WEB IMPLEMENTATION FOR BANK TO PREDICT ELIGIBLE CUSTOMER USING MACHINE LEARNING AND AWS

Submitted by

Asfand Ali **(2068882)**
Mahesh Yadav Udutha **(2056251)**

Professor
Massimo Mecella

Submitted in the partial fulfilled of the requirement for the Laboratory of Advance Programming in Computer Science and Engineering

Facoltà di ingegneria dell'informazione, informatica e statistica Engineering in Computer Science Sapienza, Rome, Italy

SAPIENZA UNIVERSITY OF ROME ITALY

Academic Year 2022/2023

ABSTRACT

Getting loans from financial organizations has become very common in today's society. For a wide range of reasons, many individuals apply for loans every day. However, not all candidates are trustworthy, and not everyone will be accepted. There are instances where borrowers fail to repay the majority of their loans to the bank, which causes enormous financial losses. Choosing to approve a loan carries a significant amount of danger. Therefore, the goal of this project is to use machine learning techniques on the loaded data set from kaggle to extract key information, and then forecast whether or not a customer will be able to repay the loan. Therefore, the objective is to determine whether you can predict the customer defaulting or not. Among the most prominent research areas in the banking and insurance industries, loan prediction for eligible customers is one of the most important in the modern environment identifying and analyzing the patterns of the dataset plays a vital role in this era. The loan prediction involves the application of various machine learning algorithms. This project will be developed using machine learning algorithms such as logistic regression, K Nearest Neighbor, Decision tree. Python programming language in Visual Studio code will be used for the implementation using GitHub and then finally deployment will be done on AWS Cloud. The proposed system expects to deliver high accuracy results and moderate loss for training and validate data.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

People want to apply for loans via the internet as data volumes grow due to digitization in the banking sector. Artificial intelligence (AI), as a common method of information scrutiny, is gaining popularity. Individuals from multiple sectors are using AI calculations to solve problems based on their industry knowledge. Banks are having a difficult time getting loans approved. Every day, there are numerous applications that bank employees must manage, and the possibility of errors is high. Most banks profit from loans but choosing deserving customers from a large number of applications is risky. Just a single blunder can result in a massive loss for a bank. Almost every bank's main business is loan distribution. This project aims to provide a loan to a deserving applicant from among all applicants. An efficient and non-biased system that saves the bank time checks every applicant on a priority basis. All other customer formalities are completed on time by bank officials, which benefits the customers. The best part is that it is beneficial to both banks and applicants[7]. Customers apply for loans first, followed by the company/bank verifies their eligibility for the loan. The company/bank then wishes to automate the loan eligibility process (in real time) based on the information provided by the customer while filling out the application form. Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and other details are included. This project used previous customer data from various banks to whom loans were approved based on a set of rules. As a result, the machine learning model is applied to that method to obtain approximate values. This project's primary goal is to forecast loan safety. KNN, logistic regression, and decision tree algorithms are used to predict loan safety. Data is expanding day by day scheduled to digitization People like to apply for loans in the banking sector throuugh the internet. For intelligence research, we are getting more attention more and more. Using individuals from different companies AI Computation handles dependent problems their production information. Banks face serious problems Loan approval issues. there are many everyday

Applications that are difficult for banks to administer employees, and the chances of some errors are also higher. Most banks profit from loans, but the choice is risky Attracting customers from the number of applications. One mistake can cost a bank a huge loss. Credit distribution is First's core business for all banks. This project aims to provide consumption A worthy applicant among all applicants. Efficient An unbiased system that cuts bank time Priority will be given to each applicant. Bank We handle all other customer guideline This has a positive effect on the customer. the best Part is that it is efficient for both banks and applicants. This system enables launching specific apps that.

deserve to be given special approval. Prediction characteristics include things like: Gender, Married, Dependents, Education, Self-Employed, Applicant Income, Co-Applicant Income, Loan Amount, Loan Term, Credit History, Property Area, and Loan Status

