computer Engineering and across interdisciplinary areas that extends the scope of Computer Engineering and benefit humanity.
- To provide stimulating learning ambience to enhance innovative ideas, problem solving ability, leadership qualities, team-spirit and ethical responsibilities.

DEPARTMENT OF COMPUTER ENGINEERING

PROGRAM EDUCATIONAL OBJECTIVE'S

1. To embed a strong foundation of Computer Engineering fundamentals to identify, solve, analyze and design real time engineering problems as a professional or entrepreneur for the benefit of society.
2. To motivate and prepare students for lifelong learning & research to manifest global competitiveness.
3. To equip students with communication, teamwork and leadership skills to accept challenges in all the facets of life ethically.

DEPARTMENT OF COMPUTER ENGINEERING

PROGRAM OUTCOMES

1. Apply the knowledge of Mathematics, Science and Engineering Fundamentals to solve complex Computer Engineering Problems.
2. Identify, formulate and analyze Computer Engineering Problems and derive conclusion using First Principle of Mathematics, Engineering Science and Computer Science.
3. Investigate Complex Computer Engineering problems to find appropriate solution leading to valid conclusion.
4. Design a software System, components, Process to meet specified needs with appropriate attention to health and Safety Standards, Environmental and Societal Considerations.
5. Create, select and apply appropriate techniques, resources and advance engineering software to analyze tools and design for Computer Engineering Problems.
6. Understand the Impact of Computer Engineering solution on society and environment for Sustainable development.
7. Understand Societal, health, Safety, cultural, legal issues and Responsibilities relevant to Engineering Profession.
8. Apply Professional ethics, accountability and equity in Engineering Profession.
9. Work effectively as a member and leader in multidisciplinary team for a common goal.
10. Communicate effectively within a Profession and Society at large.
11. Appropriately incorporate principles of Management and Finance in one's own Work.
12. Identify educational needs and engage in lifelong learning in a Changing World of Technology.

DEPARTMENT OF COMPUTER ENGINEERING

PROGRAMME SPECIFIC OUTCOME

1. Formulate and analyze complex engineering problems in computer engineering (Networking/Big data/ Intelligent Systems/Cloud Computing/Real time systems).
2. Plan and develop efficient, reliable, secure and customized application software using cost effective emerging software tools ethically.

SARASWATI Education Society's
SARASWATI College of Engineering
Learn Live Achieve and Contribute
Kharghar, Navi Mumbai - 410 210.

(Approved by AICTE, recg. By Maharashtra Govt. DTE ,Affiliated to Mumbai University)

PLOT NO. 46/46A, SECTOR NO 5, BEHIND MSEB SUBSTATION, KHARGHAR, NAVI MUMBAI-410210

Tel. : 022-27743706 to 11 * Fax : 022-27743712 * Website: www.sce.edu.in

CERTIFICATE

*This is to certify that the requirements for the mini project report entitled "**Loan Prediction System**" have been successfully completed by the following students:*

Roll numbers	Name
03	Zeeshan Ansari
40	Kanchan Mengune
48	Binitdev Pandey
68	Adnan Shaikh

In partial fulfillment of Sem –V , **Bachelor of Engineering of Mumbai University in Computer Engineering** of Saraswati college of Engineering, Kharghar during the academic year 2021-22.

Internal Guide **External Examiner**

Prof. Shatabdi Bhalerao

Mini Project Co-ordinator **Head of Department**

Prof. Bhagyashri Sonawale Prof. Sujata Bhairnallykar

DECLARATION

I declare that this written submission represents my ideas in my own words and where others ideas or words have been included. I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

1. ZEESHAN ANSARI
2. KANCHAN MENGUNE
3. BINITDEV PANDEY
4. ADNAN SHAIKH

Date:

ACKNOWLEDGEMENT

After the completion of this work, words are not enough to express feelings about all those who helped us to reach goal.

It's a great pleasure and moment of immense satisfaction for us to express my profound gratitude to **MiniProject Guide**, **Prof. Shatabdi Bhalerao**, whose constant encouragement enabled us to work enthusiastically. His perpetual motivation, patience and excellent expertise in discussion during progress of the project work have benefited us to an extent, which is beyond expression.

