CUSTOMER LOAN PREDICTION ANALYSIS

Mrs. V. SARALA [1],Mr. DASARI. SAI PAVAN MUKESH [2]

[1] Assistant Professor MCA, DEPT, Dantuluri Narayana Raju College, **Bhimavaram, Andhrapradesh**

EMAIL ID : vedalasarala21@gmail.com

[2]PG Student of MSC(CS), Dantuluri Narayana Raju College, **Bhimavaram, Andhrapradesh**

EMAIL ID : dspmukesh.2000@gmail.com

ABSTRACT

With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this paper we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this paper is to predict whether assigning the loan to particular person will be safe 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.

1. INTRODUCTION

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminate analysis. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in the Gaspé Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other. The use of this data set in cluster analysis however is not common, since the data set only contains two clusters with rather obvious separation. One of the clusters contains Iris setosa, while the other cluster contains both Iris virginica and Iris versicolor and is not separable without the species information Fisher used. This makes the data set a good example to explain the difference between supervised and unsupervised techniques in data mining: Fisher's linear discriminant model can only be obtained when the object species are known: class labels and clusters are not necessarily the same.

Nevertheless, all three species of Iris are separable in the projection on the nonlinear branching principal component. The data set is approximated by the closest tree with some penalty for the excessive number of nodes, bending and stretching. Then the so-called "metro map" is constructed. The data points are projected into the closest node. For each node the pie diagram of the projected points is prepared.

The area of the pie is proportional to the number of the projected points. It is clear from the diagram (left) that the absolute majority of the samples of the different Iris species belong to the different nodes. Only a small fraction of Iris-virginica is mixed with Iris-versicolor (the mixed blue-green nodes in the diagram). Therefore, the three species of Iris (Iris setosa, Iris virginica and Iris versicolor) are separable by the unsupervising procedures of nonlinear principal component analysis. To discriminate them, it is sufficient just to select the corresponding nodes on the principal tree.

2. LITERATURE SURVEY AND RELATED WORK

Random forest is ensemble learning method for both classification and replaces issues. The advantage of random decision forest is reduce over fitting and helps to improve the accuracy and runs efficiently on a large datasets and work on both continuous and categorical values and predict analysis of data with help of test data.

Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal are used data mining methodology to predict the likely default from a dataset that contains information about home loan applications, thereby helping the banks for making better decisions in the future [3].

Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal are used data mining methodology to predict the likely default from a dataset that contains information about home loan applications, thereby helping the banks for making better decisions in the future [3].

Aakanksha, Tamara Denning, Vivek Srikumar, Sneha Kumar Kesera[8] this paper is mainly used for voting classifier (combination of logistic regression, naïve bayes, SVM). They able to reduce the number of FP considerably. This work represents the group of generic passwords to reduce misclassification. Arutjothi [9] present a new credit scoring model, which depends on the hybrid feature selection model and C4.5 classifier. This is depend on hybrid system not only has a strong mathematical basis, but also has higher accuracy and more benefits.

3. EXISTING SYSTEM

Machine Learning implementation is a very complex part in terms of Data analytics. Working on the data which deals with prediction and making the code to predicit the future of out comes from the customer is challenging part.

Disadvantages of Existing System:

- Complexity in analyzing the data.
- Prediction is challenging task working in the model
- Coding is complex maintaining multiple methods.
- Libraries support was not that much familiar.

4. PROPOSED SYSTEM

Python has a is a good area for data analytical which helps us in analyzing the data with better models in data science. The libraries in python makes the predication for loan data and results with multiple terms considering all properties of the customer in terms of predicting.

Advantages:

- Libraries helps to analyse the data.
- Statistical and prediction is very easy comparing to existing technologies.

5. METHODOLOGIES

MODULES

DATASET

Thispaperutilizesthedatasetprovidedbyrevolutionanalyticsforthedetectionofthefraudulentcredit card transaction from Kaggle. Dataset has 51149 legal transactions and 3312 fraudulenttransactions. The dataset is divided as 60%, 20% and, 20% in the Train, Valid and Test set,respectively

DATAPREPROCESSING

For efficient implementation of the classification algorithm, data preprocessing is performedbefore feature selection. Under-sampling is performed to make the dataset balanced to avoidthe biasing of the classification algorithm towards the majority class. Feature Selection isimplementedon abalanced dataset.