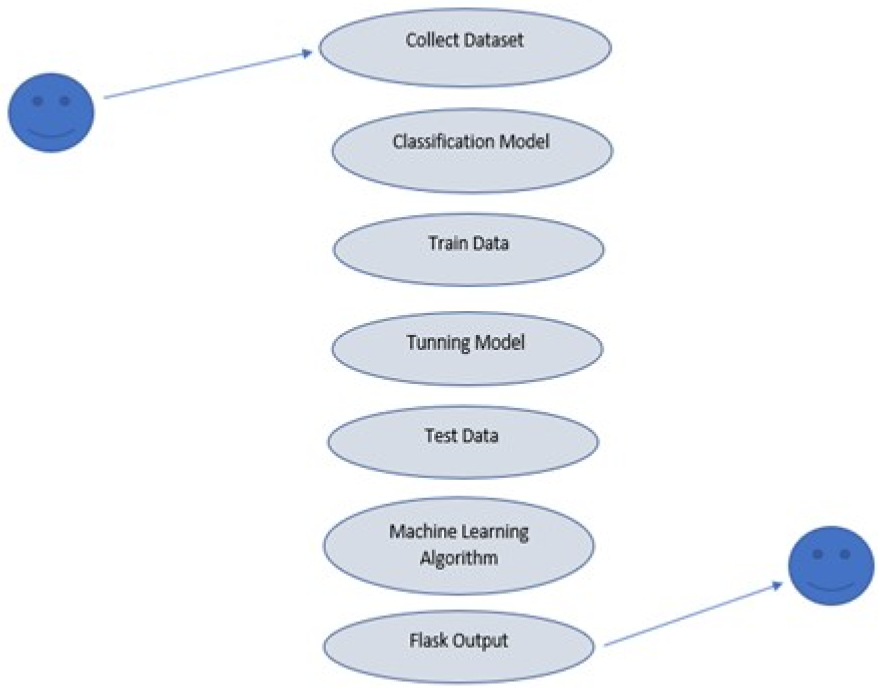


Figure 1.1: UML class diagram of Loan Prediction

1.2 PROBLEM STATEMENT

In the area of banking and finance, this project seeks to offer a solution for automating a crucial process. Numerous different kinds of personal and business loans are offered by banks and housing finance companies all over the globe. When a company applies for a loan, these firms verify the company's eligibility. Our project will automate this process by using machine learning to analyze the online form that customers must fill out in order to determine whether or not they are qualified for a loan. Details on this form include the applicant's sex, marital status, qualifications, dependents' information, yearly income, loan amount, credit history, and other information. Many banks is associated with different kinds of loan like house loan educational loan personal loan vehicle loan etc.

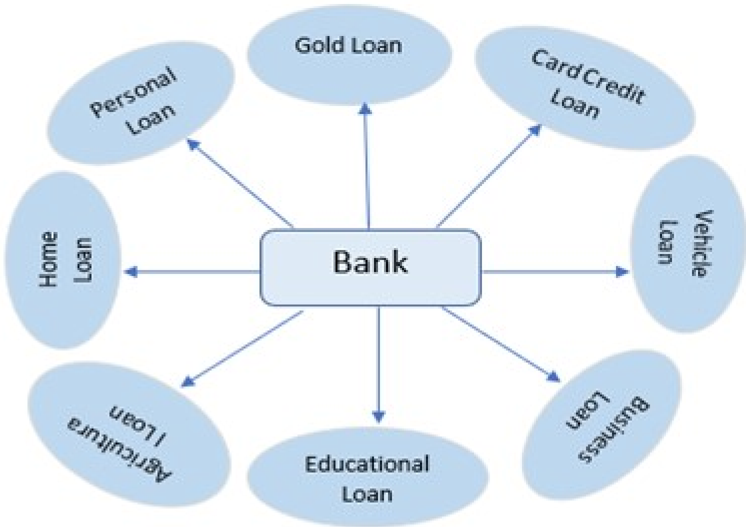


Figure 1.2: Different types of loan offered by Bank

has having different areas rural/urban and firstly a customer tries to get the loan from bank and banks check the validates for eligibility for the loan and of approve like marital status, gender, education, income, amount, credit history, and many different peoples are given the form to fill the applicant details. Therefor taking the input applicants details to verify whether applicant is eligible to apply loan or not. After build up ML model web applicant is developed for user interface that allows to user can directly check she/he is eligibility to get loan or not by entering given details.

1.3 MOTIVATION

Now a days loan prediction is very common real life problems that every bank faces leading operation customer can save many hours when loan approval process is automated when bank has robust model to predict which loan customer's loan it should be approve or reject in order to reduce loan defaulter.