We would also like to give our sincere thanks to **Prof.Sujata Bhairnallykar, Head of Department**, and **Prof. Bhagyashri Sonawale ,Mini Project co-coordinator** from Department of Computer Engineering, Saraswati college of Engineering, Kharghar, Navi Mumbai, for their guidance, encouragement and support during a project.

I am thankful to **Dr. ManjushaDeshmukh, Principal,**Saraswati College of Engineering, Kharghar, Navi Mumbai for providing an outstanding academic environment, also for providing the adequate facilities.

Last but not the least we would also like to thank all the staffs of Saraswati college of Engineering (Computer Engineering Department) for their valuable guidance with their interest and valuable suggestions brightened us.

1. ZEESHAN ANSARI
2. KANCHAN MENGUNE
3. BINITDEV PANDEY
4. ADNAN SHAIKH

ABSTRACT

In today's fast growing world everyone gets to the point where they need loan for something, in Banking, Car or any other system which gives their customers loan benefits according to their status, as technology is getting better day by day customers can apply for the loan through online system (Website or App) because of this large data (10,000+) can be piled up it will be difficult for any employee to evaluate all customers information (as well as giving additional salary just for passing loan) and depending on that approving their loan, it is necessary to have something at hand which can check whether to approve loan or not without interference of human. So, keeping this in mind we created ML (Machine Learning) Loan Prediction Project which can approve loan to customers depending on the necessary information they provide. This project goes through many steps from pre-processing data in proper format to transforming it to suitable format for applying ML algorithms, after transformation different ML algorithms are applied, selecting patterns of the algorithm which have highest accuracy and finally applying patterns on Test Dataset to predict whether to approve loan or not.

Table of Contents

List of Figures

CHAPTER 1

INTRODUCTON

1.1 GENERAL

Credits are the core business of banks. The main revenue comes directly from the mortgage's interest. The loan firms grant a loan after an intensive process of confirmation and authentication. However, they still don't have guarantee if the applicant is able to repay the loan with no complications.

We'll build an analytical model to predict if an applicant is able to repay the loaning firm or not. We will prepare the data using Jupyter Notebook and use various models to predict the target variable. Housing Finance firm deals in all home mortgages. They have presence across all town, semi urban and countryside areas. Client first apply for home loan after that firm validates the client eligibility for mortgage.

The Firm wants to systematize the credit eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Edification, Number of Dependents, Revenue, Credit Amount, Credit History and others. To systematize this process, they have given a problem to recognize the

clients sections, those are qualified for credit amount so that they can precisely target these clients.

1.2 OBJECTIVE AND PROBLEM STATEMENT

This project aims to build a predictive model to help the banks in determining if an applicant will be able to repay the loan or not.

FUNCTIONALITY

- Data analysis through univariate and bivariate analysis using pandas, matplot and seaborn.
- Data preprocessing (making data suitable for preparing model)using numpy and pandas
- Measuring data correlation
- Training different models using following algorithm:

    1. Logistic Regression

    2. Decision Tree Classifier

    3. Random Forest using Grid Search

    4. XG Boost and ADA Boost

- Applying same models on test data set to predict unknown loan status
- Saving submission file in .CSV format

CHAPTER 2

METHODOLOGY

2.1 Data Set

Attribute	Values	Type
Loan_ID	Unique ID (Nominal)	Original
Gender	Male/Female (Nominal)	Original
Married	Yes/No (Nominal)	Original
Dependents	Number family member depends on client (Numeric)	Original
Education	Graduate/Undergraduate (Ordinal)	Original
Self_Employed	Yes/No (Nominal)	Original
ApplicantIncome	Primary Income (Numeric)	Original
CoapplicantIncome	Secondary Income (Numeric)	Original
LoanAmount	Loan Amount in thousands	Original

	(Numeric)	
Loan_Amount_Term	Term of lean in months (Numeric)	Original
Credit_History	CIBIL Score (Binary)	Original
Property Area	Urban/Semi/Rural (Ordinal)	Original
Loan_Status	Loan Approval (Binary)	Original
Total Income	Applicant Income + Co-applicant Income (Numeric)	Derived
EMI	Total Income - (Loan Amount/Loan Amount Term)*1000 (Numeric)	Derived
Balanced Income	Total Income - EMI (Numeric)	Derived

2.2 ALGORITHMIC DETAILS

Algorithm:

Following ML Algorithms are going to be used in Loan prediction system:

1) Logistic regression.

2) Decision tree.

