AI model for money lending investors to make better decisions

Revanth Krishna Maddula
Software Engineering
San Jose State University
San Jose, CA
revanthkrishna.maddula@sjsu.edu

Sandeep Reddy Bhimireddy
Software Engineering
San Jose State University
San Jose, CA
sandeepreddy.bhimireddy@sjsu.edu

Kailash Nath Tallapragada
Software Engineering
San Jose State University
San Jose, CA
kailashnath.tallapragada@sjsu.edu

Saikrishna Nandikonda
Software Engineering
San Jose State University
San Jose, CA
saikrishna.nandikonda@sjsu.edu

I. INTRODUCTION

In the era of systems with artificial intelligence and development in technologies, there is a high need for making business decisions faster, in order to reduce the overload. In this, we had built a machine learning model that takes an input, set of attributes of the borrower, and predicts whether a loan disbursed by the investor in a peer to peer lending platform will charge off (unlikely to be repaid) or not.

Loans are very important for the people in need. At the same time, it is also important for the money lenders to make proper decisions in the scenario to sanction a loan or not. But money investors have to take many factors into consideration, making this process cumbersome. Hence there is a necessity to make this process escalate so that this business objective is achieved and therefore recognize and understand the needs. There are several attributes like FICO Score, number of open credit lines, annual income, debts.etc. Money Lending prediction algorithm helps to classify a particular customer into either of the two classes identified. Using machine learning algorithms, we have built a model that classifies the customer's class into charged off/fully paid.

The favorable prediction results are being achieved through the rigorous training of the machine model and thereby finetuning the parameters accordingly. Experimental results show that the accuracy was about 84%. Compared with the algorithms which were proposed earlier, this algorithm has significant improvement in accuracy and performance in real-time.

II. LITERATURE SURVEY

In the paper [1] the specialists utilized the data mining strategy for analysing information. Data mining methodology gives an incredible vision in advance of expectation frameworks, since this will instantly recognize the clients who can reimburse the advance sum inside a period. Calculations like "J48 calculation", "Bayes net", Naive Bayes" are utilized. On applying these calculations to the datasets, it was demonstrated that "J48 calculation" has high precision (right percent) of around 79% which gives the financier to choose whether the advance can be given to the customer or not. In paper [2], "advance expectation utilizing Ensemble strategy", utilized "Tree model", "Irregular backwoods", "svm model" and consolidated the over three models as Ensemble model. A model has been talked about in paper [2] with the goal that the financial divisions can concur/dismiss the credit demand from their clients. The principle technique utilized is genuine coded

hereditary calculations. The consolidated calculations from the outfit model, credit expectation should be possible in a simpler manner. It is discovered that tree calculation gives high precision of 81.25%. In paper [3], utilizing R-language, an improved hazard expectation grouping calculation is utilized to locate the awful advance clients since likelihood of default (PD) is the basic advance for the clients who desire a bank credit. Along these lines, a casing work for discovering PD in the informational collection is given by information mining strategy. R-Language has the strategy called KNN (K-closest neighbor) calculation and it is utilized for playing out various ascription counts when there are missing qualities found in the informational index. The paper [4] had utilized a tree model. It assists with discovering whether the financial segment individuals will have the option to defeat the credit issue with their clients. It gives a high exactness of 80.87%.

III. DATA PRE-PROCESSING

The dataset consists of 2,260,701 entries (customer records) and 151 columns taken from the peer to peer money lending platform-LendingClub. In order to validate the results, 20% of the dataset is taken into testing data. Since many attributes do not play an influence in deciding the final outcome, some attributes are discarded accordingly to make the predictions better. This is done by initially finding the correlation factor between the individual attributes and the final result.

This process is done on each of the attributes, so that the dataset becomes clean, and thereby discarding the attributes which don't come into picture due to the lack of correlations with the resulting outcome.