FEATURESELECTION

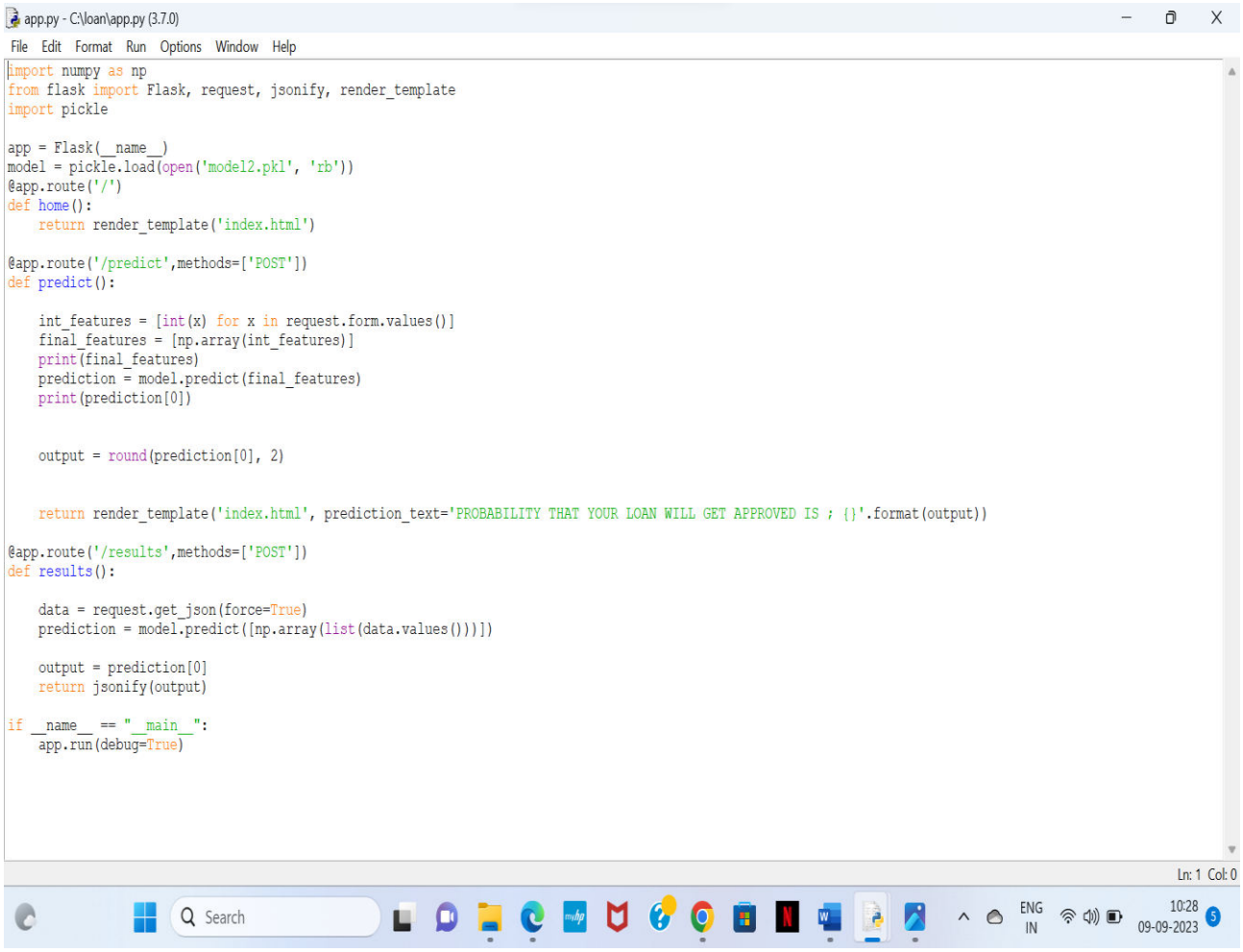
Featureselectionmethodsareusedtoremoveunnecessary,irrelevant,andredundantattributesfrom a dataset that do not contribute to the accuracy of a predictive model or which mightreduce the accuracy of the model. In this paper seven feature selection techniques namelySelect-K-best,FeatureImportance,Extratressclassifier,Person'scorrelation,MutualInformation,Stepforward selectionand Recursivefeatureeliminationareused.

