PREDICTION OF CUSTOMER LOAN ELIGIBILITY USING RANDOM FOREST ALGORITHM

SRIKANTH EADARA1 , I.PHANI KUMAR2

#1Student, Dept of CSE, VelagaNageswaraRao College Of Engineering, Ponnur(Post),Ponnur(Md)Guntur(D.T)A. Andhra Pradesh.

#2Assoc. Professor, Dept of CSE, VelagaNageswaraRao College Of Engineering, Ponnur(Post),Ponnur(Md)Guntur(D.T)A. Andhra Pradesh

ABSTRACT_A veritably important approach in prophetic analytics is used to study the problem of prognosticating loan defaulters The data is collected from the Kaggle for studying and vaticination. Random timber models have been performed and the different measures of performances are reckoned. The models are compared on the base of the performance measures similar as perceptivity and particularity. The final results have shown that the model produce different results. thus, by using a Random timber approach, the right guests to be targeted for granting loan can be fluently detected by assessing their liability of dereliction on loan. The model concludes that a bank shouldn't only target the rich guests for granting loan but it should assess the other attributes of a client as well which play a veritably important part in credit granting opinions and prognosticating the loan defaulters

1.INTRODUCTION

Finance companies deals with all kinds of loans such as house loans, vehicle loans, educational loans, personal loans etc... And has a presence across areas such as cities, towns and village areas. A Customer-first requests for a loan and after that Finance Company validates the customer eligibility for the loan ap-provel. Details like marital status, gender, education, and number of dependents, Income, Loan Amount, credit history, and others are given in the form to fill up by the applicants. Therefore, a robust model is built taking those details as input to verify whether an applicant is eligible to apply for loan or not. The target variable here is Applicants "Loan Status" and the other variables are predictors. After building the Machine Learning model a Web Application is to be developed for a user interface that allows

the user to see instantly if he/she is eligible to get a loan by entering the given details.

2.LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company Traffic Redundancy Elimination, once these things are satisfied, then next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support.

This support can be obtained from senior programmers, from book or from websites. Before building the system we have to knownthe below concepts for developing the proposed system.

[1] S. S. Sannakki and V. S. Rajpurohit, proposed a “Classification of Pomegranate Diseases Based on Back Propagation Neural Network” which mainly works on the method of Segment the defected area and color and texture are used as the features. Here they used neural network classifier for the classification. The main advantage is it Converts to L*a*b to extract chromaticity layers of the image and Categorisation is found to be 97.30% accurate. The main disadvantage is that it is used only for the limited crops.

[2] P. R. Rothe and R. V. Kshirsagar introduced a” Cotton Leaf Disease Identification using Pattern Recognition Techniques” which Uses snake segmentation, here Hu’s moments are used as distinctive attribute. Active contour model used to limit the vitality inside the infection spot, BPNN classifier tackles the numerous class problems. The average classification is found to be 85.52%.

[3] Aakanksha Rastogi, Ritika Arora and Shanu Sharma,” Leaf Disease Detection and Grading using Computer Vision Technology &Fuzzy Logic”. K-means clustering used to segment the defected area; GLCM is used for the extraction of texture features, Fuzzy logic is used for disease grading. They used artificial neural network (ANN) as a classifier which mainly helps to check the severity of the diseased leaf.

[4] Godliver Owomugisha, John A. Quinn, Ernest Mwebaze and James Lwasa, proposed" Automated Vision-Based Diagnosis of Banana Bacterial Wilt Disease and Black Sigatoka Disease "Color histograms are extracted and transformed from RGB to HSV, RGB to L*a*b.Peak components are used to create max tree, five shape attributes are used and area under the curve analysis is used for classification. They used nearest neighbors, Decision tree, random forest, extremely randomized tree, Naïve bayes and SV classifier. In seven classifiers extremely, randomized trees yield a very high score, provide real time information provide flexibility to the application.

3.PROPOSED SYSTEM

In this project we are using machine learning algorithm called Random Forest to predict loan eligibility and to train this random forest we are using below dataset

Prediction of granting the loan to the customers by the bank is the proposed model. Classification is the target for developing the model and hence using Random Forest with sigmoid function is used for developing the model. Preprocessing is the major area of the model where it consumes more time and then Exploratory Data Analysis which is followed by Feature Engineering and then Model Selection. Feeding the two separate datasets to the model, and then preceding the model.

