**EDA LOAN STATUS PREDICTION**

**MACHINE LEARNING ALGORITHM**

**By Team 17:**

Monika K J 4BD17CS066

Nanditha K N 4BD17CS070

Spoorthy M N 4BD17CS107

Navya Gadina 4BD17CS072

Nethravathi B 4BD17CS404

**TABLE OF CONTENTS**

1. Introduction
   1. Overview

1.2 Purpose

1. Literature Survey
   1. Existing Problem

2.2 Proposed Solution

1. Theoretical Analysis
   1. Block Diagram

3.2 Software Designing

1. Experimental Investigation
2. Flowchart
3. Result
4. Advantages and Disadvantages
5. Applications
6. Conclusion

10. Future Scope

**1.INTRODUCTION**

There are various areas in which Machine Learning can be used in financial sectors like customer segmentation and profitability, high risk loan applicants, predicting payment default, marketing, credit analysis, ranking investments, fraudulent transactions, optimizing stock portfolios, cash management and forecasting operations, most profitable Credit Card Customers and Cross Selling. There are many different types of loans you have to take into account when you’re looking to borrow money and it’s important to know your options. Loan categorization refers to the process of evaluation loan collections and assigning loans to groups or grade based on the perceived danger and other related loans properties. The process of continual review and classification of loans enables monitoring the quality of the loan portfolios and to take action to counter fall in the credit quality of the portfolios. It is required for banks to use more complicated internal classification schemes than the more standardized schemes that bank managers need for reporting reasons and that are intended to make easy observing and interbank evaluation. There are many types of loans such as: Open-ended loans are loans that you can have a loan of more and more. Credit cards and lines of credit are the famous types of open-ended loans. You have a credit limit that you can buy with both of these two types of loans. In any time, you can purchase automatically your available credit will decrease. since you make expenditure, you're on hand increases permitting you to use the credit more and more. Closed-ended loans, this type of loans cannot be on loan once they’ve been repaid. while you make expenditure on closed-ended loans, the balance of the loan became downward. though, you don’t have any existing credit you can employ on closed-ended loans. As an option, if you want to lend more money, you’d have to make application for other loan. widespread types of closed-ended loans involve auto loans, mortgage loans, and student loans. There are two most important goals for prediction and description. Prediction involves using some variables in data set to predict unknown values of other variables and Description concentrates on finding patterns describing the data that can be interpreted by human. The derived knowledge must be new, not obvious, relevant and can be applied in the field where this knowledge has been obtained. It is also the process of extracting useful information from raw data. There are huge number of phases in the prediction based on Machine Learning, and this prediction problem used most of them, Data collection is the first phase, by this phase data should be collected not usually a less data set, it should be huge data set according to the requirements one should collect or create the data for the prediction. Data Preprocessing is the second phase and this contain a lot of sub-phases for the processing of the data, it includes importing libraries, Data Visualization, Data Transformation, Feature Scaling, Splitting and Label Encoding. Data Splitting, in this phase the data is to be split into two as train\_data and test\_data for the training of the model. Then the Fourth phase is Model Training, Supervised learning allows for processing data with target attributes or labeled data. These attributes are mapped in historical data before the training begins and the last phase is Model evaluation and Testing and it is to develop the simplest model able to formulate a target value fast and well enough. A data scientist can achieve this goal through model tuning. That’s the optimization of model parameters to achieve an algorithm’s best performance. Machine Learning techniques aid to distinguish between borrowers who pay back loans at the appointed time from those who don't. It also helps to expect when the borrower is at default, whether providing loan to a particular customer will result in bad loans. All processes related to banking sector could be analyzed using Machine Learning techniques to detect the customer’s behaviour. It also helps to analyse whether the customer will make prompt or delay payment if the credit cards are sold to them.

**1.1 OVERVIEW**

Nowadays, there are many risks related to bank loans, for the bank and for those who get the loans. The analysis of risk in bank loans need to understand what is the meaning of risk. In addition, the number of transactions in banking sector is rapidly growing and huge data volumes are available which represent the customers behaviour and the risks around loan are increased. Data Mining is one of the most motivating and vital area of research with the aim of extracting information from tremendous amount of accumulated data sets. Here a new model for classifying loan risk in banking sector by using Machine Learning concepts. The model has been built using data form banking sector to predict the status of loans. Six algorithms have been used to build the proposed model: Random Forest, Logistic Regression, Decision Tree, KNN, Naïve Bayes. By using this algorithm, a Flask model has been implemented and tested. The results have been discussed and a full comparison between algorithms was conducted. Logistic regression was selected as best algorithm based on accuracy.

**1.2 PURPOSE**

Our aim from the project is to make use of pandas, matplotlib, & seaborn libraries from python to extract the libraries for machine learning for the loan prediction. Secondly, to learn how to build the model using machine learning algorithm. And in the end, to predict whether the loan applicant can repay the loan or not using voting ensemble techniques of combining the predictions from multiple machine learning algorithms and withdrawing the conclusions.