FEATUREIMPORTANCE

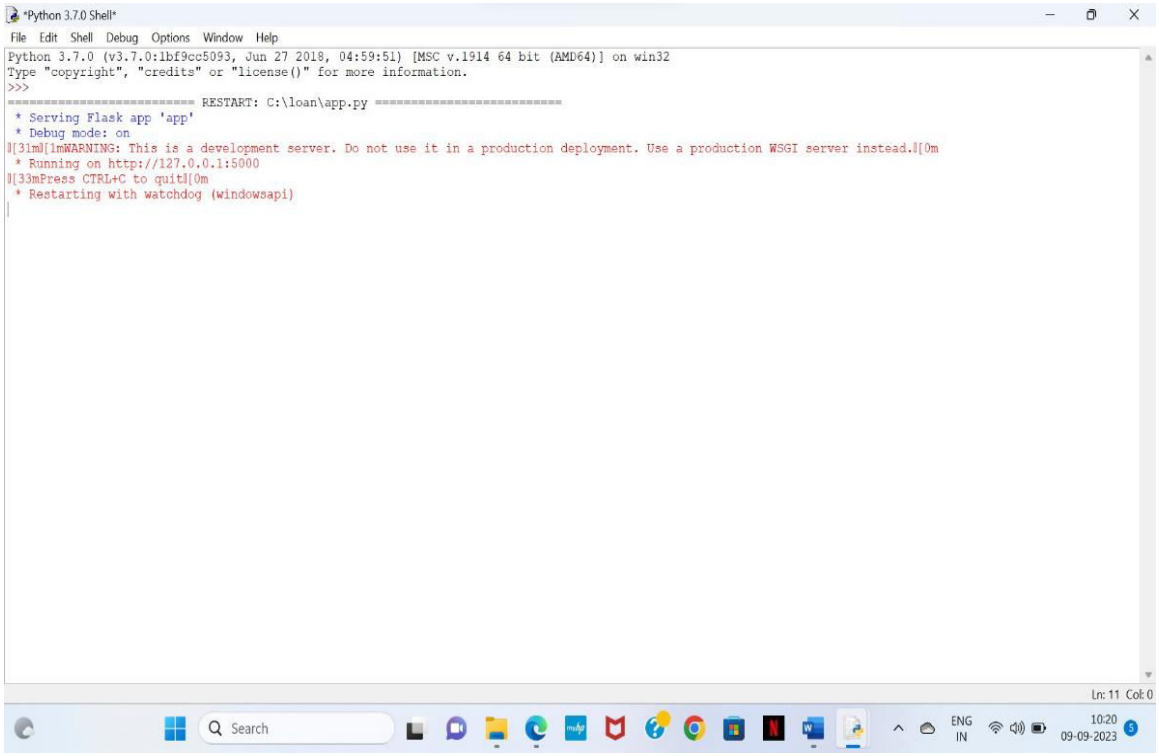
Featureimportanceisaclassoftechniquesforassigningscorestoinputfeaturestoapredictivemodelthatindicatestherelativeimportanceofeachfeatureatthetimeofmakingaprediction.Itreducesthenumberofinputfeatures.Inthispaper,featureimportanceisimplementedusinganextratreeclassifierfromthedecisiontree.ExtraTrees issimilartoRandomForest,itbuildsmultiple trees and splits nodes using random subsets of features, but unlike Random Forest,ExtraTreesampleswithoutreplacementand nodesarespliton random