CHAPTER 2

LITERATURE REVIEW

2.1 REVIEW OF THE LOAN PREDICTION

classification model is used to predict outcomes Yes or No. We have use supervised machine learning to design the pattern of dataset which will help the algorithm to predict anyother new dataset. we have used test data set, which is used to evaluate the performance of built algorithms. Classification algorithms create predictors while learning. The training dataset is used to teach the classifier and the class designations that go with them (variables). The classifier is used for categorization in classifying. Here Classification algorithms' accuracy is estimated using test data. The components of classification can be transferred to the new datasets tuples if the accuracy is deemed adequate, which will be determined by the the degree of confidence or tolerance. The classifier-training method utilizes these pre-classified examples to identify all of the variables (parameters) required for correct judgment or analysis. The algorithm then encodes these factors (variables) into a data structure.Neural networks, which have been successfully applied in a variety of areas such as banking institutions, database mining, and computer security, can also be used to solve the default prediction problem. Recently, many machine learning ' standard-oriented approaches to debt default prediction have been developed[**?**] A significant portion of these approaches are based on Logistics Regression, classification, or a combination of these and other ensemble methods. Preliminary literature review reveals that previous studies focused on the loan default of online lending clubs and online lending institutions for credit risk assessment by using historical data of customers and FICO score to measure the risk associated with the borrowers, implying that the data collected online was insufficient and difficult to clean. Loan default trends have been long studied from a socio-economic viewpoint. Most research papers believe in empirical modelling of these complex systems in order to be able to predict the loan default rate for a specific individual. The use of machine learning for such jobs is a trend which we are perceiving now Previously lenders and financial institutions employed highly professional people to evaluate a borrower's worthiness before approving or rejecting a loan application. However, recently these institutions have started employing various models for loan evaluation in order to decide whether to reject or approve a loan to a borrower based on their credit score and ability to repay a given loan. They chose models based on machine learning models and artificial neural networks for accurately predicting credit defaulters among borrowers. Many peer-to-peer lending platforms encourage borrowers to form online groups in order to reduce information asymmetry. The practical results of the research show that the closer the location of the borrowers in the group, the lower the likelihood of default. In addition, if they have a definite connection in real life, joining the group will significantly reduce the borrower default probability. According to the borrower's social 8 media disclosure, borrowers with many friends are more likely to obtain loans, and the default probability is very low. Because of the vast amount of data that the banking industry works with, data analysis and transformation into useful knowledge has become a job that is beyond human capability. Data mining techniques can be adopted in answering business problems by finding patterns, associations and correlations which are masked in the business information stored in the databases. By employing data mining techniques to investigate the patterns and trends associated with them, bank executives can predict, with high accuracy, how customers will respond to adjustments in interest rates, which customers are likely to receive new product offers, which customers will be at a higher risk for defaulting on a loan, Lenders in the lending industry and financial institutions typically assess loaners' repayment ability as well as the risks of lending money to borrowers from customer. Based on the repayment ability and risks associated, the lenders, especially the banks, can balance the interest rates of the loans which are issued to the client or borrowers.

2.2 MACHINE LEARNING

Machine learning is a type of algorithm that allows software applications to become more precisely unpredictable without being explicitly programmed. A subset of AI supports the notion that a system can learn from data, recognize patterns, and make decisions to achieve optimal solutions with minimal human intervention. There are two types of ML algorithms, supervised machine learning algorithms and unsupervised machine learning algorithms.

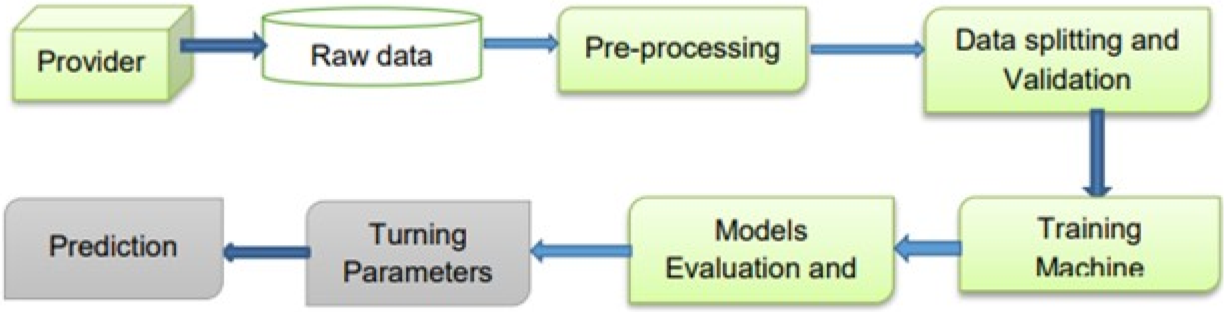


Figure 2.1: Machine Learning Process

2.3 MACHINE LEARNING PROCESS

2.4 EXISTING SYSTEM

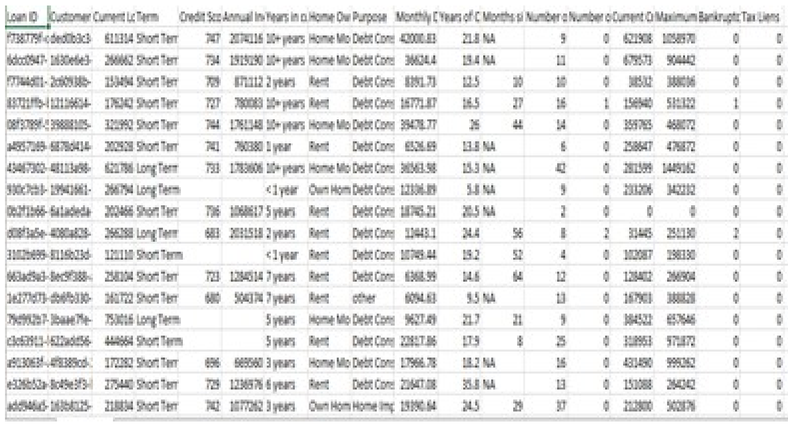
Existing systems uses traditional methods of Planning and strategic management for predicting the loan to the customer which delays the process as well as the outcome is not efficient. To overcome this solution in this project we have used Machine Learning algorithms which gain amazing results by using different machine learning models.

