**PROJECT REPORT**

**Loan Prediction**

Submitted by

**Kadithampalli Santhosh Registration Number : 11917070**

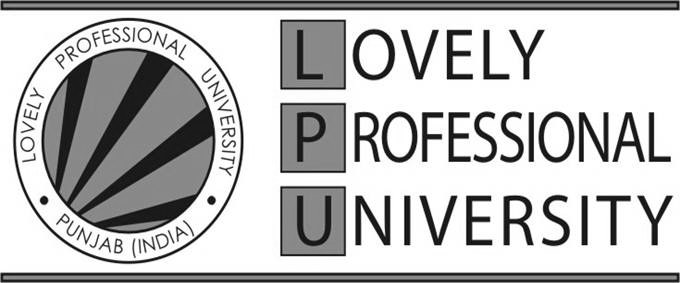
**Yeduru Lakshmi Naveen Reddy Registration Number :11906402**

**Course Code INT246**

Under the Guidance of

Sagar Pande (Department of computer science engineering)

# School of Computer Science and Engineering



1

**CONTENTS:**

Declaration ……………………….. 3

Acknowledgement …………………….…. 4

Literature Survey ……………………….. 5

Abstract ………………………... 5

Introduction ……………………….. 6

Motivation ………………………. 10

Problem Statement ……………………… 11

Problem Methodology ……………………… 11

Implementation ……………………… 12

Conclusion ……………………… 20

Github link ……………………… 21

References ……………………… 21

2

**DECLARATION**

We hereby declare that the project work entitled Loan Prediction is an authentic record of our own work carried out as requirements of Project for the award of B.Tech degree in Computer science engineering from Lovely Professional University, Phagwara, under the guidance of sagar pande sir, during August to November 2020. All the information furnished in this project report is based on our own intensive work and is genuine.

K.santhosh(11917070)

Y.Naveen(11906402)

**School of Computer Science and Engineering,**

Lovely Professional University,

Phagwara, Punjab

3

Acknowledgement:

I am really really grateful to my English Teacher (Teacher’s Name) for advising me and introducing the project to me in a easy to understand way which has helped me complete my project easily and effectively on time.

I am dearly obliged to (Name) for giving me an opportunity to work on this project which has provided valuable information about (Topic of your Project).

Thank you.

4

Literature Survey:

" Loan Approval Prediction based on Machine Learning

Approach" by Santhosh And Naveen,in the year 2021.The

main objective of this paper is to predict whether assigning

the loan to particular person or not. This paper is divided into

four sections (i)Data Collection (ii) Comparison of machine

learning models on collected data (iii) Training of system on

most promising model (iv) Testing

ABSTRACT :

- In our banking system, banks have many products to sell but main source of

income of any banks is on its credit line. So they can earn from interest of those

loans which they credits. A bank’s profit or a loss depends to a large extent on

loans i.e. whether the customers are paying back the loan or defaulting. By

predicting the loan defaulters, the bank can reduce its Non-performing Assets.

This makes the study of this phenomenon very important. Previous research in

this era has shown that there are so many methods to study the problem of

controlling loan default. But as the right predictions are very important for the

maximization of profits, it is essential to study the nature of the different methods

5

and their comparison. The aim of this project was to compare the various Loan

Prediction Models and show which is the best one with the least amount of error

and could be used by banks in real world to predict if the loan should be approved

or not taking the risk factor in mind. A very important approach in predictive

analytics is used to study the problem of predicting loan defaulters

1. Collection of Data, (ii) Data Cleaning and (iii) Performance Evaluation.

INTRODUCTION :

A Prediction Model uses data mining, statistics and probability to

forecast an outcome. Every model has some variables known as predictors

that are likely to influence future results. The data that was collected from

various resources then a statistical model is made. It can use a simple linear

equation or a sophisticated neural network mapped using a complex

software. As more data becomes available the model becomes more refined

and the error decreases meaning then it’ll be able to predict with the least

risk and consuming as less time as it can. The Prediction Model helps the

banks by minimizing the risk associated with the loan approval system and

helps the applicant by decreasing the time taken in the process.

The main objective of the Project is to compare the Loan Prediction

Models made implemented using various algorithms and choose the best one

out of them that can shorten the loan approval time and decrease the risk

6

associated with it.