**2. LITERATURE SURVEY**

Data mining is the process of analysing data from different perspectives and extracting useful knowledge from it. It is the core of knowledge discovery process. Different data mining techniques include classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression etc. Classification is the most commonly applied data mining technique, which employs a set of preclassified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to classification technique. This approach frequently employs Decision tree based classification Algorithm. In classification, a training set is used to build the model as the classifier which can classify the data items into its appropriate classes. A test set is used to validate the model.

**2.1 EXISTING PROBLEM**

The previous models have high time complexity and space complexity whereas this model is constrained with the lot of advantages and with a higher accuracy than any other model already proposed. In this model we used Machine learning algorithm named logistic regression which give an accuracy more than 80% of the previously predicted problem and there is an user friendly user interface to check loan score for the people who are about get the loan, and lot of the previous models haven't included the UI (User interface) which is so friendly and convenient for the users.

**2.2 PROPOSED SOLUTION**

**MACHINE LEARNING (Logistic Regression):**

Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature. It is most commonly used when the data in question has binary output, so when it belongs to one class or another. So we have used this model to build our application and we have created an UI using the Flask for the loan status prediction, this UI will allow the users to predict the loan status very easily and the User interface is user friendly not at least one complication in using the interface, and it can be used just by entering some necessary details into the UI in real time it'll give the predicted value like if the customer is beneficial to take a loan and how often does he pays the loan interest amount to the bank. Basically this model will give the predicted value when a customer with details will pay the loan back to bank, by just taking some necessary details of the customer in real time, and those details will be collected by bank employee within minutes.

**3. THEORETICAL ANALYSIS**

* Logistic regression is a mathematical modeling approach used in describing the relationship of several independent variables to a dichotomous dependent variable or a limited dependent variable. The logit function is employed because the dependent variable„default‟ is dichotomous, whereas the proposed covariates were mixture of continuous and categorical random variables. Thus the model was chosen over others due to the data structure and purpose. Also the independent variables need not be interval, nor normally distributed, nor linearly related, nor equal variance within each group. The logit model is a derivative of the odds function. The odd of a function is the ratio of the probability of success to that of failure. Thus

𝑂𝑑𝑑𝑠(𝑌 = 1) =𝑃(𝑌 = 1/ 𝑋 = 𝑥)/𝑃(𝑌 = 0/ 𝑋 = 𝑥)

Where 𝑂𝑑𝑑𝑠(𝑌 = 1)is the odds of defaul t; 𝑃(𝑌 = 1) is the probability that default occurs given a set of explanatory variables and 𝑃(𝑌 = 0) is the probability of non-default given set of explanatory variables. If the odds of default is greater than one it means there is a higher probability of default compared to that of non-default. A value less than one indicate a higher probability of non-default than that of default. Given the binary response variable (default or non-default), the probability distribution of the number of defaults in a given loan portfolio size, for given values of explanatory variables is binomial. Thus the probability that the number of default of a given portfolio size **n** is exactly equal to size *x* is given by

𝑃(𝑋 = 𝑥) =𝑛!𝑥! \* 𝑝𝑥 𝑞𝑛−𝑥 /(𝑛 − 𝑥)!

Where 𝑞 = 𝑃 𝑦(0) =; probability of non-default). This means that given a portfolio size n and probability of default P(Y=1) from a financial institution, one can use the theorem to perform risk control analysis.

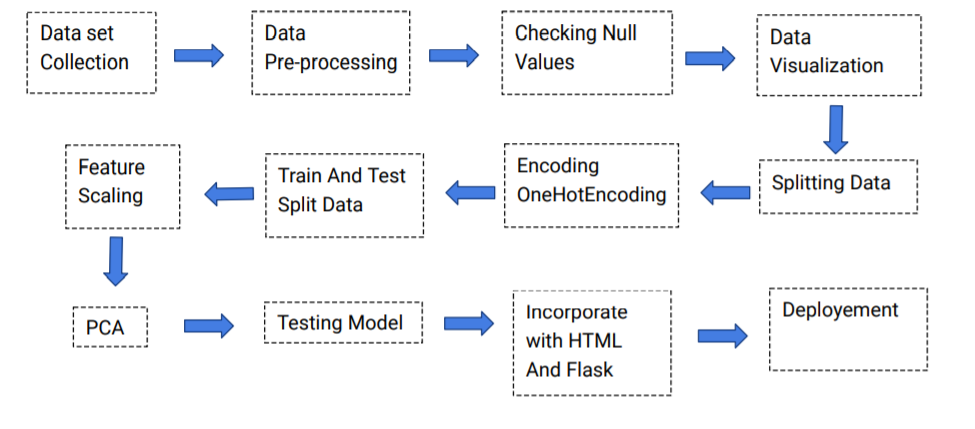
* While selecting the algorithm that gives an accurate prediction we gone through lot of algorithms which gives the results abruptly accurate and from them we selected only one algorithm for the prediction problem that is Logistic regression and it is based on binomial probability theory. The algorithm is selected based on its accuracy. Accuracy is defined as the ratio of the number of samples correctly classified by the classifier to the total number of samples for a given test data set. The formula is as follows