CHAPTER 3

DATA SET

The dataset for our implementation has been downloaded from the website name as Kaggle which has good collection public data. And dataset consists important features like loan _ID, gender, marital status, dependents, level of education, self-employment, applicant and applicant incomes, loan amount, loan amount and term, and credit history.

The titles of the columns with the int64 data type are Property Area. Each row contains values with Loan _ID being unique, Gender having either a masculine or female value, The dependents values are ‘ standard+, the credit history value is 0 or 1, the married and self-employed values are either yes or no values, the education input is either graduate or not graduate, and the remaining values are all of the numerical type[3]. The dataset consists of an acceptable loan application. The dataset consists 614 records with 13 attributes. It has labelled data for customer loan eligibility may become comes outcomes Yes/No. so Yes means customer are able to get loan and No means customer are not eligible to get loan.



Loan ID	Customer	Current Lc	Term	Credit Sco	Annual In	Years in c	Home Ow	Purpose	Monthly D	Years of C	Months si	Number o	Number o	Current C	Maximum	Bankruptc	Tax Liens
f738779f-c	ded0b3c3-	611314	Short Terr	747	2074116	10+ years	Home Mo	Debt Con	42000.83	21.8	NA	9	0	621908	1058970	0	0
6dcc0947-	1630e6e3-	266662	Short Terr	734	1919190	10+ years	Home Mo	Debt Con	36624.4	19.4	NA	11	0	679573	904442	0	0
f7744d01-	2c609338b-	153494	Short Terr	709	871112	2 years	Rent	Debt Con	8391.73	12.5	10	10	0	38532	388036	0	0
83721ffb-	12116614-	176242	Short Terr	727	780083	10+ years	Rent	Debt Con	16771.87	16.5	27	16	1	156940	531322	1	0
08f3789f-	39888105-	321992	Short Terr	744	1761148	10+ years	Home Mo	Debt Con	39478.77	26	44	14	0	359765	468072	0	0
a4957169-	6878d414-	202928	Short Terr	741	760380	1 year	Rent	Debt Con	6526.69	13.8	NA	6	0	258647	476872	0	0
43467302-	48113a98-	621786	Long Term	733	1783606	10+ years	Home Mo	Debt Con	36563.98	15.3	NA	42	0	281599	1449162	0	0
930c7cb3-	19941661-	266794	Long Term			< 1 year	Own Hom	Debt Con	12106.89	5.8	NA	9	0	231206	342232	0	0
0b2f1b66-	6a1adeda-	202466	Short Terr	736	1068617	5 years	Rent	Debt Con	18745.21	20.5	NA	2	0	0	0	0	0
d08f3a5e-	4080a828-	266288	Long Terr	683	2031518	2 years	Rent	Debt Con	12443.1	24.4	56	8	2	31445	251130	2	0
3102b699-	8116b23d-	121110	Short Term			< 1 year	Rent	Debt Con	10749.44	19.2	52	4	0	102087	198330	0	0
663ad9a3-	8ec9f388-	258104	Short Terr	723	1284514	7 years	Rent	Debt Con	6368.99	14.6	64	12	0	128402	266904	0	0
1e277d73-	db6fb330-	161722	Short Terr	680	504374	7 years	Rent	other	6094.63	9.5	NA	13	0	167903	388828	0	0
79d992b7-	3baae7fe-	753016	Long Term			5 years	Home Mo	Debt Con	9627.49	21.7	21	9	0	384522	657646	0	0
c3c63913-	622add56-	444664	Short Term			5 years	Rent	Debt Con	22817.86	17.9	8	25	0	318953	571872	0	0
a913063f-	4f8389cd-	172282	Short Terr	696	669560	3 years	Home Mo	Debt Con	17966.78	18.2	NA	16	0	431490	999262	0	0
e326b52a-	8c49e3f3-	275440	Short Terr	729	1236976	6 years	Rent	Debt Con	21647.08	35.8	NA	13	0	151088	264242	0	0
add946a5-	163b8125-	218834	Short Terr	742	1077262	3 years	Own Hom	Home Imp	19390.64	24.5	29	37	0	212800	502876	0	0

Figure 3.1: Train Data

Basically, Data test has many attributes such as salary, marital status loan without the loan approval status so if we want to deploy the test data we have build the trained data.