The aim of this Paper is to provide quick, immediate and easy way to choose the

deserving applicants. Dream housing Finance Company deals in all loans. They

have presence across all urban, semi urban and rural areas. Customer first apply

for loan after that company or bank validates the customer eligibility for loan.

Company or bank wants to automate the loan eligibility process (real time)

based on customer details provided while filling application form. These details

are Gender, Marital Status, Education, Number of Dependents, Income, Loan

Amount, Credit History and other. This project has taken the data of previous

customers of various banks to whom on a set of parameters loan were approved.

So the machine learning model is trained on that record to get accurate results.

Our main objective of this project is to predict the safety of loan.

We are using SVM, logistic Regression Model for this.

SVM : Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is

7

termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



Source: <https://static.javatpoint.com/tutorial/machine-learning/images/support-vector-machine-algorithm.png>

SVM can be of two types:

Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

8

Hyperplane and Support Vectors in the SVM algorithm:

Hyperplane: There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

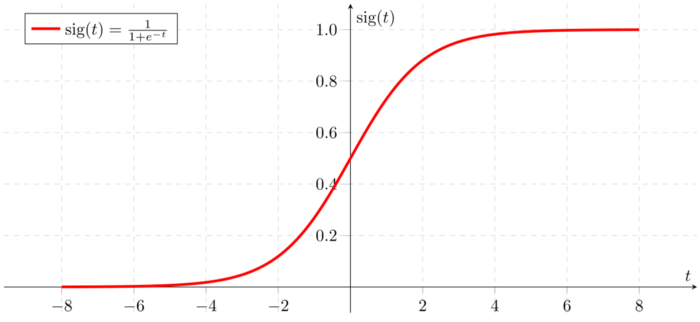
Support Vectors:

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

Logistic Regression:

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of cyber security are classification problems, such as attack detection, logistic regression is a useful analytic technique.

9



Source: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRuEf6uwdmpi8XPrkrWaAtuTwPHPkLLmV7pJQ&usqp=CAU

Types of Logistic Regression

1. Binary Logistic Regression

The categorical response has only two 2 possible outcomes. Example: Spam or Not

2. Multinomial Logistic Regression

Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)

3. Ordinal Logistic Regression

Three or more categories with ordering. Example: Movie rating from 1 to 5

MOTIVATION:

Loan approval is a very important process for banking organizations. The

system approved or reject the loan applications. Recovery of loans is a major

contributing parameter in the financial statements of a bank. It is very difficult

10

to predict the possibility of loan approval by the customer. Using Machine

learning we predict the loan approval.

PROBLEM STATEMENT:

Banks, Housing Finance Companies and some NBFC deal in various types of

loans like housing loan, personal loan, business loan etc in all over the part of

countries. These companies have existence in Rural, Semi Urban and Urban

areas. After applying loan by customer these companies validates the eligibility

of customers to get the loan or not. This project provides a solution to automate

this process by employing machine learning algorithm. So the customer will fill

an online loan application form. This form consist details like Sex, Marital

Status, Qualification, Details of Dependents, Annual Income, Amount of Loan,

Credit History of Applicant and others.

PROPOSED METHODOLOGY:

The following shows the pseudo code for the proposed

loan prediction method

1. Load the data

2. Determine the training and testing data

11

3. Data cleaning and pre-processing.

a) Fill the missing values with mean values regarding

numerical values.

b) Fill the missing values with mode values regarding

categorical variables.c) Outlier treatment.

4. Apply the modelling for prediction

a) Removing the load identifier

b) Create the target variable (based on the requirement).

In this approach, target variable is loan-status

c) Create a dummy variable for categorical variable (if

required) and split the training and testing data for

validation.

d) Apply the model: SVM method, logistic regression, Linear regression

IMPLEMENTATION:

Loan Dataset is very useful in our system for prediction of more accurate result.