3) Random Forest.

4) XGBOOST

5) ADABOOST

Logistic Regression

I) Logistic regression is a classification algorithm which uses logistic function or sigmoid function which takes value in the range [0, 1].

II) Sigmoid function: f(x) = $\frac{1}{1+e^{-x}}$ ---------- (1)

III) Projecting extreme values i.e f: $[-inf, +inf]$ → [0,1].

IV) Logistic regression work much like Linear regression taking Input values (x) are combined linearly using weights or coefficient values

(referred to as the Greek capital letter Beta) to predict an output value (y=f(x)).

V) Equation of logistic regression:

$$y = \frac{e^{b0 + b1*x}}{1 + e^{b0 + b1*x}} = \frac{1}{1 + e^{-(b0 + b1*x)}} \text{---------- (2)}$$

VI) Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in input data has an associated b coefficient (a constant real value).

VII) The coefficients (Beta values b) of the logistic regression algorithm must be estimated from your training data. This is done using maximum-likelihood estimation, which uses maxima-minima technique to find value of parameters such that error rates are as low as possible.

e.g.: Consider we want to create a model which predict loan to be pass or not on basis of Income.

We will find probability of Loan pass (1) and rejected (0) depending on Income.

i.e. $p(x) = p(L = 1|Income = value) \frac{e^{b0 + b1*x}}{(1 + e^{b0 + b1*x})}$

(Loan pass when certain value limit in Income identified).

After simplifying this equation we get $\ln\left(\frac{p(x)}{1-p(x)}\right) = b0 + b1 * x.$ ----(3)

After applying maximum likelihood on given function

(Since, $p(x; b0, b1) = L(b0, b1; x)$) using our train set we will find value of b0 and b1 and we will apply these values with value of x to find corresponding value of y in test set. We will consider Loan pass if f(x) >= 0.5 and rejected if f(x) < 0.5.

$$y = \frac{e^{b0 + b1*X}}{(1 + e^{b0 + b1*X})}$$

$$y = \frac{e^{-100 + 0.6*150}}{(1 + e^{-100 + 0.6*X})}$$

$(consider, b0 = -100 \; and \; b1 = 0.6)$

$f(x) = y = 0.0000453978687$

$since, f(x) < 0.5 \; Loan \; is \; Rejected.$

Decision Tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

This algorithm split the node into n-nodes depending on best cost of split (lowest cost). Different type of cost function can be used, for classification decision tree we used Gini and Entropy, Gini index is given by:

$$\boldsymbol{G} = \mathbf{1} - \sum\left(\boldsymbol{pk}(\mathbf{1} - \boldsymbol{pk})\right) \text{ ---------- (4)}$$

Here, pk is proportion of same class inputs present in a particular group. A perfect class purity occurs when a group contains all inputs from the same class, in which case pk is either 1 or 0 and G = 0, where as a node having a 50–50 split of classes in a group has the worst purity, so for a binary classification it will have pk = 0.5 and G = 0.5.

Entropy is given by: $\boldsymbol{E} = -(\sum \boldsymbol{pklog}(\boldsymbol{pk}))$ ---------- (5)

Here, pk means same as in Gini index but the value lies from [0, 1]. A perfect class purity occurs when a group contains all inputs from the same class, in which case pk is either 1 or 0 and E = 0, where as a node having a 50–50 split of classes in a group has the worst purity, so for a binary classification it will have pk = 0.5 and E = 1.

Random Forest

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging and bootstrap (0.632) sampling. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. We will be using Greed Search to tune the model i.e. finding the best values for hyper parameters and RepeatedStratifiedKFold sampling which is an iterative sampling method combining both the Stratified and KFold Cross Validation method.

XGBOOST

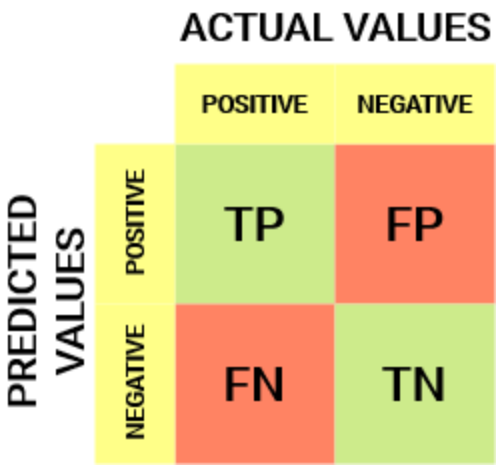
This algorithm only works with the quantitative variable. It is a gradient enhancing algorithm which forms solid rules for the model by boosting the weak beginners to a strong learner. It is a fast and well-organized algorithm which newly conquered machine learning because of its high performance and speed.