3.1 IMPLEMENTATION

3.1.1 RANDOM FOREST ALGORITHM

First, Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach.

The author gives 4 links to help people who are working with decision trees for the first time to learn it, and understand it well. The decision tree is a decision support tool. It uses a tree-like graph to show the possible consequences. If you input a training dataset with targets and features into the decision tree, it will formulate some

set of rules. These rules can be used to perform predictions. The author uses one example to illustrate this point: suppose you want to predict whether your daughter will like an animated movie, you should collect the past animated movies she likes, and take some features as the input. Then, through the decision tree algorithm, you can generate the rules. You can then input the features of this movie and see whether it will be liked by your daughter. The process of calculating these nodes and forming the rules is using information gain and Gini index calculations.

The difference between Random Forest algorithm and the decision tree algorithm is that in Random Forest, the process es of finding the root node and splitting the feature nodes will run randomly.

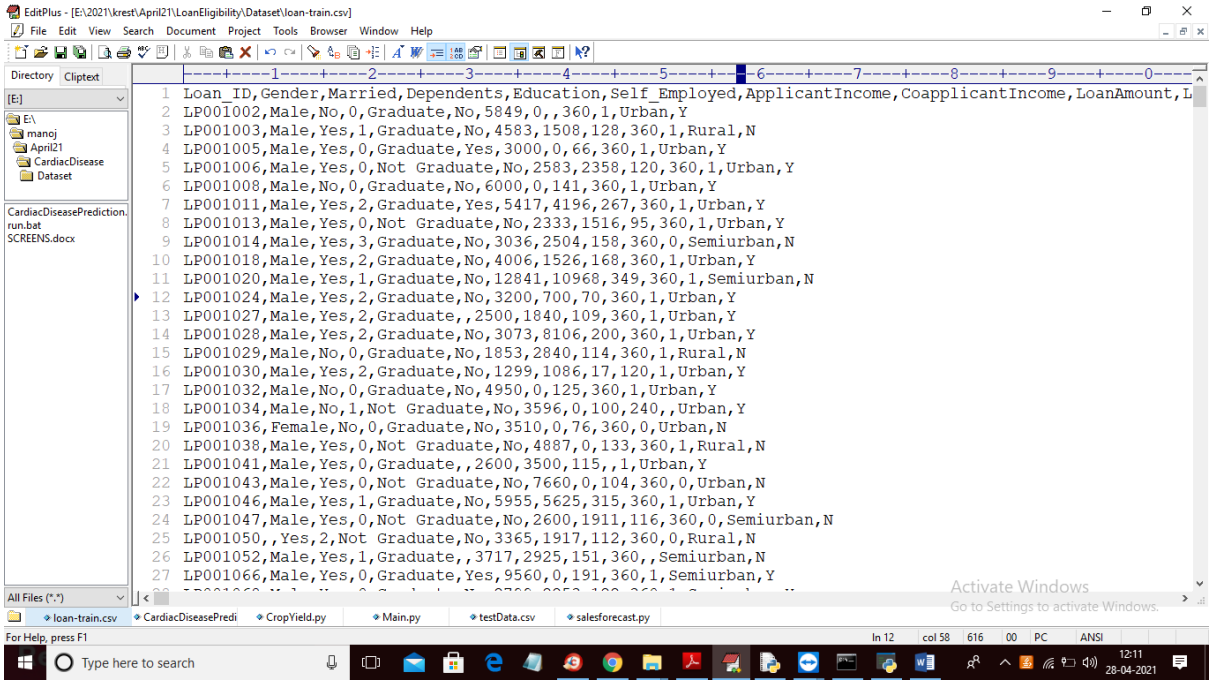
3.1.2 Why Random Forest algorithm?

The author gives four advantages to illustrate why we use Random Forest algorithm. The one mentioned repeatedly by the author is that it can be used for both classification and regression tasks. Overfitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won't overfit the model. The third advantage is the classifier of Random Forest can handle missing values, and the last advantage is that the Random Forest classifier can be modeled for categorical values.