Using the loan Dataset the system will automatically predict which costumer’s

loan it should approve and which to reject. System will accept loan application

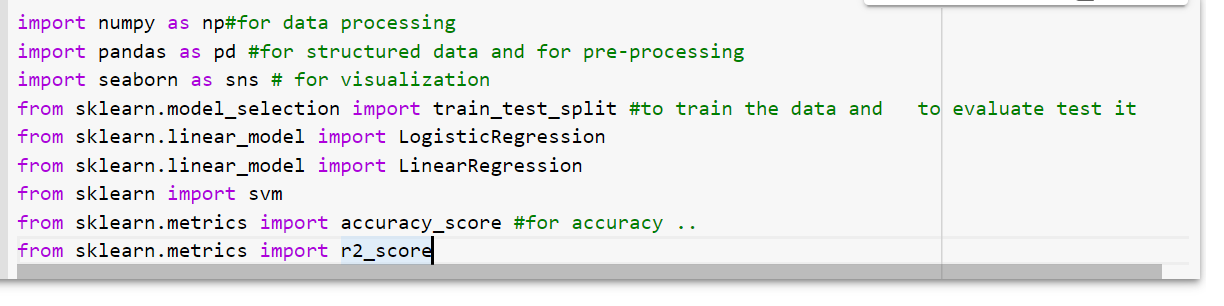
form as an input. Justified format of application form should be given as an

input to get processed.

Importing Dependencies

We imported whatever libraries are required for this project

12



1.we used numpy as np for data processing and also pandas

2.seaborn is used for data visualization

3.And importing train\_test\_split function to train the data and evaluate the test

4.The model we are training is Support Vector Machine

5.To calculate accuracy score we imported accuracy\_score

6.Logistic regression used to understand the relationship between the

dependent variable and one or more independent variables

7. Linear regression analysis is used to predict the value of a variable based on

the value of another variable.

8. R-squared will give you an estimate of the relationship between movements

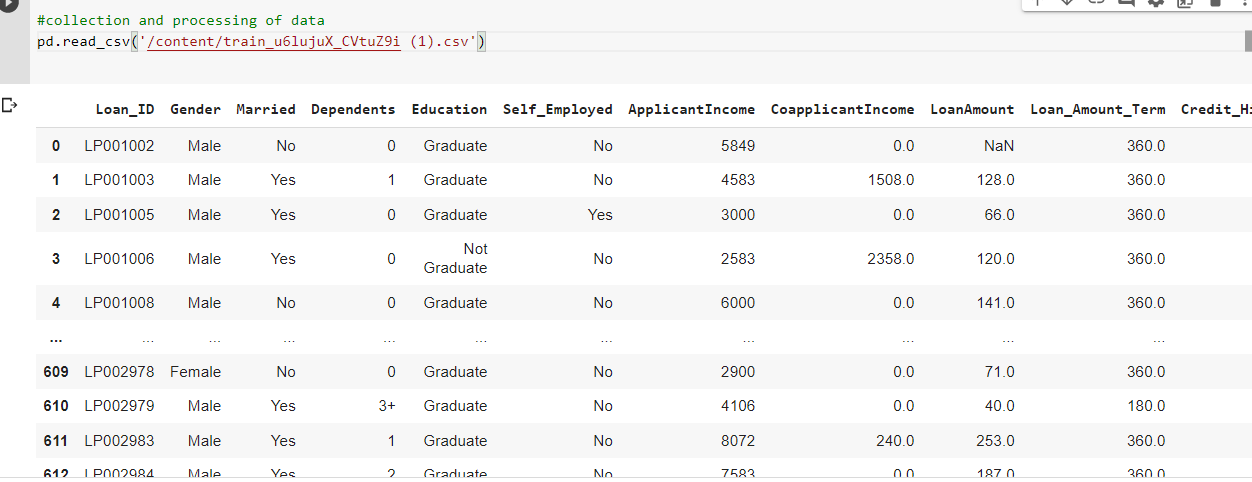
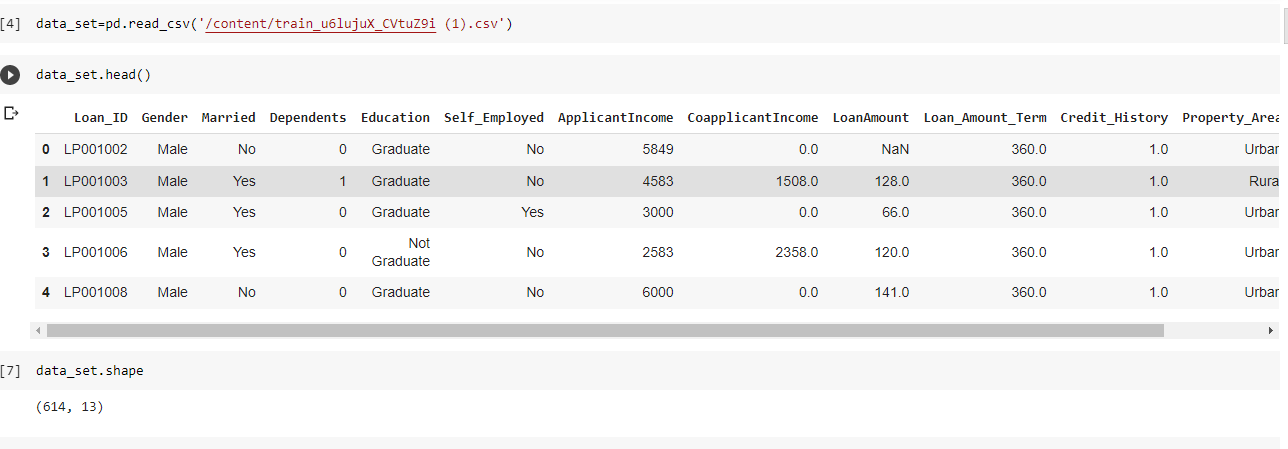
of a dependent variable based on an independent variable's movements.