ADA BOOST

ADA Boost algorithm, short for Adaptive Boosting, is an Enhancing method used as a Collective Method in Machine Learning. It is called an Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to imperfectly classified instances. Boosting is used to lessen the bias as well as variance for supervised learning. It

works on the principle of learners growing sequentially. But for the first, each subsequent beginner is grown from previously grown beginners. In simple words, weak beginners are transformed into strong ones.

True Positive Rate and False Positive Rate



Confusion Matrix

From the confusion matrix, we can derive some important metrics

Sensitivity / True Positive Rate / Recall:

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

Sensitivity tells us what proportion of the positive class got correctly classified.

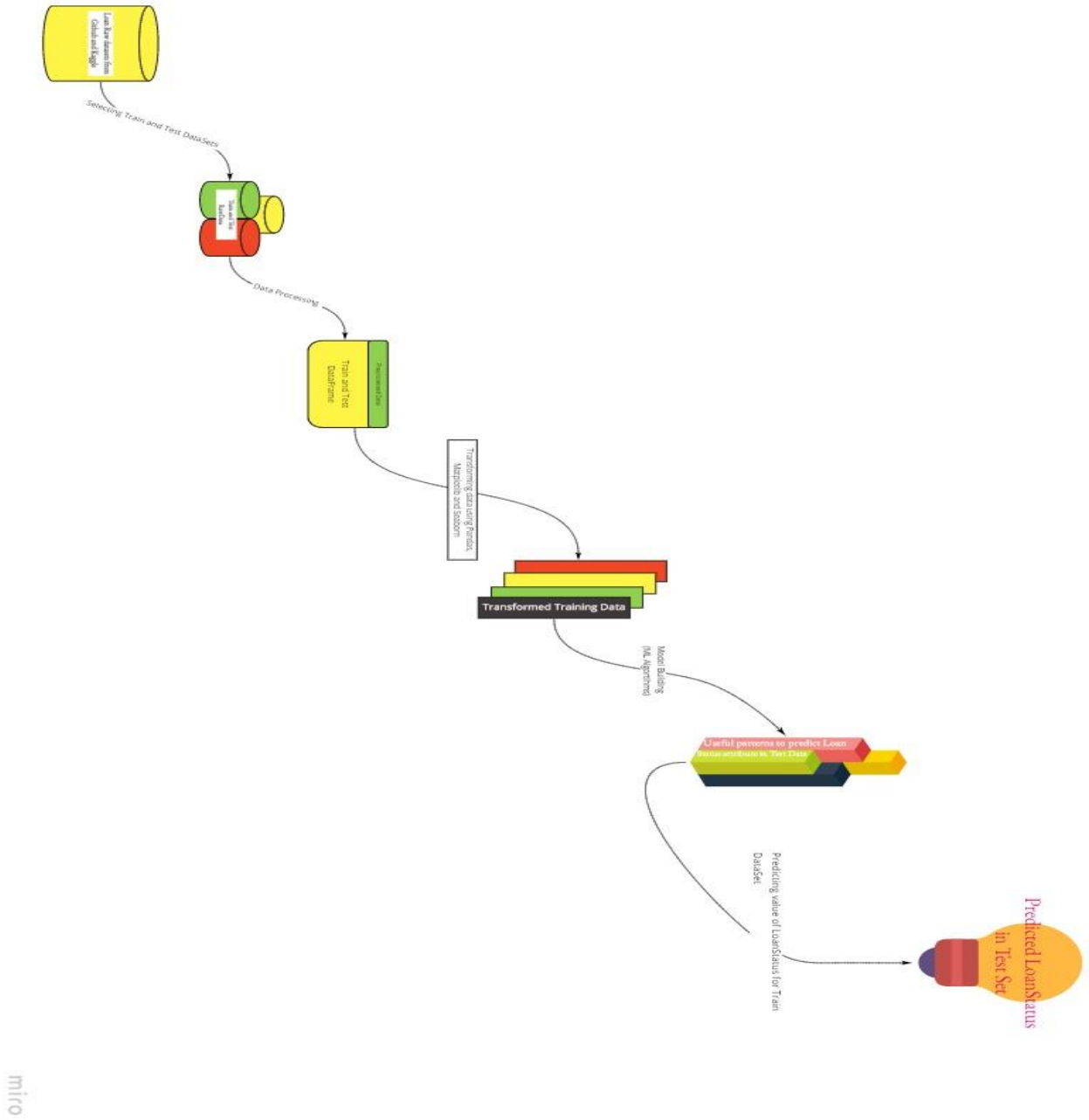
False Positive Rate:

$$FPR = \frac{FP}{TN + FP} = 1 - Specificity$$

FPR tells us what proportion of the negative class got incorrectly classified by the classifier.

A lower FPR is desirable since we want to correctly classify the negative class.

2.3 DESIGN DETAILS:



CHAPTER 3

SYSTEM REQUIREMENT

3. HARDWARE AND SOFTWARE REQUIREMENTS

3.1 HARDWARE REQUIREMENTS:

1. RAM : 8 GB+, 2700 MHz, DDR4 Minimum
2. Hard Drive : SSD Required

3.2 SOFTWARE REQUIREMENTS:

1. Anaconda 3
2. Jupyter Notebook
3. Google collab
4. Python 3.7
5. Pandas
6. Numpy
7. Seaborn
8. Skit-Learn

CHAPTER 4

IMPLEMENTATION AND RESULTS

IMPLEMENTATION AND RESULTS:

SR No.	Algorithms	Accuracy	Running Time	AUC Score
1.	Logistic Regression	81.11%	133ms ± 3.05ms	0.83
2.	Decision Tree	78%	34ms ± 915 µs	0.71
3.	Random Forest	81.40%	4.27s ± 83.2ms	0.69
4.	XGBOOST	76.54%	211ms ± 10ms	0.70
5.	ADABOOST	81.11%	228ms ± 5.83ms	0.74

Table 1.

Random Forest, Logistic Regression and ADABOOST gave the highest accuracies, Random Forest estimation time is highest and lowest AUC score because of modelling of large number of decision trees but these factors help the Random Forest to achieve highest accuracy and Logistic Regression estimation time is lowest out of these three classifiers which can be useful in evaluation of large data set