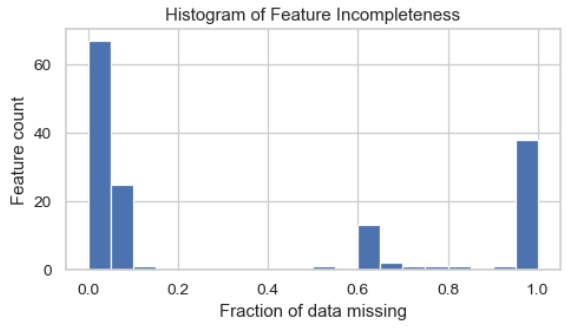

Fig 1.Incomplete Data visualization

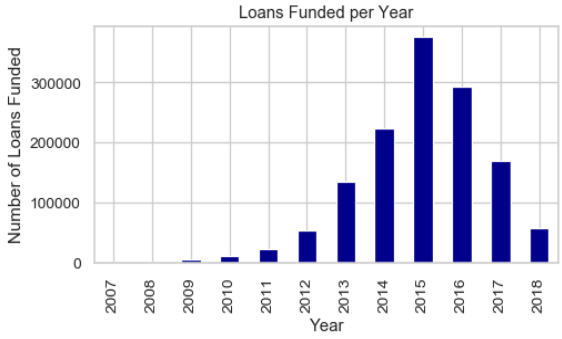

Fig 2. Data visualization of Loans

Sample analysis of a attribute "interest rate":

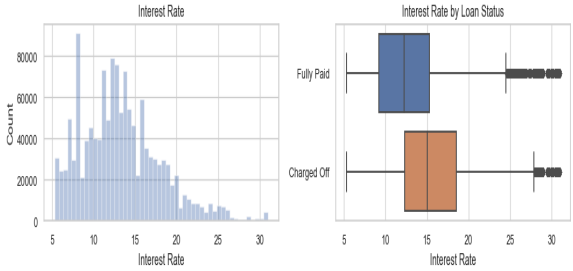

Fig 3.Analysis on interest rate attribute

IV. MACHINE LEARNING MODELS

A. Logistic Regression with SGD:

Stochastic gradient descent is an algorithm which will consider one point as random and changes the weights, which is considered as dataset.

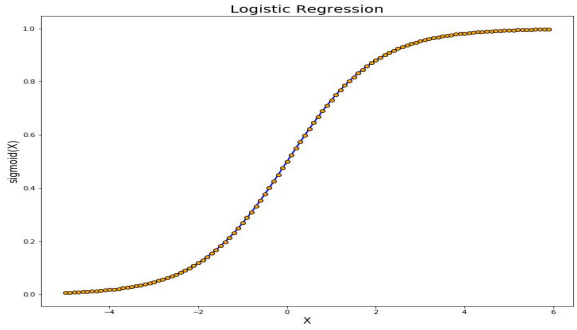

Fig 4. Logistic Regression Graph

Logistic regression is used for categorical classification, which is exactly the case here, where we have two classes. Logistic regression can be done by setting up the parameter loss variable.

B. Random Forest Classifier

This classifier is a mixture of non-related entities, where the trees are not necessarily related, it works better with the unknown data also by giving good predictions.

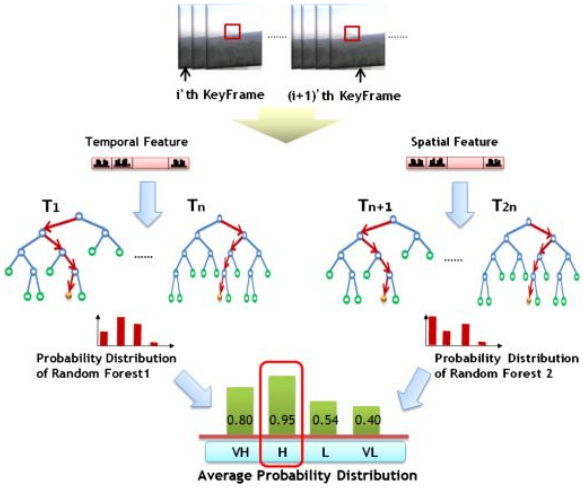

Fig 5. Random Forest Classifier

Pros of using Random Forest:

- The model is fast in being trained by multiple decision trees.
- It gives accurate predictions, with the data which is either missing or having outliers.

C. Gradient Boosting:

This is a machine learning technique which is used for classification and regression problems, which will produce a prediction model like an ensemble of less accurate prediction models, like decision trees.

Grouping calculations every now and again utilize logarithmic misfortune, while relapse calculations can utilize squared mistakes. Inclination boosting frameworks don't need to determine another misfortune work each time the boosting calculation is included, rather any differentiable misfortune capacity can be applied to the framework.