13

Data collection and Processing

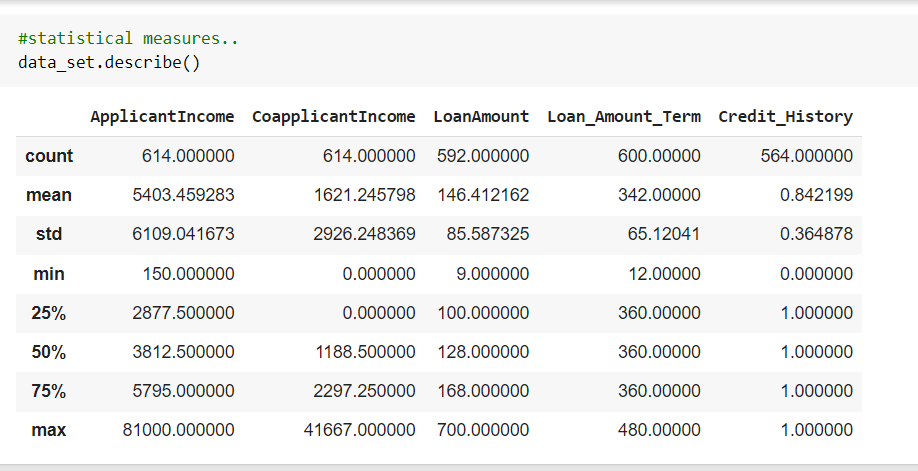
The training set was imported in csv format and a simple function is

applied to check whether it’s working or not.



For statistical measures we used describe ,

14



Data cleaning

In Data cleaning the system detect and correct corrupt or inaccurate records

from database and refers to identifying incomplete, incorrect, inaccurate or

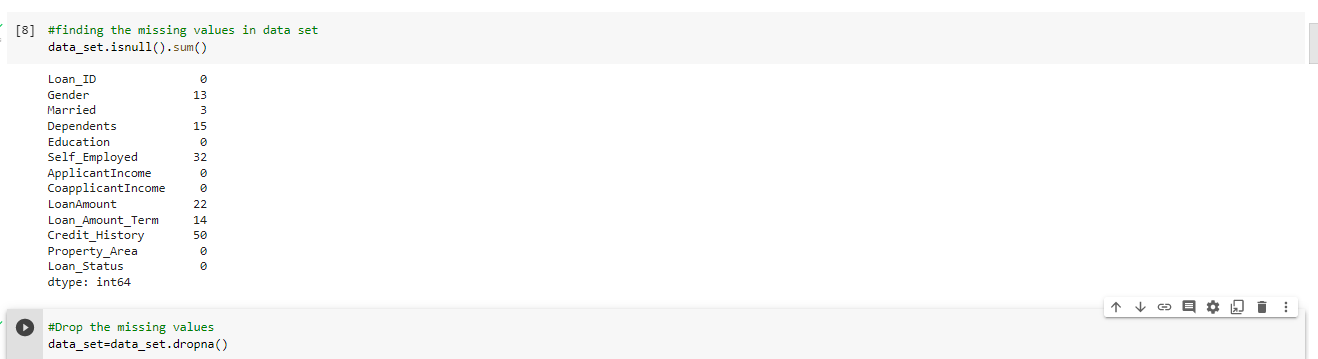
irrelevant parts of the data and then replacing , modifying or detecting the dirty

or coarse data. In Data processing the system convert data from a given form to

a much more usable and desired form i.e. make it more meaningful and

informative.