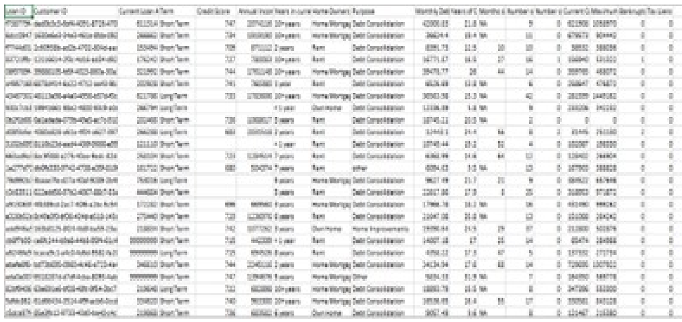


Figure 3.2: Test Data

The data gathering, preprocessing, selection of key features, training of the model, and testing of the model's performance are the main stages in this work. The information for this effort is gathered from a public repository using data from kaggle' customers. the 13 attributes of the gathered dataset. Output label, which indicates the state of the customer loan, is one of those attributes. One indicates that the applicant's credit has been approved, and zero indicates that it has not. Figure 2 depicts the project's process. Data collection is the first step, followed by data preprocessing, feature selection, data splitting into train and test parts, model training, performance assessment of the model, and comparison of the results[1]. Step 1 is data collection, step 2 is data preprocessing, step 3 is feature selection, step 4 is data splitting into train and test parts, step 5 is model training and step 6 is model performance evaluation and finally work concluded with comparison of decision tree, support vector machine, k nearest neighbor, and decision tree techniques accuracy in predicting customer loan eligibility Accuracy is one of the performance evaluation metrics for machine learning model. It gives the value which is the ration of number of correct predictions with total number of predictions[2].

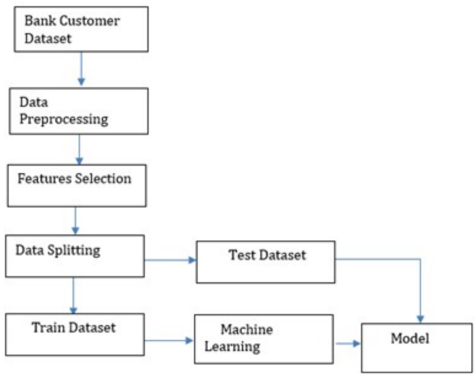


Figure 3.3: Workflow of the work

CHAPTER 4

METHODOLOGY

4.1 DATA CLEANING AND PROCESSING

Missing values in the dependent's column are filled with the column's average value, those in the applicant income attribute are filled with the column's lowest value, those in the loan _amount _term field are filled with the field's maximum term period, and those in the credit_history field are filled with 0. Data processing raises the level of data quality generally.

4.2 FEATURE EXTRACTION

The feature selection approach is used to choose more the right characteristics to determine a customer's credit eligibility. Application of the feature selection technique and analysis of variance on dataset to enhance the classifier's efficiency.

4.3 DATA SPLITTING

The foundation of a machine learning algorithm is data. In this study, the data is split in half, 80:20, for training and evaluating the model. The model was trained using the train dataset, and the model was validated using the test dataset portion.

4.4 DATA COLLECTION

The training set and testing set are created from the dataset that was gathered for forecasting loan failure. On the training set, the decision tree-created data model is applied, and test set forecasting is carried out based on the test pick fineness. Bank Loan eligibility prediction is conducted using attributes describes eligibility the attributes is associated with information of person's age, gender, educational background, properties, credit card information etc. and class attribute is bank loan eligibility prediction. where there are following attributes.

Variable	Description
Loan id	Unique loan id
Gender	Male/Female
Married	Yes/No
Dependent	Number of Dependent
Education	Graduate/ Undergraduate
Self employed	Yes/ No
Applicant income	Applicant Income
Loan Amount	Amount loan in Hundred/ Thousands
Loan Amount Term	Term of loan in Months
Credit History	Credits history meets guidelines
Property Area	Urban/Rural
Loan Status	Loan Approved Yes/No

Figure 4.1: Table Variable vs Description

4.5 TRAIN MODEL ON TRAINING DATA SET

The model is then trained using the two datasets - training and assessment. Our train collection is divided into two parts: tract train and testimony. The model is then trained on this portion to make the testimony part easier[4]. As we have the true sooth sayings for the testimony portion (which we do not have for the test dataset), we can verify our sooth sayings in this manner. So entries of new application will able to act test data which are filled at the time of submitting the application while performing the test model will determine whether loan will be approved or not on the basis of the various training dataset.

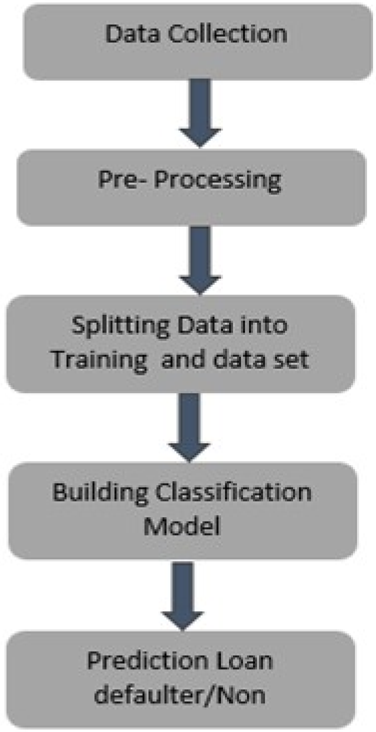


Figure 4.2: Chronology of Data

4.6 CORRELATING ATTRIBUTES

Based on the correlation between attributes, it was discovered that they were more likely to repay their debts[5]. Individual and essential attributes include property area, education, loan measure, and originally credit History, which is regarded as important. In the Python platform, the characteristics are associated using scatter-plot and Boxplot. The approach is used:

The proposed model is used to determine the output, i.e., predicting whether the loan will be authorized in the end.

