143

by Mmmm Llll

Submission date: 09-May-2024 10:23PM (UTC-0500)
Submission ID: 2375668215
File name: FOML_Paper_210701143.docx (95.23K)
Word count: 3468
Character count: 18183

MACHINE LEARNING IN FINANCIAL SERVICES: PREDICTING LOAN APPROVAL OUTCOMES

Abstract – *The financial sector is faced with a huge volume of data. Loans are considered the major profitable area for banks, with most of their profit coming from loans. While loans provide higher profit, the process of loan approval is challenging and error-prone, as customers who are approved for loans may lead to loan defaults. Machine learning can help overcome this situation by providing a solution through the use of machine learning techniques and models. In this process, the machine is trained with a predefined set of data containing customer details and important data to be considered for loan approval. The machine then classifies new user data as either accepted or rejected for a loan. In this study of comparing different machine learning methods like Logistic Regression, Gradient Boosting, SVM, Random Forest, and Voting Classifier, Random Forest was the best at predicting loan approvals accurately. This research is beneficiary for banks trying to improve their loan approval processes with large amounts of loan applications on a daily basis. Even with the current credit scoring systems, there are still loan defaults. This study uses machine learning to predict future loan defaulters by analyzing past customer data.*

Key Words: Loan Approval, Machine learning, Random Forest, Logistic Regression, Gradient Boosting, Support Vector Machine (SVM), Voting Classifier

I. INTRODUCTION

Loans can be thought of as a sum of money that can be borrowed from the banks or any other financial service providers in order to manage their planned or any events that are not planned. In the perspective of banks, Loans are considered to be the primary source of profit or income. Since there are large numbers of loan applications on a daily basis, it is seen as a big threat to the financial sector due to the increasing count of loan defaults by the individuals. In general, these loan defaults mean the incapability of an applicant to repay the loan amount on correct time, which occur due to the wrong selection of people for providing loans. As the data is huge, manual data collection, processing, validating and providing accurate results become very hectic and time consuming process and largely considered to be error-prone due to many reasons. When picking an applicant for loan approval, they must take into account certain bank policies. The bank has to determine who will be the best applicant for approval after considering several factors. It is challenging and risky to physically verify each individual before recommending them for loan approval. Hence the machine learning approach used in this study helps to predict whether an individual has the ability to repay the loan based on the past customer's data and provides results as accepted or rejected. This study majorly focuses on training and comparing different machine learning models and chooses the best model that accurately predicts the loan approval of a customer. This study uses five different models of machine learning, which are notably Logistic Regression, Gradient Boosting, SVM, Random Forest, and Voting Classifier. It does comparison by fitting the data into the model and predicting whether they produce accurate results by using performance measurement like accuracy and f1 score, which also helps us to come to a conclusion that which model gives better outcome and choose the best model according to our requirements.

II. LITERATURE SURVEY

The study [1] investigated how machine learning algorithms may be used to forecast a loan's condition. Among the algorithms tested was SVM, which had an accuracy rating of 83%. However, the study also showed that a number of variables, like data quality, hyper parameter selection, and the presence of databases in the model or data, might affect how effective SVM is. According to the study's findings overall, SVM has the potential to be a useful tool for predicting loan status, but its effectiveness will rely on a variety of circumstances. Based on the study's [2] predictive analysis of anticipated events, a collaborative system that prioritizes wait times for customers has been developed. A new loan application system emerges every day. The bank will evaluate a customer's loan request and determine whether or not to accept it based on the customer's qualifications. In this case, the final clearance is determined by forecasting historical customer data using machine learning algorithms. Accurately forecasting the customer's repayment of loans and documenting it for future prediction are made possible by the Random Forest (RF) Algorithm. The recommended strategy decreases time complexity while improving precision. In this work [3], patterns are extracted from a shared dataset of loans that have been offered via machine learning (ML) techniques so as to estimate future borrowers who default on loans. Analysis will be done using prior data from customers, including age, financial status, loan amount, and the duration of occupation. Various machine learning methods, notably Random Forest, Support Vector Machine, K-Nearest Neighbor, and Logistic Regression, were employed to identify the most relevant features, or the ones that have the most impact on the prediction outcome. The aforementioned algorithms are evaluated using common measures and contrasted with one another. The accuracy of the random forest method is superior. In this study [4], we employ a machine learning approach that, based on the prior record of the individual to whom the loan amount has previously been certified, will forecast who is dependable for a loan. Predicting whether or not a certain person will be approved for a loan is the main goal of this effort. The suggested [5] loan approval predicting method is a machine learning-based web service that offers consumers instant loan acceptance forecasts. In addition to calculating a credit score known as a CIBIL score, the application employs logistic regression to forecast the likelihood of loan acceptance. With every aspect considered, the approval of loans predicting technique is a useful tool for people and financial organizations that want to evaluate loan applications fast and decide on the best course of action. It makes use of machine learning to offer predictions that are trustworthy and accurate, and it additionally provides consumers a straightforward and readily available method to access this capabilities. This paper [6] uses gradient boosting, an Extreme Gradient Boosting method, to forecast loan default. The estimate is backed by data on loans from the internet-based Super Lender. We examine details gathered from the loan application along with demographic data. The investigation's F1-Score, Accuracy, Recall, and Precision are a few of the crucial assessment criteria we outline here. By employing predictive modeling to separate high-risk applicants from an extensive quantity of loan applications, this study offers a solid basis for loan credit approval. For the purpose to assess whether it would be feasible to approve each individual loan request, this research [7] suggests combining ensemble learning techniques with machine learning models. This method can improve the precision with which qualified applicants are chosen from an already existing list. As a result, the issues raised above about loan approval procedures can be resolved by using this procedure. The concept is beneficial to applicants as well as bank employees because it significantly shortens the time needed to authorize loans.