ROC AUC CURVE

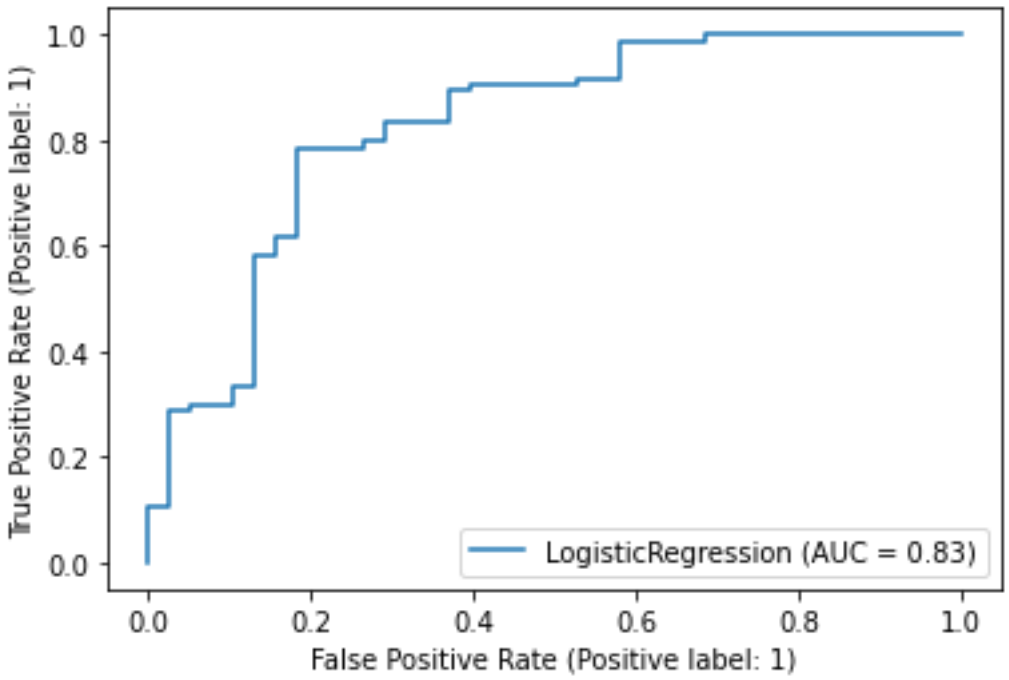


Figure 1. Logistic Regression

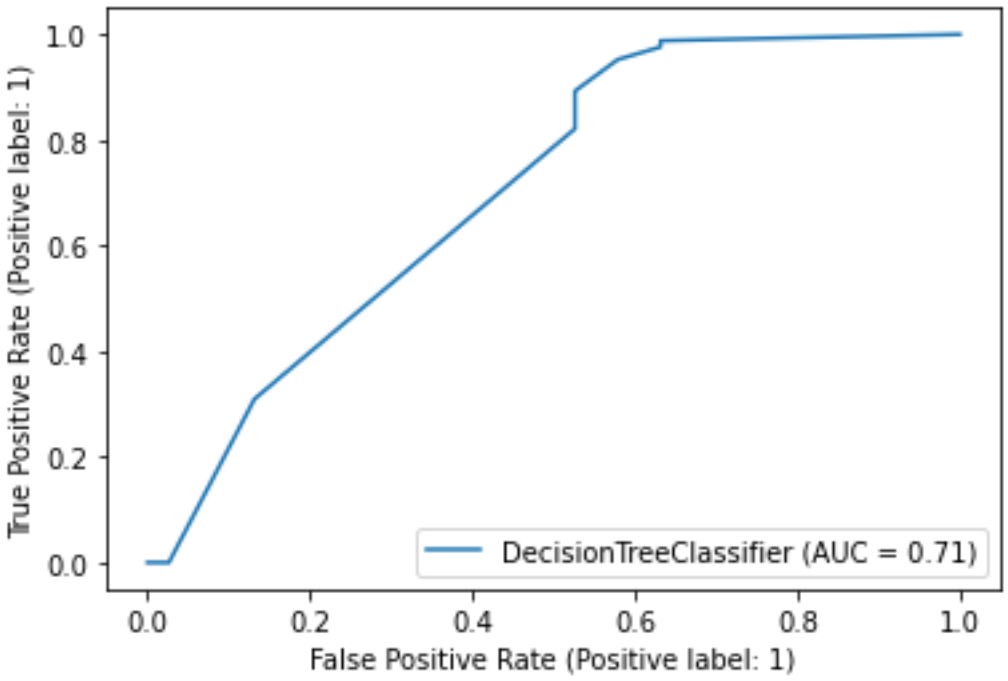


Figure 2. Decision Tree Classifier

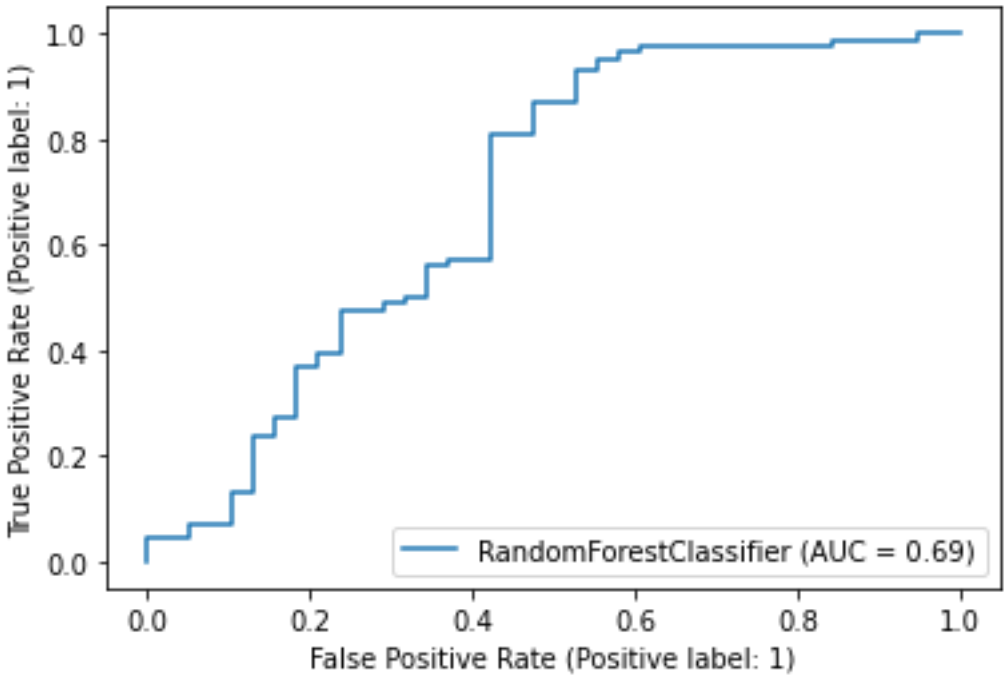


Figure 3. Random Forest Classifier

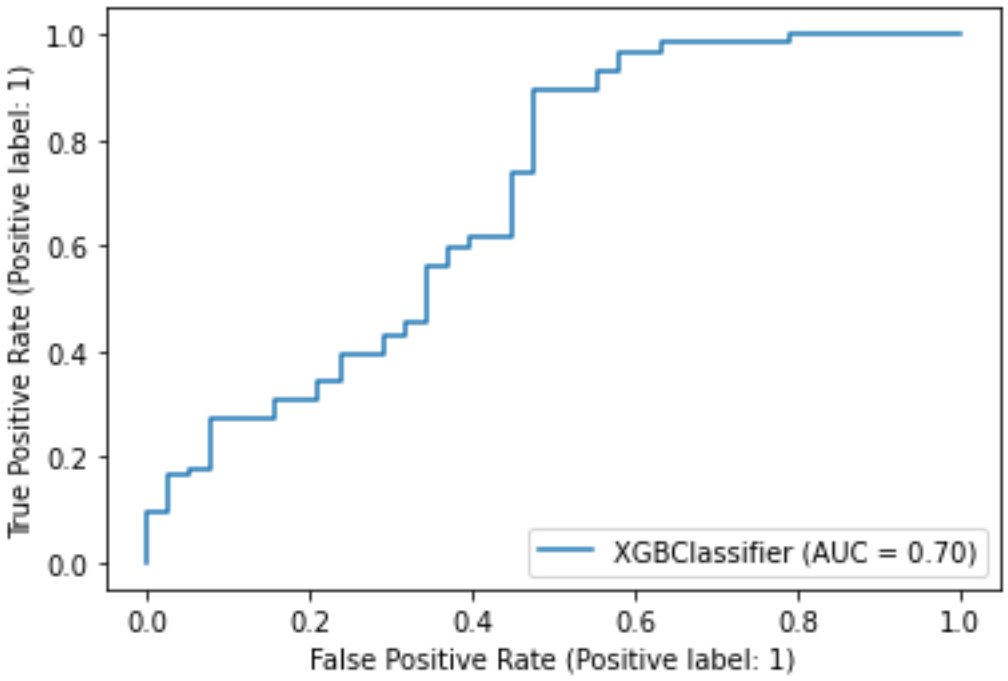


Figure 4. XGBoost Classifier

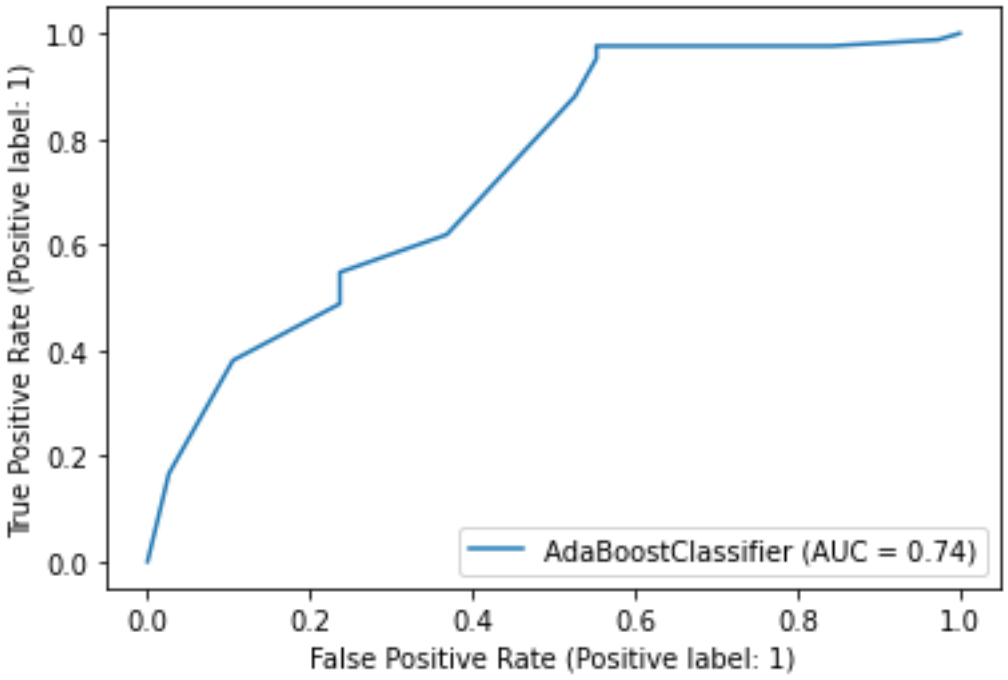


Figure 5. ADA Boost Classifier

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

CONCLUSION:

We did exploratory data Analysis on the features of this dataset and observe how each feature is distributed. We did bivariate and univariate analysis to see impact of one another on their features using charts. We analyzed each variable to check if data is cleaned and normally distributed. We cleaned the data and removed NA values we also generated hypothesis to prove an association among the independent variables and the Target variable. And based on the results, we assumed whether or not there is an association. We calculated correlation between independent variables and found that applicant income and loan amount have significant relation. We created dummy variables for constructing the model we constructed models taking different variables into account and found through odds ratio that credit