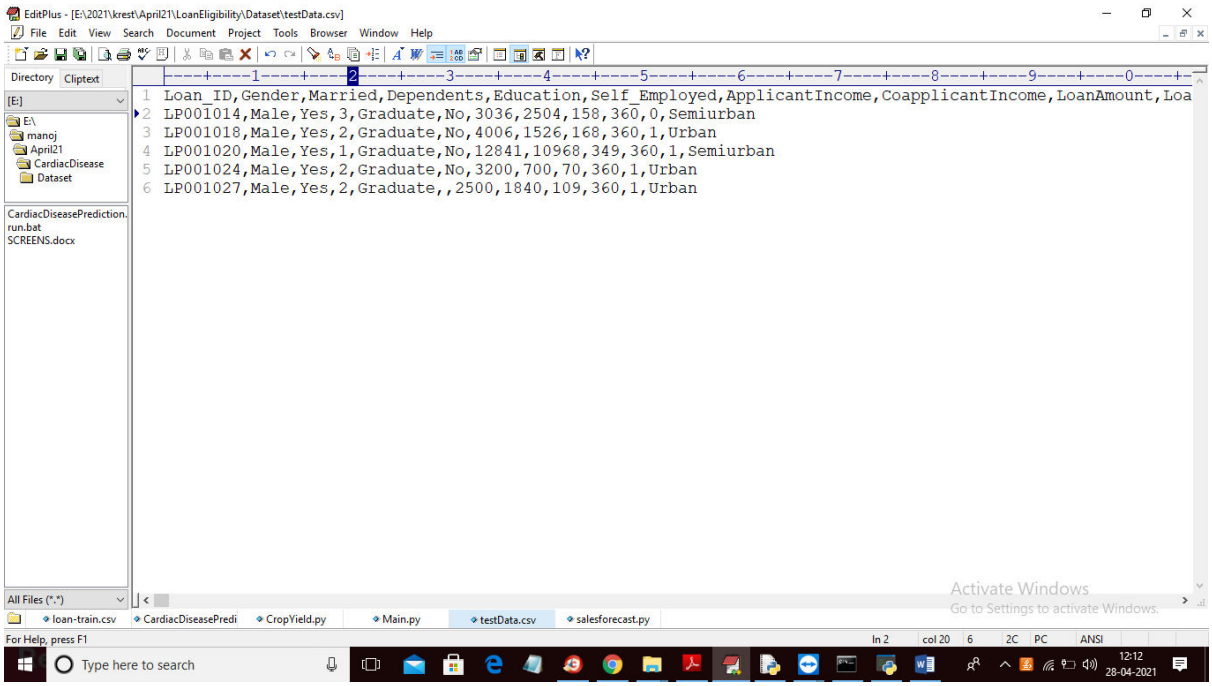
4.RESULTS AND DISCUSSION

4.1 DATASET

In this project we are using machine learning algorithm called Random Forest to predict loan eligibility and to train this random forest we are using below dataset