III. PROPOSED METHODOLOGY

This study compares different machine learning models to predict the loan approval for individuals based on 13 features which are present in a dataset. Here are the steps that are performed to make this analysis of different models.

1. Data Preprocessing

Banks get their large amount of profit from loan as many applicants get loan and are liable to repay the loan and the amount of interests are also increasing on a monthly or yearly basis. This study uses a dataset which is available in kaggle.com. The dataset is “loan_approval_dataset.csv” which has the following columns loan_id, no_of_dependents, education, self_employed, income_annum, loan_amount, loan_term, cibil_score, residential_assets_value, commercial_assets_value, luxury_assets_value, bank_asset_value, loan_status where loan_status is the output feature.

This study uses pandas and numpy for creating the dataframe and preprocessing the data. The dataset is visualized by using the head() method and describe method. The head() method displays the first five data or rows in the dataset in order to get an idea about the dataset. The describe() method is used to view the count, mean , std, min, 25%, 50%, 75%, max of each column in the dataset. This gives an overview of the dataset.

The info() method is used to get the information about the non null count and the type of data of each column in the dataset. It also provides the count of how many integer and objects are there in the dataset.

The isnull().sum() method is used to visualize the sum of the non-null values of each column of the dataset.

Using a for loop to run through the columns of the dataset, this study uses unique() and len() method to find the length of the unique values in the dataset of each column, and print them on the console and visualize them to understand the unique values in the columns of the dataset.

To find the count of each value in each column, run a for loop through the columns of the dataset and used value_counts() method for each column.

2. Data encoding

Using OrdinalEncoder, this study encode the columns with object datatype to an integer datatype with 1's and 0's values to ease the process of the training and testing of the model.

Initially the loan_status column's value notably 'Rejected' and 'Approved' to 0.0 and 1.0 respectively. Similarly the other two columns with only two outputs are also encoded with 1.0 and 0.0

3. Data Visualization

This study majorly focuses on the comparison of different models for the prediction of loan approval and choosing the better model that gives better accuracy than others. So to attain this, it uses data visualization tools like matplotlib and seaborn.

Initially, the data is flattened using flatten() method , to convert every resultant 2 dimensional arrays into a single linear vector.

Using matplotlib, the data in each column is visualized using distplot which is otherwise known as distribution plot, which depicts the variation in the data distribution.

Along with the distplot, the boxplot is also created and visualized using matplotlib, which gives the features that are important in the classification of loan approval.

Using seaborn, the correlation matrix is created and visualized that provides the correlation between each and every columns in the dataset. After visualizing the heatmap using seaborn, the values less than or equal to 0.00 is removed and again the correlation matrix is visualized using heatmap.

With the aid of seaborn's scatterplot the relation between the loan term and cibil score is visualized and it is found that when the loan term is lower than 5 and the cibil score is less than 550 the chance of loan approval is 1 in some region. When the loan term is above 5 and the cibil score is less than 550, when when there is increase in loan term and the person is having less cibil score, their chance of getting a loan is very much less or zero.

On the other hand, when the cibil score is greater than 550, the chance of getting a loan is higher irrespective of the loan term. So, this cibil score acts as the important feature in the prediction of loan approval process.

4. Training and Testing of the data:

In this study, initially the training and testing data is split and stored in X and y variable. Using StandardScaler from sklearn.preprocessing, the training and testing data of the input features is given to a function called fit_transform().This StandardScaler() methos gives us the mean and the standard deviation of the data where the data is scaled and centered to have a mean of 0 and a standard deviation of 1.

In the first level, Random forest is implemented by using the parameter grid that defined the n_estimators, max_depth, min_samples_split, and min_samples_leaf. This random forest is a popular machine learning technique tha combines the output of multiple decision trees to reach a single result.

After performing the training, the best_params_ is called to get the best parameters among the grid. And the random forest classifier is called with the best parameters to get the better accuracy.

After training with the help of fit() method, the prediction is done using the predict() method and the score for accuracy and f1 score is calculated for the model.