6. RESULTS AND DISCUSSION SCREEN SHOTS

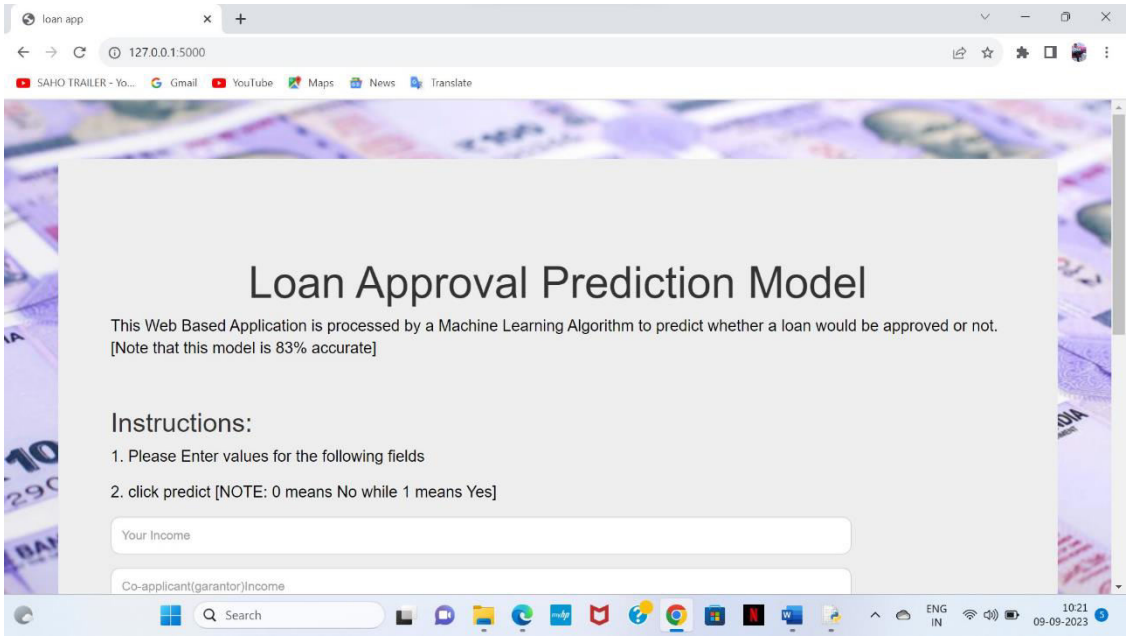
MAIN SCREEN

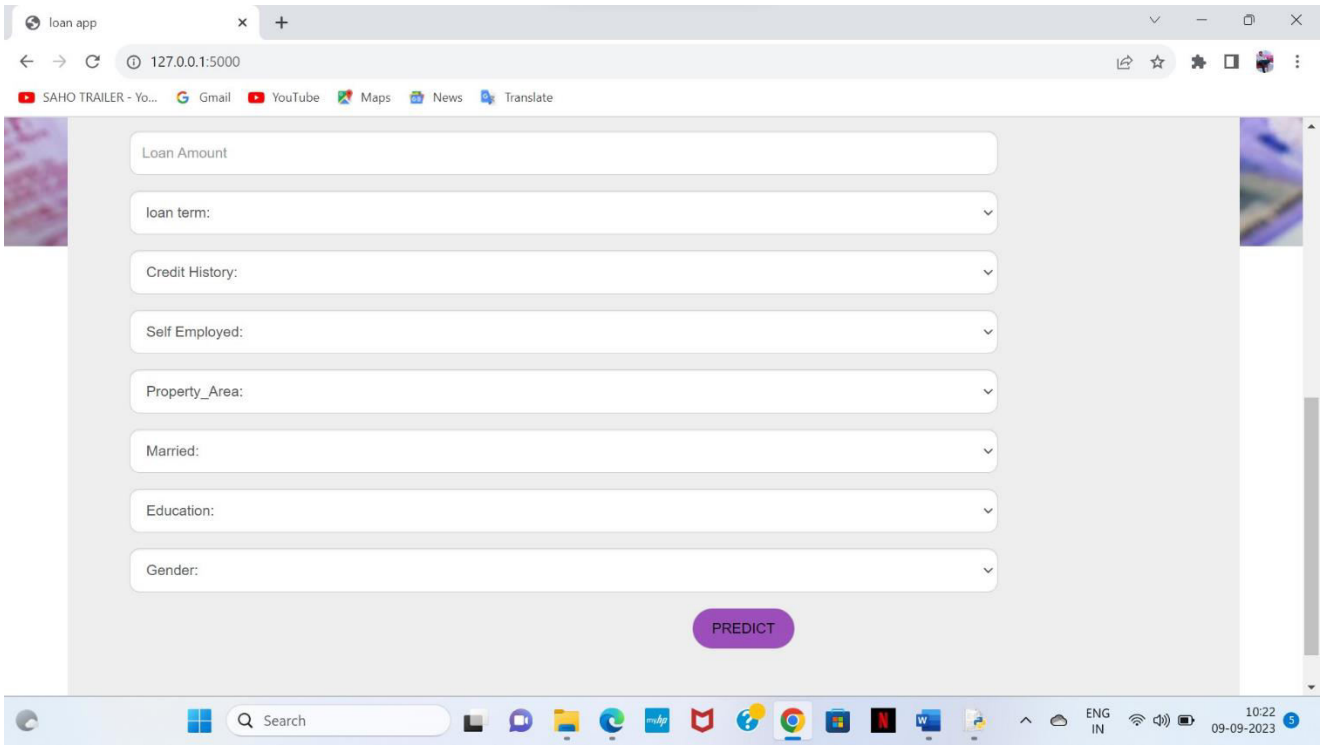


OUTPUT SCREEN

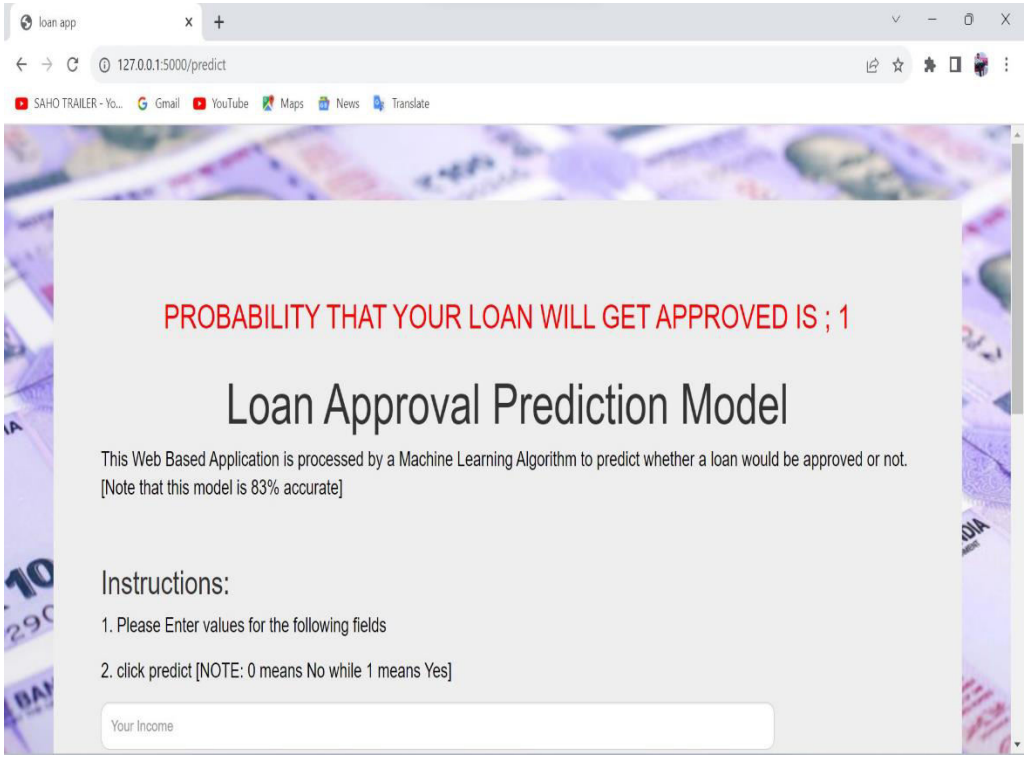


HOME SCREEN





LOAN APPROVAL SCREEN



7. CONCLUSION AND FUTURE SCOPE

In this paper, we have proposed customer loan prediction using supervised learning techniques for loan candidate as a valid or fail to pay customer. In this paper, various algorithms were implemented to predict customer loan. Optimum results were obtained using Logistic Regression, Random Forest, KNN, and SVM, decision Tree Classifier. Compare these five algorithms random forest is the high accuracy. From a correct analysis of positive points and constraints on the part, it can be safely ended that the merchandise could be an extremely efficient part. This application is functioning properly and meeting to all or any Banker necessities. This part is often simply obstructed in several different systems. There are numbers cases of computer glitches, errors in content and most significant weight of option is mounted in machine-driven prediction system, therefore within the close of future the therefore called software system might be created more secure, reliable and dynamic weight adjustment. In close to future this module of prediction can be integrated with the module of machine-driven processing system.

With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this paper we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets.

The system is trained on old training dataset in future software can be made such that new testing data should also take part in training data after some fix time.