history is creating the most impact on loan giving decision finally, we got a model with co-applicant income and credit history as independent variable with highest accuracy. We tested the data and got the accuracy of 81%. t. In future we can integrate these models with credit system app, we can increase model accuracy by introducing attributes and can also predict future losses.

REFERENCES

[1] Sara Zahia, Boujemâa Achchab “Modeling Car Loan Prepayment” International Workshop on Statistical Methods and Artificial Intelligence (IWSMAI) April 6-9, 2020, Warsaw, Poland

[2] Jasmin Kevric, Samed Jukic, Abdulhamit Subasi “An effective combining classifier approach using tree algorithms for network intrusion detection” springer article

[3] Eman A. Toraih, Rami M. Elshazli, Mohammad H. Hussein, Abdelaziz Elgam, Mohamed Amin, Mohammed El-Mowafy, Mohamed El-Mesery, Assem Ellythy, Juan Duchesne, Mary T. Killackey, Keith C. Ferdinand, Emad Kandil, Manal S. Fawzy “Association of cardiac biomarkers and comorbidities with increased mortality, severity, and cardiac injury in COVID-19 patients: A meta-regression and decision tree analysis” in Journal of Medical Virology · June 2020

[4] Guestrin, Tianqi Chen “XGBoost: A Scalable Tree Boosting System” research article