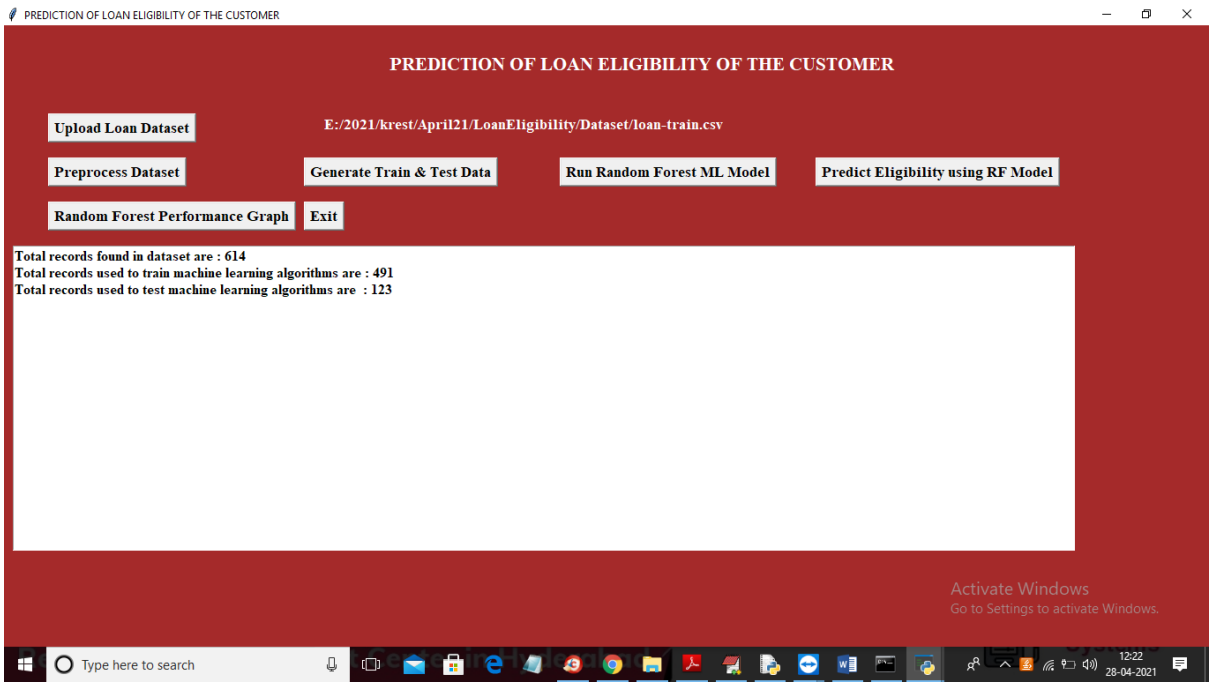


In above dataset in first row we can see dataset column names and in other rows we have dataset values and in last column we have class label as Y or N where Y means eligible and N means not eligible and now we used above dataset to train machine learning model and after training we will upload test dataset and then application will predict class label Y or N and below is test dataset screen shots



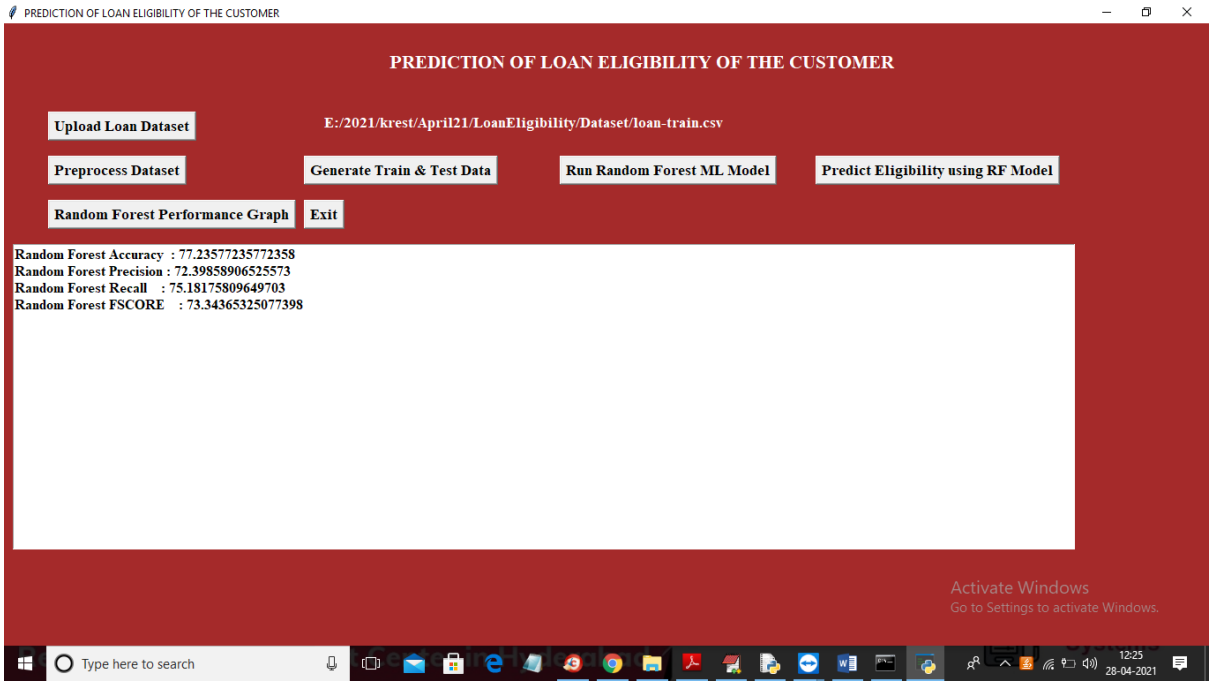
In above test data we don't have any N or Y class label and by analysing above records machine learning will predict eligibility.