8. REFERENCES

[1] Yu Jin and Yudan Zhu, "A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending," School of Information, Zhejiang University of Finance and Economics, 310018 Hangzhou, China.

[2] https://www.kaggle.com/telco-churn

[3] Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal "Loan default forecasting using data mining" Department of Information Technology, St. Francis Institute of Technology, Mumbai, India (2020)

[4] Octave Iradukunda, Haiying Che, Josiane Uwineza, Jean Yves Bayingana, Muhammad S Bin-Imam, Ibrahim Niyonzima "Malaria Disease Prediction Based on Machine Learning" School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China (2019).

[5] G. Arutjothi, Dr. C. Senthamarai "Prediction of Loan Status in Commercial Bank using Machine Learning Classifier" department of computer applications government arts college (Autonomous) Salem, India (2017.)

[6] Mohammad Ahmad Sheikh, Amit Kumar Goel, Tapas Kumar "An Approach for Prediction of Loan Approval using Machine Learning Algorithm" School Of Computer Science And Engineering Galgotias University Greater Noida, India (2019).

[7] Xin Li, Xianzhong Long, Guozi Sun, Geng Yang, and Huakang Li "Overdue Prediction of Bank Loans Based on LSTM-SVM"Jiangsu Key Lab of Big Data and Security and Intelligent Processing Nanjing University of Posts and Telecommunications, Nanjing, 210023, China.

[8] Aakanksha, Tamara Denning, Vivek Srikumar, Sneha Kumar Kesera "secrets in source code: reducing false positives using ML" software engineering (Microsoft) school of computing, USA (2020)