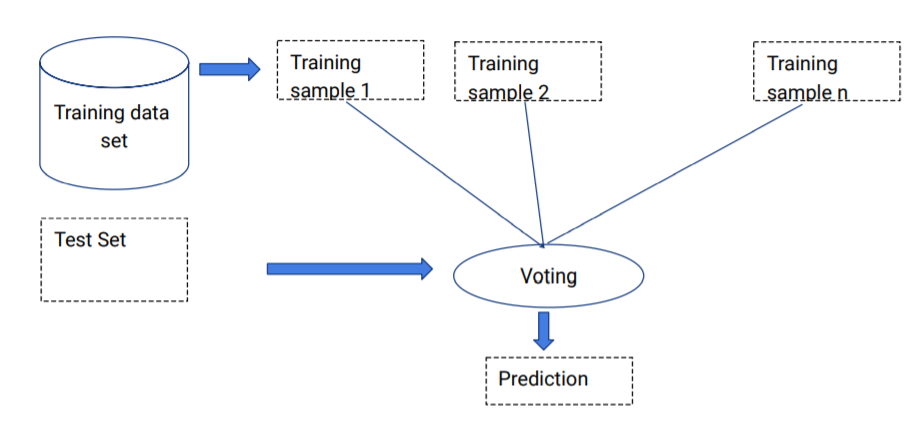
Accuracy=TP+TN/TP+TN+FT+FN

At first we got like lot of worst accuracies because we tried lot of algorithms for the best accurate algorithm , finally after all of that we tried the best suitable algorithm which gives the prediction accurately is Naive Bayes Classifier. And developed it to use as a real time prediction probelm for the loan status prediction. In statistics, a receiver operating characteristic (ROC), is a two dimensional graphical plot that illustrates the performance of a binary classifier system. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. ROC curve can intuitively represent the performance of classifier.

FPR=FP/FP+TN TPR=TP/TP+FP

**3.1 BLOCK DIAGRAM:**





**3.2 SOFTWARE DESIGNING:**

● Jupyter Notebook Environment

● Spyder Ide

● Machine Learning Algorithms

● Python (pandas, numpy, matplotlib, seaborn, sklearn)