4.2 RESULTS

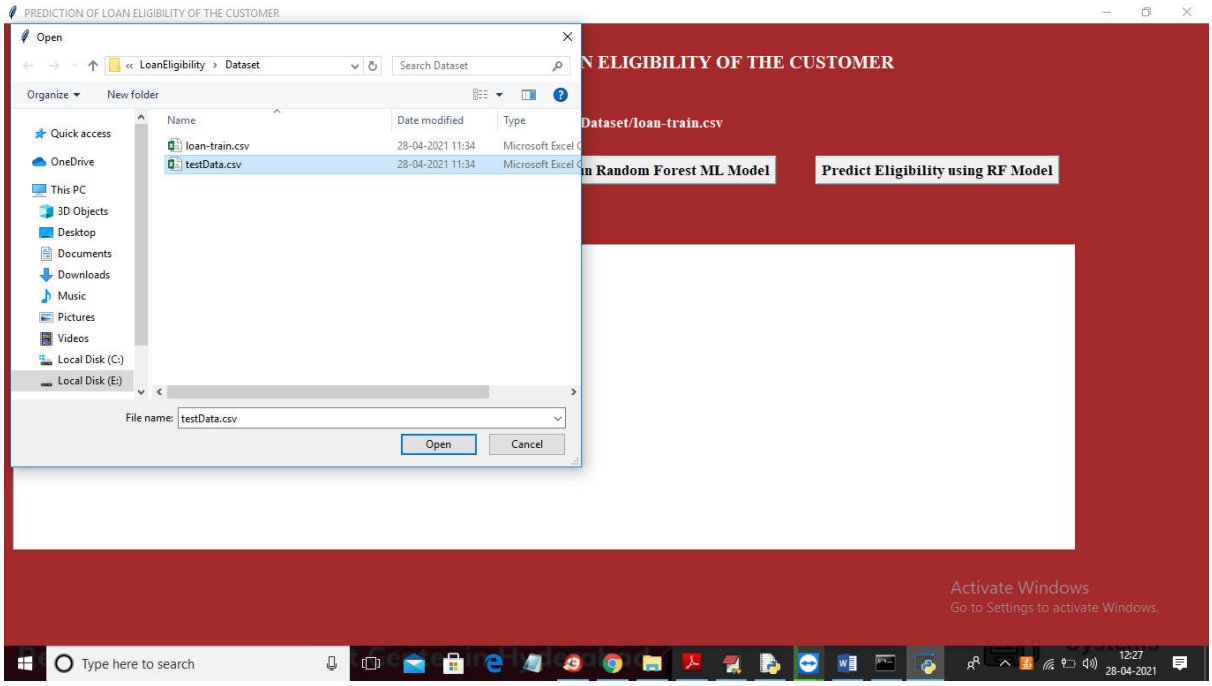


In above screen dataset contains 614 records and using 491 records to train ML and 123 records to test ML accuracy. In below graph we can see importance of each attribute with other attribute by using graph correlation metric