[5] Weiwei lin, Ziming wu, Longxin Lin, Angzhan Wen, and Jin li “An Ensemble Random Forest Algorithm for Insurance Big Data Analysis” Digital Object Identifier 10.1109/ACCESS.2017.2738069

[6] Iftikhar Ahmaed, Mohmmad Basheri, Muhmmad Javed Iqbal, and Aneel Rahim “Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection” Digital Object Identifier 10.1109/ACCESS.2018.2841987

[7] Robert E. Schapire “The Boosting Approach to Machine Learning an Overview” AT&T Labs Research Shannon Laboratory 180 Park Avenue, Room A203 Florham Park, NJ 07932 USA www.research.att.com/ schapire December 19, 2001

[8] R. Bekkerman. The present and the future of the kdd cup competition: an outsider’s perspective.

[9] R. Bekkerman, M. Bilenko, and J. Langford. Scaling Up Machine Learning: Parallel and Distributed Approaches. Cambridge University Press, New York, NY, USA, 2011.

[10] J. Bennett and S. Lanning. The netflix prize. In Proceedings of the KDD Cup Workshop 2007, pages 3–6, New York, Aug. 2007.

[11] L. Breiman. Random forests. Maching Learning, 45(1):5–32, Oct. 2001.

[12] https://pandas.pydata.org/docs/

[13] https://numpy.org/doc/

[14] https://scikit-learn.org/0.21/documentation.html

[15] https://matplotlib.org/stable/contents.html

[16] https://seaborn.pydata.org/