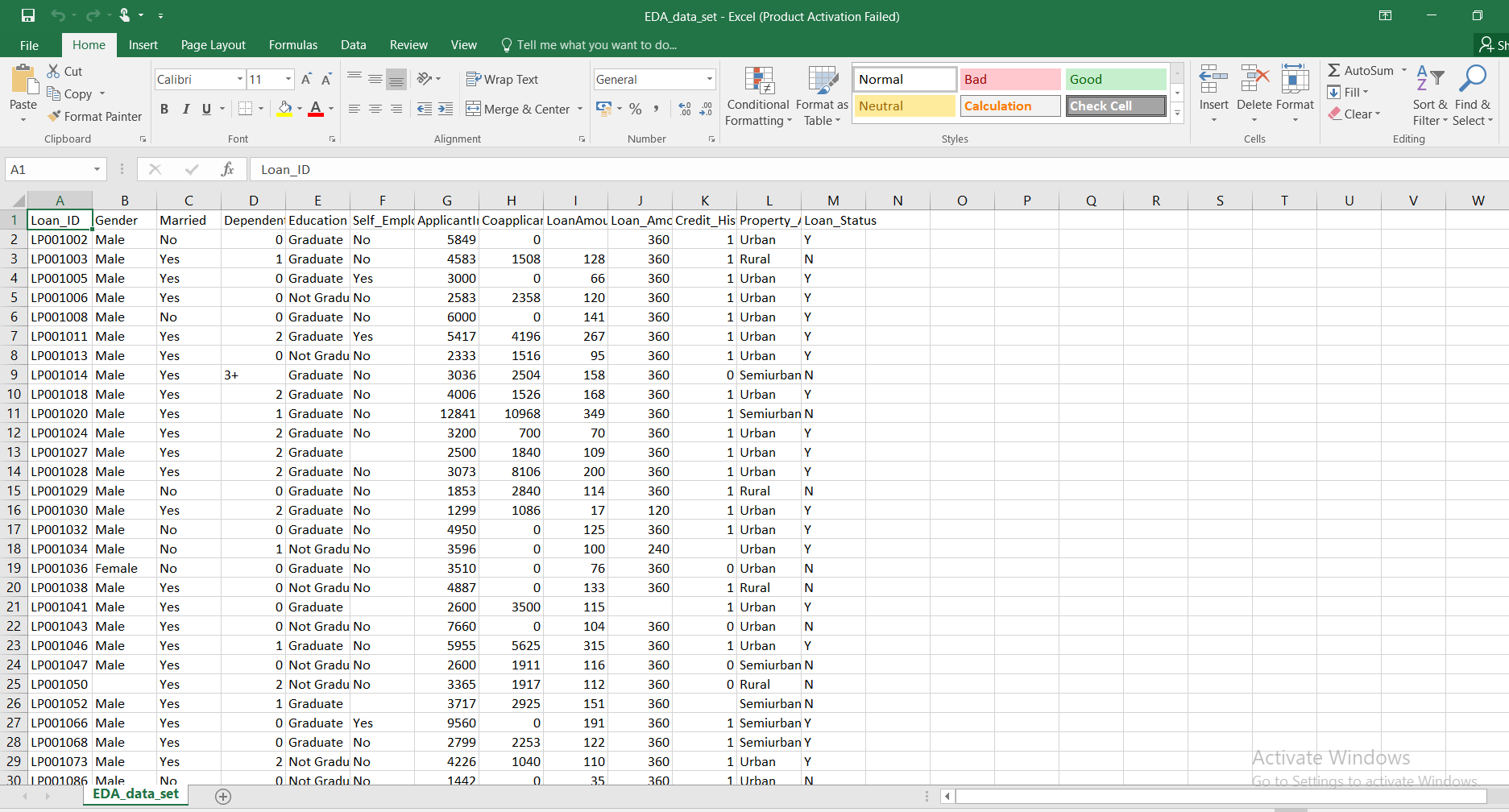
● HTML

● Flask

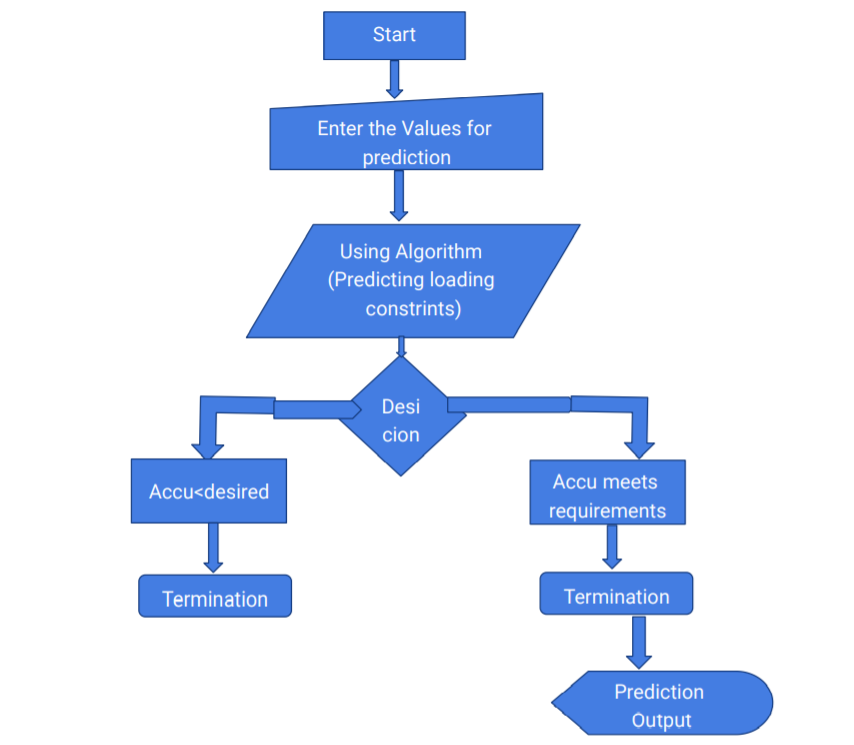
We developed this loan status prediction by using the Python language which is an interpreted and high level programming language and using the Machine Learning algorithms. For coding we used the Jupyter Notebook environment of the Anaconda distributions and the Spyder, it is an integrated scientific programming in the python language. For creating an user interface for the prediction we used the Flask. It is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions, and a scripting language to create a webpage is HTML by creating the templates to use in the functions of the Flask and HTML.

**4. EXPERIMENTAL INVESTIGATION**

It contains more than 615 original loan data of users with 16 attributes. After that, the missing values are filled in by means of mode interpolation, and the duplicate or meaningless attributeare deleted, finally we have retained to 13 attributes. Those attributes were shown below in the screenshot of the data set we used.