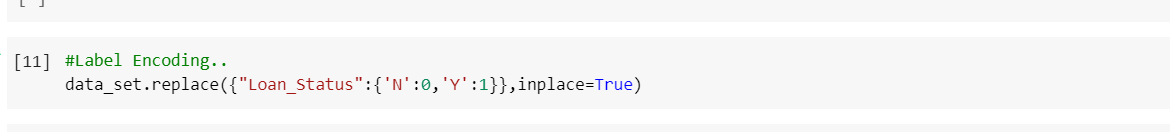
15

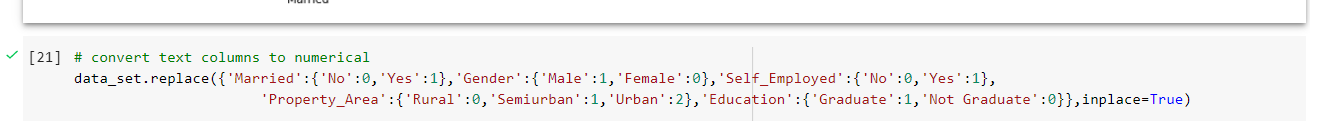


Categorical Variable analysis

Label Encoding

We know machine/model cannot understand categorical values so we have to modify them to numerical values, for that purpose we used replace function from pandas

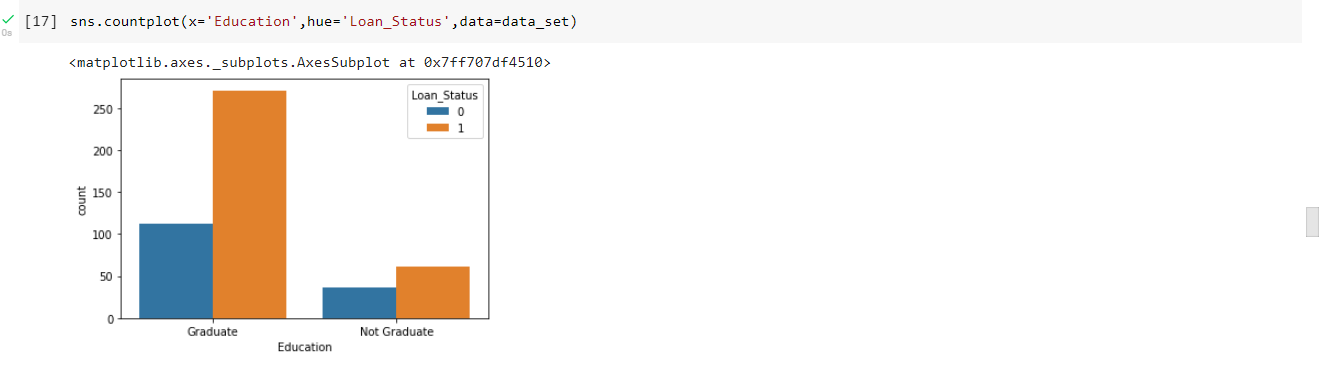




16

Data Visualization

At earlier said that we had used seaborn for visualization



Here the people who are educated they have been accepted more for the loans

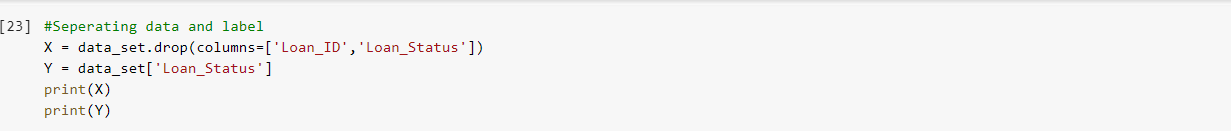
And Coming to gender male has been accepted more than female.



17

Seperating DATA and LABEL

Because here we don’t need loan\_id column ,no use of it ,so we are dropping it out of the table.