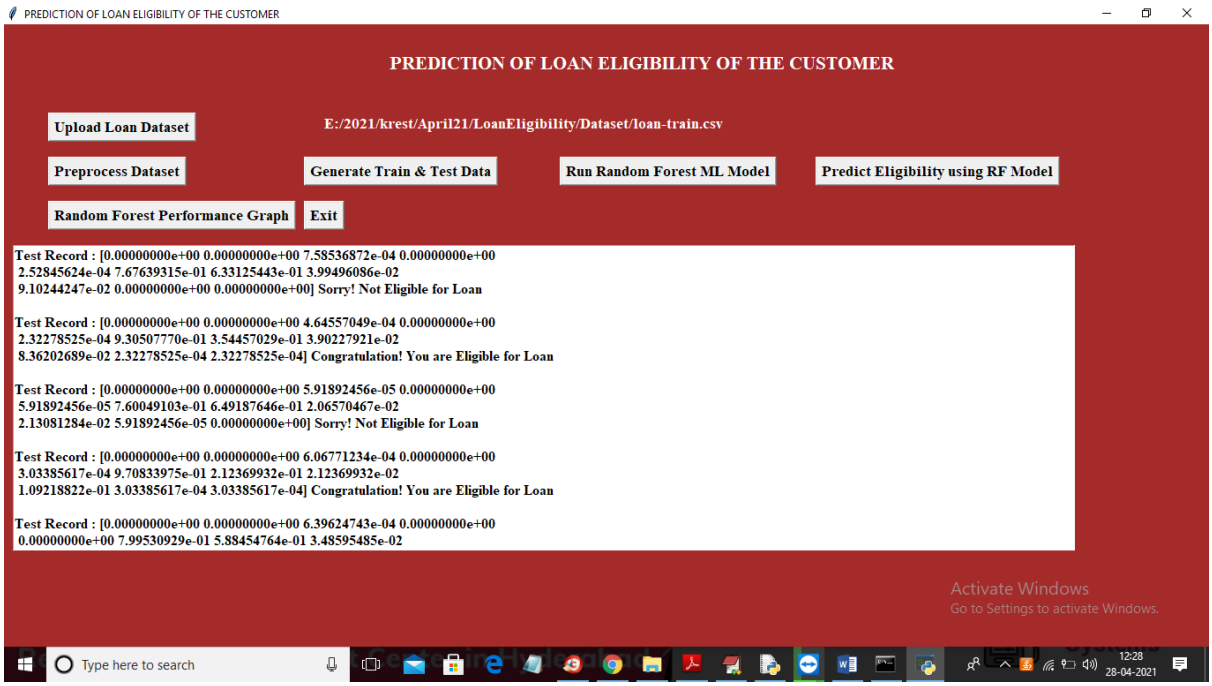
In above graph whatever column in x-axis and y-axis having value >0 will be consider as important features or column. Now click on 'Run Random Forest Ml Model' to build random forest model on above dataset



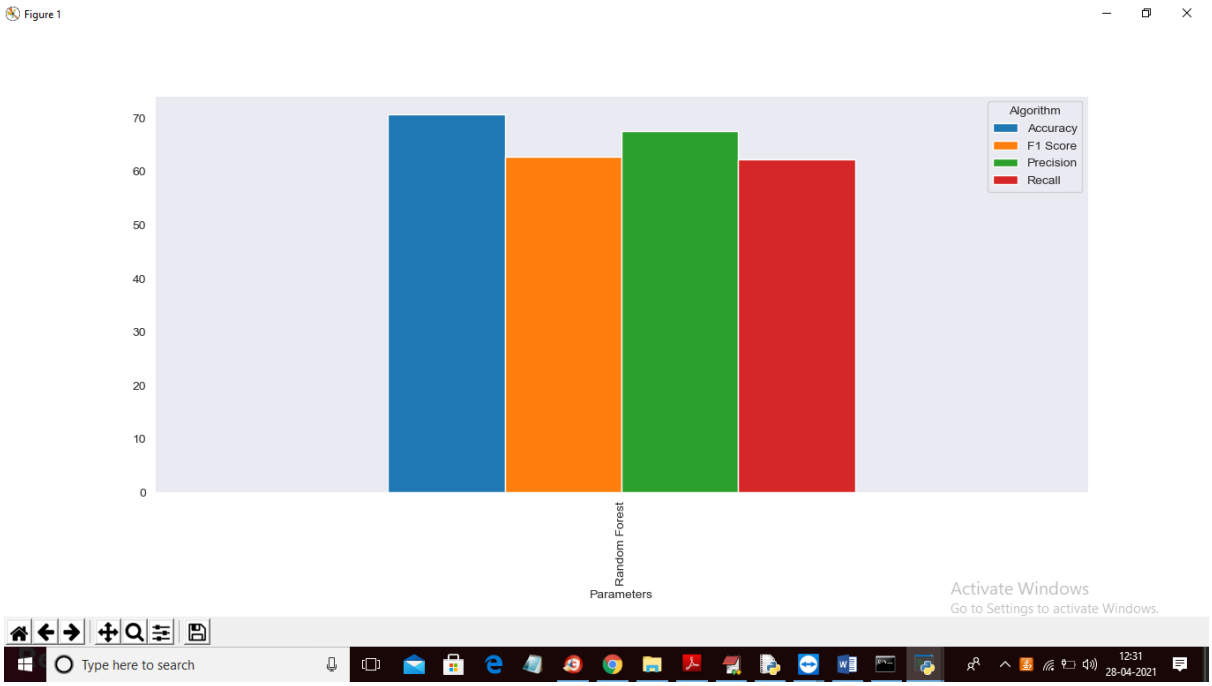
In above screen random forest model generated with 77% accuracy and we can see its precision, recall and FSCORE value and now click on 'Predict Eligibility using RF Model' button to upload test data and perform eligibility prediction