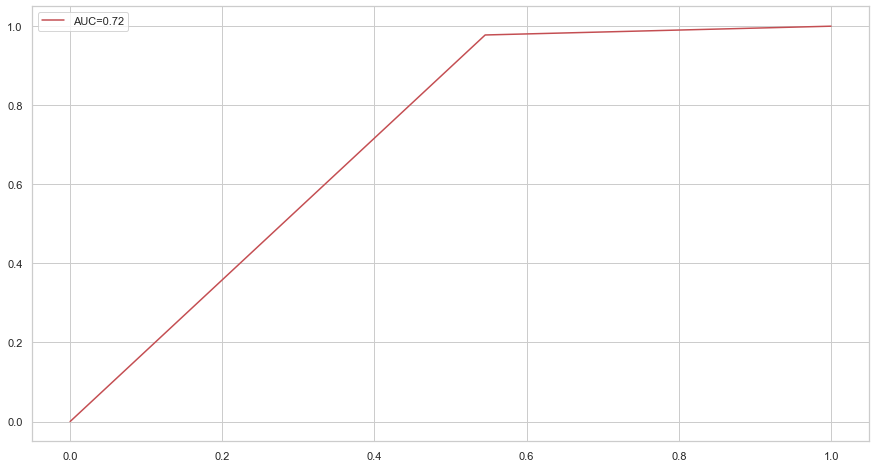


**5.FLOWCHART:**

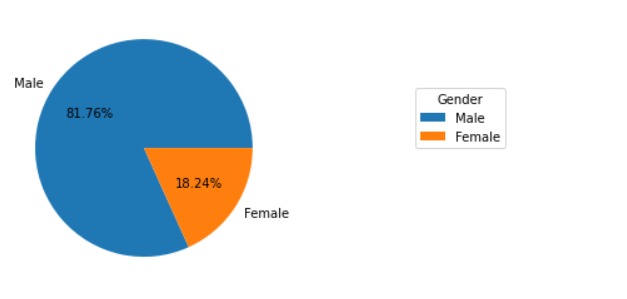
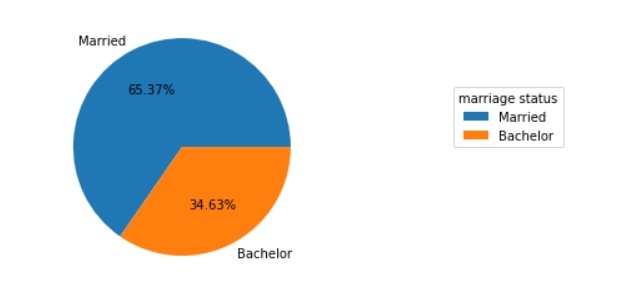


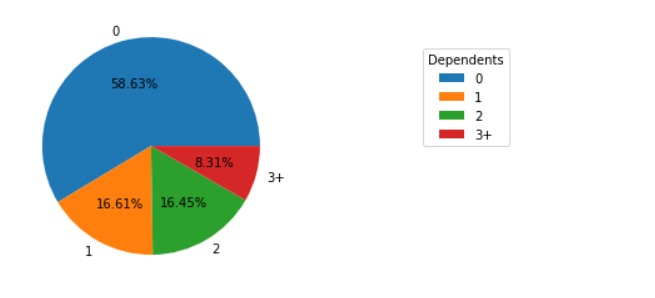
**6.RESULT:**

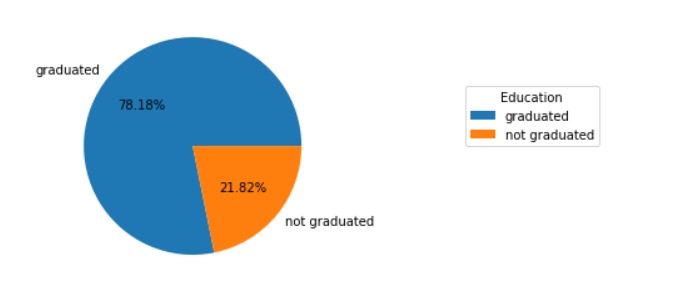
In this paper, the Logistic regression algorithm is used to predict its performance, and compared with all other machine learning methods namely the decision tree, the Naïve Bayes, KNN, Random Forest. The obtained results are displayed in Table below. The results show that, the performance of logistic regression have comparable performance than that of Naive Bayes, KNN, random forest and decision tree, but the logistic regression still performs the best, with an accuracy of 83%, higher than the other models. The ROC curve of the prediction model based on logistic regression are all above 0.50, indicating that the model has strong ability of generalization. The ROC curve of the logistic regression is shown below, the curve have 0.72 as AUC but the accuracy of this model is best out of the all algorithms at 83%. The point on the ROC curve has least classification errors, and the total number of false positive examples and false negative examples is the lowest. The given pie shows who are the people likely to take a loan and gets approved and pays the loan amount with interest to the bank from the data set we've took for the prediction.