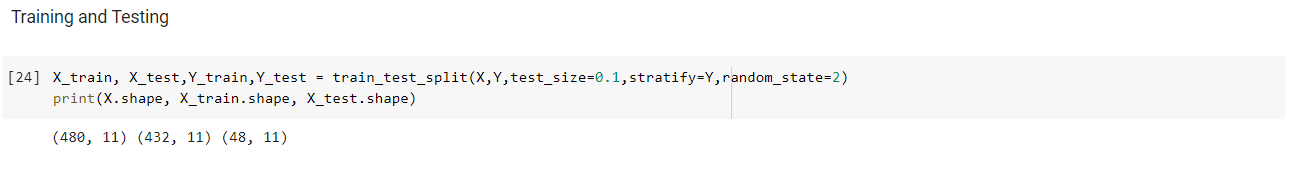
Training and Testing

Typically , Here the system separate a dataset into a training set and testing set

,most of the data use for training ,and a smaller portions of data is use for

testing. after a system has been processed by using the training set, it makes the

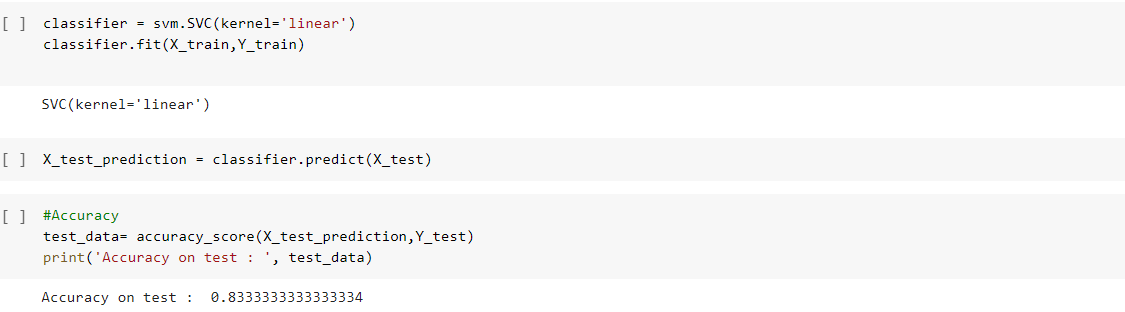
prediction against the test set.



18

Models Used

1.SVM

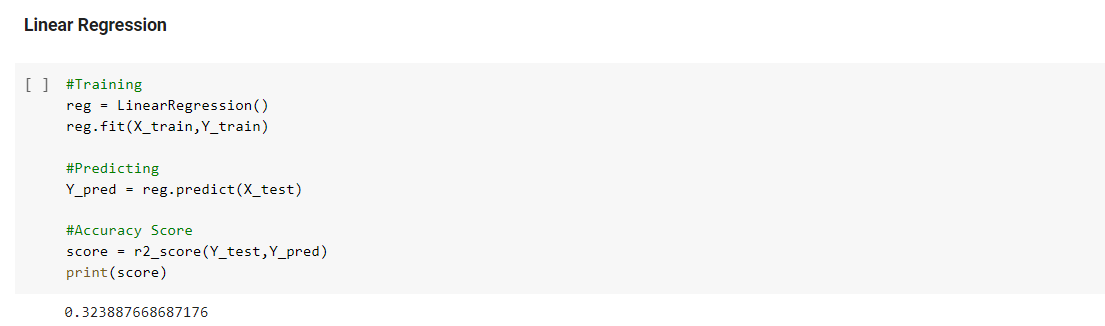


2.Logistic Regression



19

3.Linear Regression



Conclusion

The predictive models based on SVM,Logistic Regression,Linear Regression

give the accuracy as 83.3%,83.3% and 32.3% respectively ,

So here, it can be concluded with confidence that the SVM and Logistic

Regression models are extremely efficient and give a better result when

compared to other model. It works correctly and fulfills all requirements of

bankers. This system properly and accurately calculate the result. It predicts the

loan is approve or reject to loan applicant or customer very accuratly.

20

Github Link:

<https://github.com/naveen9581/Loan-Prediction>

References

<https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRuEf6uwdmpi8XPrkrWaAtuTwPHPkLLmV7pJQ&usqp=CAU>

<https://static.javatpoint.com/tutorial/machine-learning/images/support-vector-machine-algorithm.png>

21