In above screen selecting and uploading 'testData.csv' file and then click on 'Open' button to load test data and then will get below prediction result



In above screen in square bracket we can see normalized test values and after square bracket we can see the prediction result as eligible or not eligible. You can scroll down above text area to view all predicted records and now click on 'Random Forest Performance Graph' button to get below graph



In above graph we can see accuracy, precision, recall and FSCORE values of random forest and graph y-axis represents %value where accuracy got 80% and Precision got 65%. Each metric bar colour name you can see from top right side

5.CONCLUSION

As a result, the proposed model automates the procedure of determining the creditworthiness of the applicant. It focuses on data comprising the essential points of loan applicants. The random forest model is employed in this system. Random forest analysis is a supervised learning method in Machine Learning. As a consequence, it is useful for forecasting the proper result in the current world scenario and also helps the bank to put the money in the right hands and also helps the people in receiving loan in a lot faster pace. The key advantage of this approach is that it provides greater accuracy.

REFERENCE

[1] Toby Segaran, “Programming Collective Intelligence: Building Smart Web 2.0 Applications.” O’Reilly Media.

[2] Drew Conway and John Myles White,” Machine Learning for Hackers: Case Studies and Algorithms to Get you Started,” O’Reilly Media.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer ,Kindle

[4] PhilHyo Jin Do ,Ho-Jin Choi, "Sentiment analysis of real-life situations using loca- tion, people and time as contextual features," International Conference on Big Data and Smart Computing (BIGCOMP), pp. 39–42. IEEE, 2015.

[5] Bing Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, May 2012.

[6] Bing Liu, "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions," Cambridge University Press, ISBN:978-1-107-01789-4.

[7] Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng, "CNN for situations understanding based on sentiment analysis of twitter data," Procedia computer science, 111:376–381, 2017.CrossRef.

[8] K I Rahmani, M.A. Ansari, Amit Kumar Goel, "An Efficient Indexing Algorithm for CBIR,"IEEE- International Conference on Computational Intelligence & Communication Technology ,13-14 Feb 2015.

[9] Gurlove Singh, Amit Kumar Goel ,"Face Detection and Recognition System using Digital Image Processing" , 2nd International conference on Innovative Mechanism for Industry Application ICMIA 2020, 5-7 March 2020, IEEE Publisher.

[10] Amit Kumar Goel, Kalpana Batra, Poonam Phogat," Manage big data using optical networks", Journal of Statistics and Management Systems "Volume 23, 2020, Issue 2, Taylors & Francis.

AUTHOR'S PROFILE

SRIKANTH EADARA pursing M. Tech in Computer Science and Engineering from Velaga Nageswara Rao College Of Engineering, Ponnur. Affiliated to JNTUK, KAKINADA

I.Phani Kumar ,Qualifications: Ph.D ,M.Tech, having 13 years of *teaching experience,present he is working as Assoc.Prof in Velaga Nageswara Rao College of Engineering,Ponnur,Guntur(D.t),A.P,maild:phanikumar148@gmail.com,