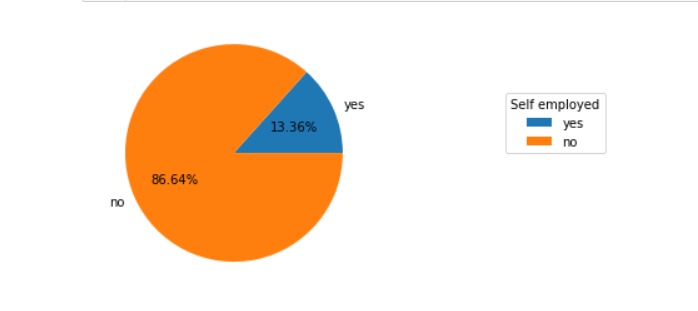
****

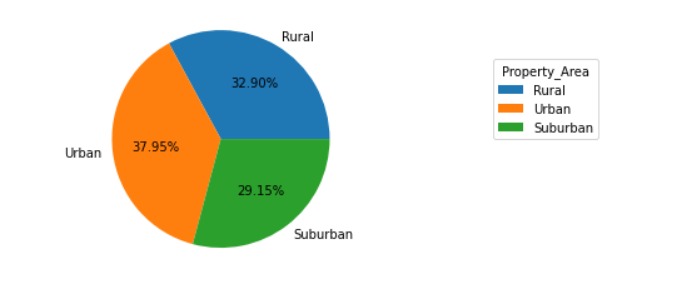


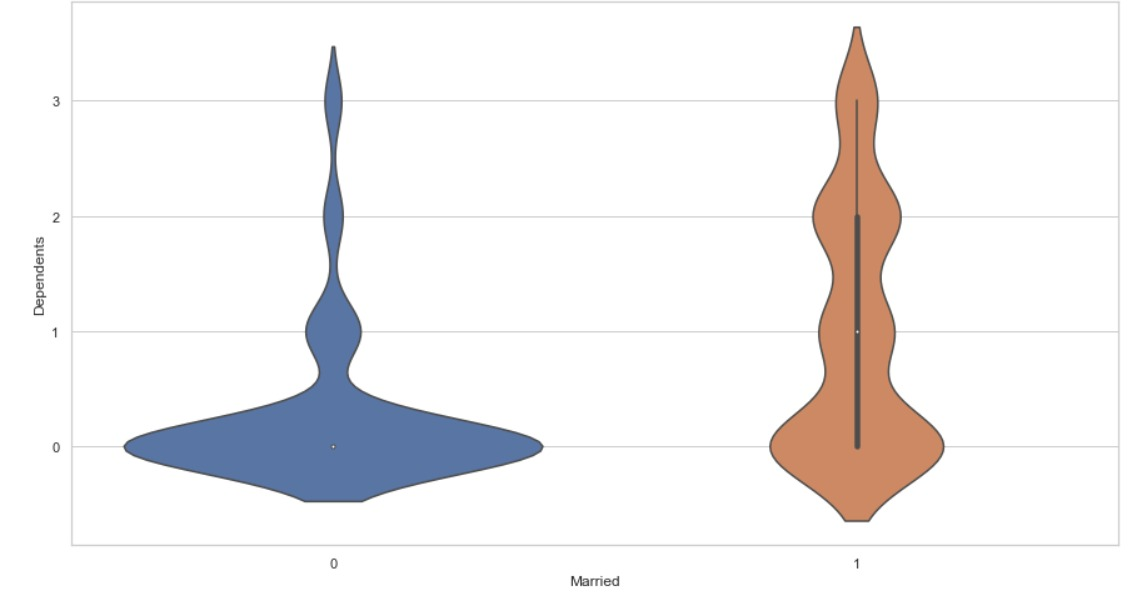
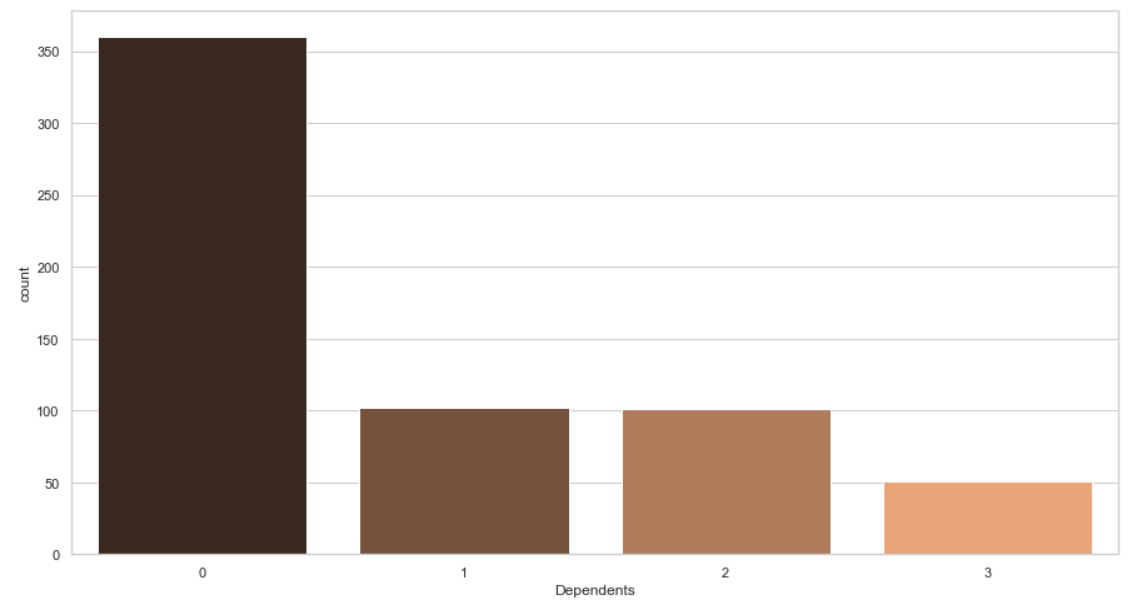
 