In the following section, we get the results after finally entering the data of customers into web page built on flask server..

Proposed Model

We have developed the software application for approval/not approval of loan and we have also used machine learning to have symmetric model from existing dataset and it has been used for test the dataset consists two data train and dataset so model has been used testing dataset for result and Decision Tree uses and builds model for existing training the dataset.

4.7 SELECTING THE MODELS

4.7.1 Logistic Regression

A classified dependent variable's outcome is predicted using logistic regression. As a result, the conclusion must be categorical or distinct. It can be Yes or No, true or False, and so on, but instead of providing the precise values as 0 and 1, it provides the probabilistic values that fall between 0 and 1.

4.7.2 Decision Tree classifier

It is a technique for obtaining all possible solutions to an issue based on constraints[6]. Because the model's depiction resembles a tree, it is known as a decision tree. Figure 3 depicts the design of a decision tree model. It is a tree-structured classifier in which internal nodes reflect dataset characteristics, branches represent judgment criteria, and each leaf node symbolizes the result. it uses for asking the loan approved or not.

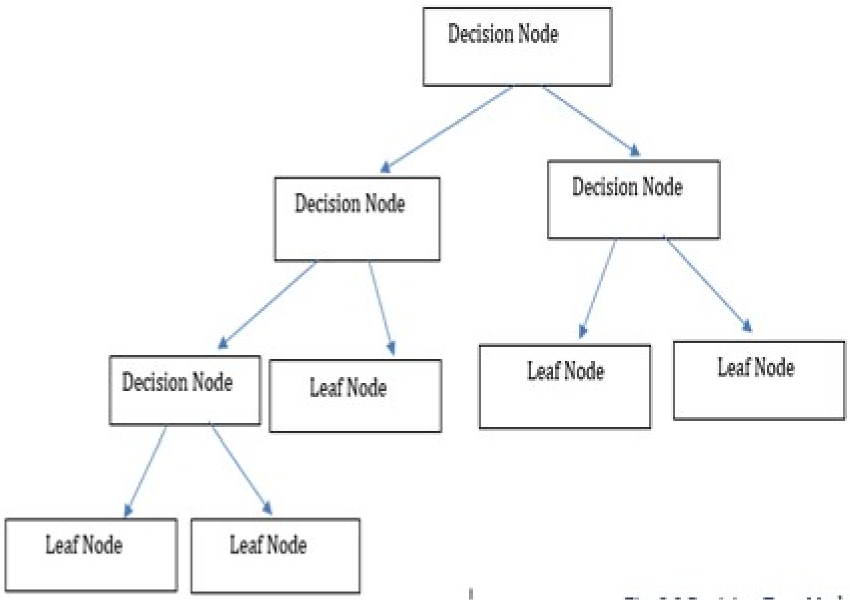


Figure 4.3: Decision Tree Model

CHAPTER 5

SUMMARY

5.1 RESULTS

In this section we can see the attributes are displayed in a form manner and this is need to be filled and according to the given values either loan will be approved or rejected. In this section, We have used machine learning algorithms on a loan dataset and predicted the result. using HTML, CSS with python jupyter notebook at the flask server. It shows the loan prediction system of the applicants based on the value entered by the bank employee.

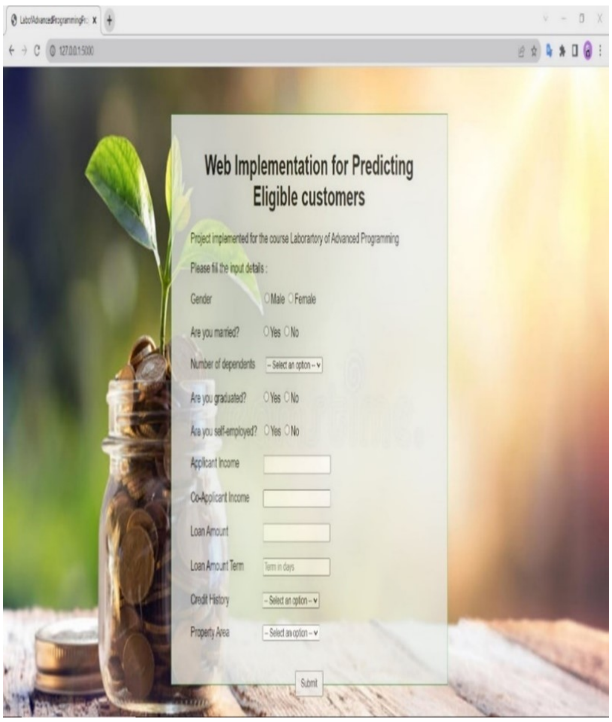


Figure 5.1: Final Layout

The attributes is select gender either male or female second attributes is marital status married or single after dependent attributes it means applicant is dependent some attributes are education of customer employment status applicant income loan amount term credit history residential area of applicant and show the status of loan.