Gradient boosting frameworks have two other vital parts: a feeble student and an added substance segment. Inclination boosting frameworks use choice trees as their powerless students.

Relapse trees are utilized for the frail students, and these relapse trees yield genuine qualities. Since the yields are genuine qualities, as new students are included into the model the yield of the relapse trees can be included to address blunders in the forecasts.

The added substance segment of an angle boosting model originates from the way that trees are added to the model after some time, and when this happens the current trees aren't controlled, their qualities stay fixed.

The new tree's yield is then affixed to the yield of the past trees utilized in the model. This procedure is rehashed until a formerly indicated number of trees is reached, or the misfortune is decreased beneath a specific edge.

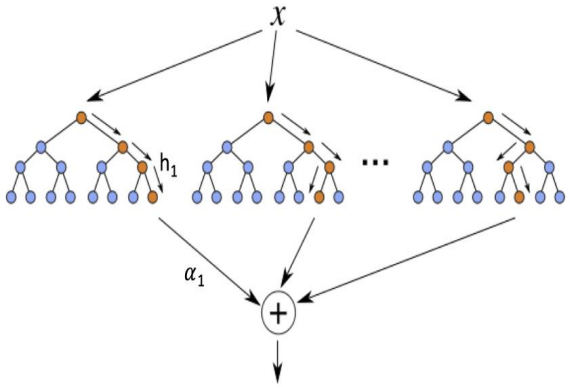

Fig 6.Gradient Boosting

V. ARCHITECTURE

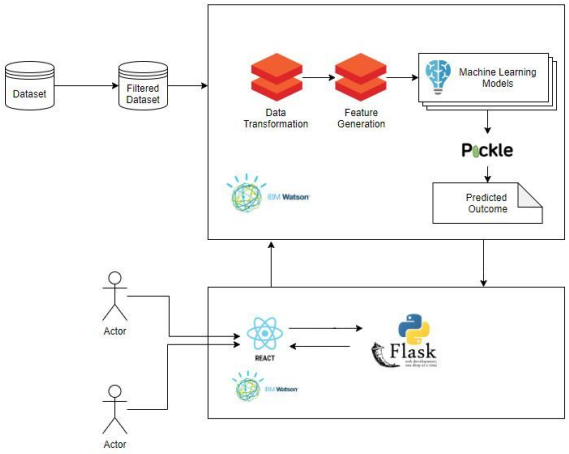

Fig 7. Model Architecture

VI. RESULT

The results obtained from our model to predict that loan will be charged off is given below.

Model	**Accuracy (%)**
Logistic Regression with SGD	74.21%
Random Forest Classifier	78%
Gradient Boosting	84.38%

Table 1. Model accuracy table

As the table suggests that the gradient boosting model has the highest accuracy of all the models tried, It has been used to validate the business model created.

VII. FURTHER WORK

As the model currently used has only an accuracy of 84 percent, It would be highly unmanageable to use this in a real life fully equipped business scenario.

To make it more effective and more business oriented, there is a definite need to build a model with high accuracy as close as 90 percent.

VIII. REFERENCES

[1] A. Goyal and R. Kaur, "A survey on Ensemble Model for Loan Forecast", International Journal of Engineering Trends and Applications (IJETA), vol. 3(1), pp. 32-37, 2016.

[2] A. J. Hamid and T. M. Ahmed, , "Developing Prediction Model of Loan Risk in Banks using Data Mining".

[3] G. Shaath, "Credit Risk Analysis and Prediction Modeling of Bank Loans using R".

[4] A. Goyal and R. Kaur, "Accuracy Prediction for Loan Risk Using Machine Learning Models"

[5] Kumar Arun, Garg Ishan, Kaur Sanmeet, "*Loan Approval Prediction based on Machine Learning Approach,*" IOSR Journal of Computer Engineering (IOSR-JCE, Medea, Algeria, 2016, pp. 18-21

[6] J.M. Chambers. Computational methods for data analysis. Applied Statistics, Wiley, 1(2):1–10, 1077

[7] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. R News, 2(3):9–22, 2002

[8] X.Francis Jency, V.P.Sumathi, Janani Shiva Sri , "An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients", International Journal of Recent Technology and Engineering (IJRTE), November 2018