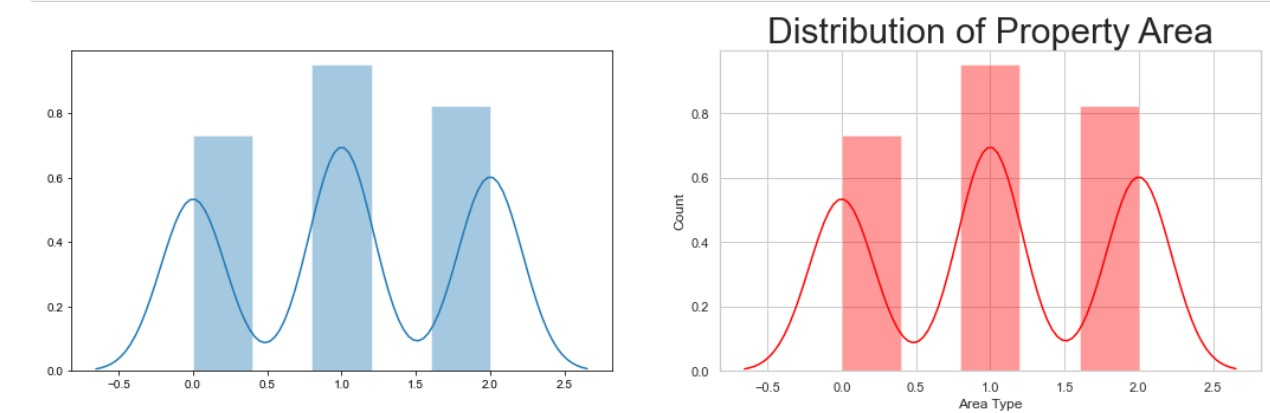










|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No.** | **Algorithm used** | **Accuracy** | **ROC Curve** |
| 1. | Logistic Regression | 0.837 | 0.716 |
| 2. | KNN | 0.829 | 0.75 |
| 3. | Decision Tree Classifier | 0.691 | 0.63 |
| 4. | Naïve Bayes | 0.829 | 0.71 |

**7**. **ADVANTAGES** **AND DISADVANTAGES**

Advantages:

● Easy and simple User Interface for the bank people who is going to evaluate the customer loan status.

● Logistic Regression give the accurate result of the prediction up to

83% which is the algorithm we used for prediction.

● It is composed using the HTML and Python for the web usage in real time.

● It can work in real time and predict as soon as the necessary details for prediction are given to the model.

Disadvantages:

● It could not work anywhere like a web-application, if one is using other should be quiet.

● Needs more than a single value for the prediction.

**8. APPLICATIONS**

● It is widely used for managing risks in the banking industry. Bank executives need to know the credibility of customers they are dealing with in real time.

● To have an idea of customer relationship cycle such as customer acquisition, increasing value of the customer and customer retention.

● It is one of the most widely used areas of data mining in the banking industry. The consumer behaviour with reference to product, price and distribution channel can be analysed by the marketing department.

● Due to tremendous growth in data the banking industry deals with, analysis and transformation of the data into useful knowledge has become a task beyond human ability.

● So we use Machine Learning Algorithms to analyse the data and propose what banks and loan lending companies need to achieve their needs.

**9. CONCLUSION**

In this paper, the logistic algorithm is adopted to build a UI model for predicting loan status and the results are compared with other algorithms of Naïve Bayes, KNN, random forest and decision tree. The experiment shows that the logistic regression algorithm performs outstanding than the other algorithms in the prediction of loan default and has strong ability of generalization. There is no definitive guide of which algorithms to use given any situation. What may work on some data sets may not necessarily work on others. Therefore, always evaluate methods using cross validation to get a reliable estimate.