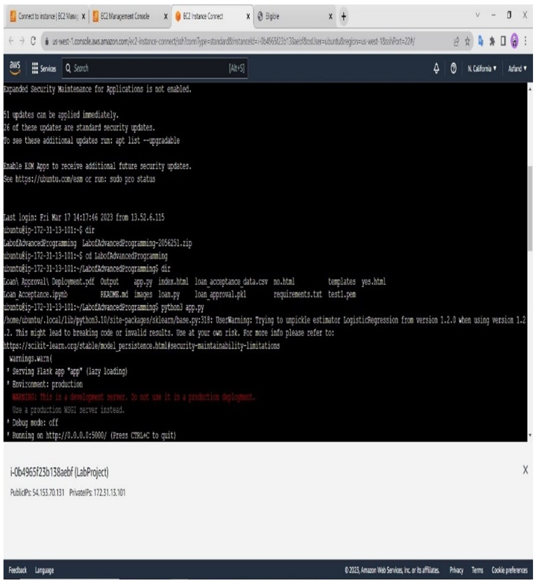


Figure 5.2: AWS Output

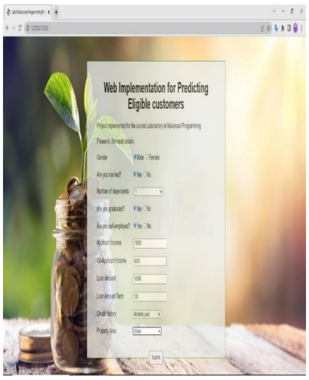


Figure 5.3: Prediction Page Area Rural

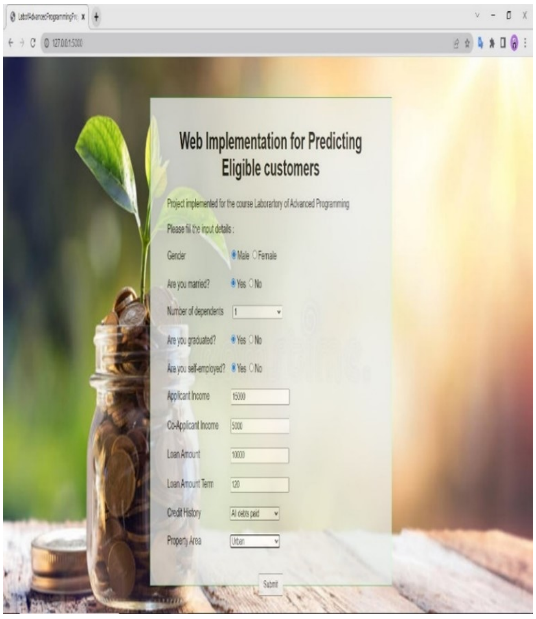


Figure 5.4: Prediction Page Area Urban

After filling in the mentioned details, we can click the submit button and there is used for proposed. Model which will give us predict. output it has been observed two cases loan approval and Loan. rejection based on different. attributes. The dataset has been used in this work hence attributes contribute to be better. to be an accurate real time prediction.

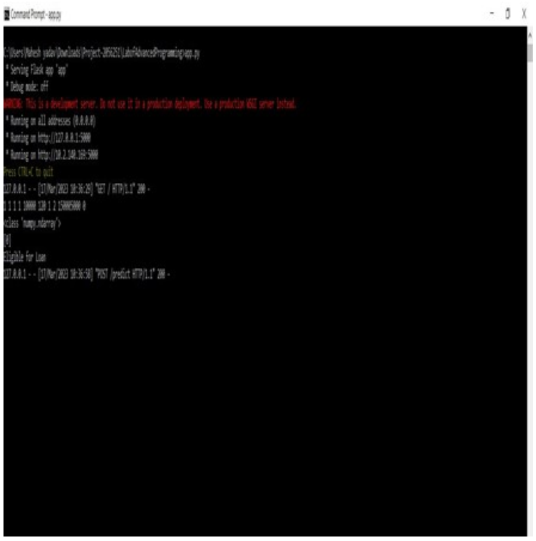


Figure 5.5: Cmd output eligible

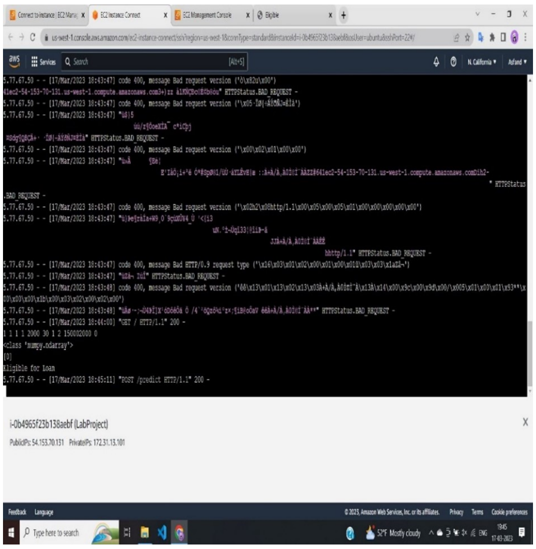


Figure 5.6: Console Output Eligible

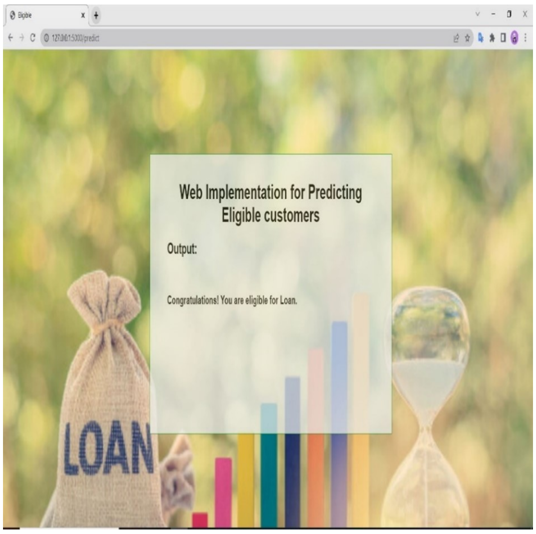


Figure 5.7: A case of Loan Approval

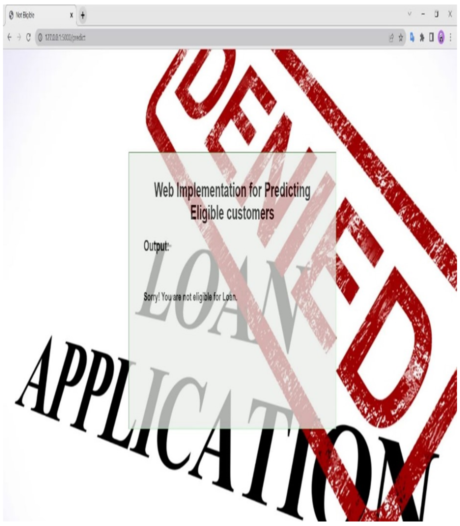


Figure 5.8: A case of Loan Rejection

So the result has been obtained according out train and test the data which our data has been measured as per requirement.

5.2 CONCLUSION

At the start, we have cleaned the dataset. Then exploratory data analysis and feature engineering was performed. Then different model were created which predicted whether the applicant would repay the loan or not.

5.3 LEARNING EXPERIENCE

The project development provided us with a sense of new technologies that we were not familiar with at the beginning of this project. we have learned to work with Jupyter notebooks, Scrum project management, Functional point Analysis, COCOMO-2, Github and usage of different Python libraries for Machine Learning while deployment on AWS EC2. Plus, we understood the concepts of Machine Learning by building models.

5.4 FUTURE ENHANCEMENT

Using different machine learning techniques can be implemented and to get good results so machine learning model is valuable for customer loan eligibility prediction and all existing model doesn't give 100 percent accuracy so it means there is need to design the innovative model for customer loan eligibility prediction .In the future a deep model is about propose for customer loan eligibility for future work.

REFERENCES

[1] Kumar Arun, Garg Ishan, and Kaur Sanmeet. Loan approval prediction based on machine learning approach. *IOSR J. Comput. Eng*, 18(3):18–21, 2016.

[2] Anshika Gupta, Vinay Pant, Sudhanshu Kumar, and Pravesh Kumar Bansal. Bank loan prediction system using machine learning. In *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, pages 423–426. IEEE, 2020.

[3] Ch Naveen Kumar, D Keerthana, M Kavitha, and M Kalyani. Customer loan eligibility prediction using machine learning algorithms in banking sector. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, pages 1007–1012. IEEE, 2022.

[4] Mohammad Ahmad Sheikh, Amit Kumar Goel, and Tapas Kumar. An approach for prediction of loan approval using machine learning algorithm. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 490–494. IEEE, 2020.

[5] Mohammad Ahmad Sheikh, Amit Kumar Goel, and Tapas Kumar. An approach for prediction of loan approval using machine learning algorithm. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 490–494. IEEE, 2020.

[6] Vishal Singh, Ayushman Yadav, Rajat Awasthi, and Guide N Partheeban. Prediction of modernized loan approval system based on machine learning approach. In *2021 International Conference on Intelligent Technologies (CONIT)*, pages 1–4. IEEE, 2021.

[7] CN Sujatha, Abhishek Gudipalli, Bh Pushyami, N Karthik, and BN Sanjana. Loan prediction using machine learning and its deployement on web application. In *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pages 1–7. IEEE, 2021.