**10. FUTURE SCOPE**

In future the logistic regression algorithm can be applied on other data sets available for loan approvals to further investigate its accuracy. A rigorous analysis of other machine learning algorithms other than these can also be done in future to investigate the power of machine learning algorithms for loan status prediction. In further study, we will try 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 and a great UI support system making it complete web application model.

**11.BIBLIOGRAPHY**

●Kaggle platform - <https://www.kaggle.com/psvishnu/loan-prediction-part-1/data>

●Docsity - https://www.docsity.com/en/eda-loan-status-prediction/5584982/

●GitHub - https://github.com/SmartPracticeschool/llSPS-INT-2394-EDA-Loan-Status-

**HTML FILE**:

  <!DOCTYPE html>

<html >

<!--From https://codepen.io/frytyler/pen/EGdtg-->

<head>

<meta charset="UTF-8">

<title>Loan Status Prediction</title>

<link href='https://fonts.googleapis.com/css?family=Pacifico' rel='stylesheet' type='text/css'>

<link href='https://fonts.googleapis.com/css?family=Arimo' rel='stylesheet' type='text/css'>

<link href='https://fonts.googleapis.com/css?family=Hind:300' rel='stylesheet' type='text/css'>

<linkhref='https://fonts.googleapis.com/css?family=Open+Sans+Condensed:300' rel='stylesheet' type='text/css'>

<link rel="stylesheet" href="{{ url\_for('static', filename='css/style22.css') }}">

<style>

.login{

top: 20%;

}

</style>

</head>

<body>

<div class="login">

<h1>Loan Status Prediction</h1>

<!-- Main Input For Receiving Query to our ML -->

<form action="{{ url\_for('y\_predict')}}"method="post">

</select>

<input type="number" name="Loan\_ID" placeholder="Loan\_ID" required="required" />

<select name="Gender" required="required">

<option value="">Select Gender </option>

<option value="Male">Male</option>

<option value="Female">Female</option>

</select>

<select name="Married" required="required">

<option value="">Select Married </option>

<option value="Yes">Yes</option>

<option value="No">No</option>

</select>

<select name="Dependents" required="required">

<option value="">Select Dependents </option>

<option value="0">0</option>

<option value="1">1</option>

<option value="2">2</option>

<option value="3">3</option>

</select>

<select name="Education" required="required">

<option value="">Select Education </option>

<option value="Graduate">Graduate</option>

<option value="Not Graduate">Not Graduate</option>

</select>

<select name="Self\_Employed" required="required">

<option value="">Select Self\_Employed </option>

<option value="Yes">Yes</option>

<option value="No">No</option>

</select>

<input type="number" name="ApplicantIncome" placeholder="ApplicantIncome" required="required" />

<input type="number" name="CoapplicantIncome" placeholder="CoapplicantIncome" required="required" />

<input type="number" name="LoanAmount" placeholder="LoanAmount" required="required" />

<input type="number" name="Loan\_Amount\_Term" placeholder="Loan\_Amount\_Term" required="required" />

<input type="number" name="Credit\_History" placeholder="Credit\_History" required="required" />

<select name="Property\_Area" required="required">

<option value="">Select Property\_Area </option>

<option value="0">Rural</option>

<option value="1">Semiurban</option>

<option value="2">Urban</option>

</select>

<button type="submit" class="btn btn-primary btn-block btn-large">Predict</button>

</form>

<br>

<br>

{{ prediction\_text }}

</div>

</body>

</html>

**App.py file**:

import numpy as np

from flask import Flask, request, jsonify, render\_template

from joblib import load

app = Flask(\_\_name\_\_)

model= load('logisticregressor.save')

trans1=load('transform1')

trans2=load('transform2')

trans3=load('transform3')

trans4=load('transform4')

scaler = load('scaler')

@app.route('/')

def home():

return render\_template('index.html')

@app.route('/y\_predict',methods=['POST'])

def y\_predict():

'''

For rendering results on HTML GUI

'''

x\_test = [[x for x in request.form.values()]]

print(x\_test)

test=trans1.transform(x\_test)

test=test[:,1:]

test=trans2.transform(test)

test=test[:,1:]

test=trans3.transform(test)

test=test[:,1:]

test=trans4.transform(test)

test=test[:,1:]

print(test)

#test = scaler.transform(test)

prediction = model.predict(scaler.transform(test))

print(prediction)

if prediction[0] == 1:

output = 'Congrats,you are eligible for loan'

else:

output = 'sorry,you are not eligible for loan'

return render\_template('index.html', prediction\_text='Loan\_Status:{}'.format(output))

'''@app.route('/predict\_api',methods=['POST'])

def predict\_api():

#For direct API calls trought request

data = request.get\_json(force=True)

prediction = model.y\_predict([np.array(list(data.values()))])

output = prediction[0]

return jsonify(output)'''

if \_\_name\_\_ == "\_\_main\_\_":

app.run(debug=True)