It provides accuracy of about 0.9566744730679156 and f1 score of about 0.9651272384542884

Then the confusion matrix is created for the random forest model and the classification_report method is called and the output is printed as follows:

	precision	recall	f1-score	support
0.0	0.94	0.94	0.94	323
1.0	0.97	0.96	0.97	531
accuracy			0.96	854
macro avg	0.95	0.95	0.95	854
weighted avg	0.96	0.96	0.96	854

A heatmap is drawn between the True and Predicted values which helps to visualize how much values are correctly predicted and how many outliers points or noise points are present in the model.

For implementing other models, using import statement, logistic regression from linear_model, gradient boosting classifier, voting classifer from ensemble and svc from svm are imported. The accuracy_score and f1_score is imported from sklearn.metrics.

All four models like logistic regression, gradient boosting classifier , scv and ranfom forest classifer is called and stored in four different variables.

Voting classifier is called with the above four models to calculate which model is best using voting mechanisms. The voting method used here is "hard voting" which is used to produce majority voting.

All the above five models are trained using fit() method, which is giving the necessary input and output features and training the model with predefined set of data from the dataset.

Here a list is created with the five models' variable called 'models' and another list is created with the names of the five models called model_names and an empty list called scores is created.

Using a for loop, run through the different models in the model_names and models list, so that each model is tested using predict() method and their accuracy and f1 score is calculated and appended to the scores list as a key value pair like 'Model' : model name, 'Accuracy' : accuracy and 'F1 Score' : f1_score.

The scores list is converted into a dataframe using pd.DataFrame method and produces results as follows:

The Logistic Regression model produces accuracy of 0.927400 and f1 score of 0.941065

The Gradient Boosting model produces accuracy of 0.949649 and f1 score of 0.959624

The SVM model produces accuracy of 0.950820 and f1 score of 0.959770

The Random Forest model produces accuracy of 0.950820 and f1 score of 0.960227

The Voting Classifier model produces accuracy of 0.953162 and f1 score of 0.961315

Then these scores like accuracy and f1 scores are plotted using matplotlib, as a bar chart which are given in the results section.

5. Important Features

The important feature in the dataset is found using feature_importances_ method by giving the column values from the random forest classifier model and the values are sorted to get the most important feature. To visualize this, a graph is plotted between input features and importance, where we found that cibil score has much importance when compared with other features.

Then the top k features are selected by slicing the list to top 5 features and printing them, we found cibil_score, loan_term, luxury_assets_value, residential_assets_value, and commercial_assets_value to be the top 5 features, which has more impact on predicting the output, that is whether the loan is approved or rejected.

IV. EXPERIMENTATION AND RESULTS

This study had made some comparison and obtained the accuracy and f1 scores for five different models and those values are listed in the following table:

	Model	Accuracy	F1 Score
0	Logistic Regression	0.927400	0.941065
1	Gradient Boosting	0.949649	0.959624
2	SVM	0.950820	0.959770
3	Random Forest	0.950820	0.960227
4	Voting Classifier	0.953162	0.961315

Table 1: Scores of different models

From the above table 1, we can able to see that the f1 score for Random Forest is higher and it compares the five different models with their accuracy and f1 scores, which provides an efficient way to compare the different models. This analysis helps researchers, banking systems, individuals who are trying to apply for a loan to get a fast and efficient loan approval prediction in minutes, which will save lot of their time and efforts.

Model Accuracy Comparison:

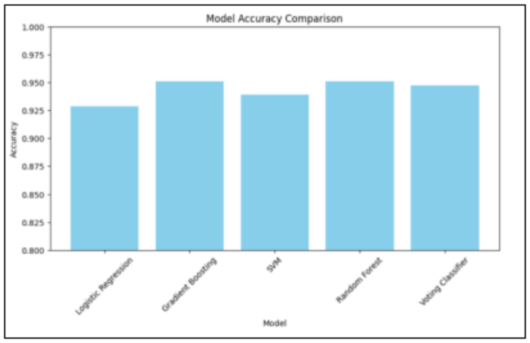

Figure: 1 Model Accuracy Comparison

The figure 1 depicted above explains the accuracy of various models that is given in a bar chart, which clearly describes which model gives better accuracy in the process of approval of a loan by the bank to the individuals who are applying for a loan.

In the x co-ordinate, different models are placed and in the y co-ordinate accuracy ranges from 0.800 to 1.000, where all the models lies above the point 0.925, that implies all the five models produces better accuracy and able to predict the correct individual who will be able to repay the loan on the scheduled time period.

Model F1 Score Comparison:

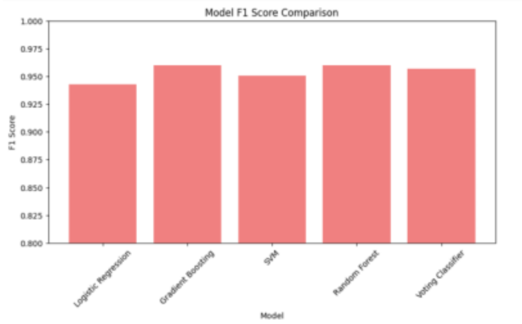

Figure: 2 Model F1 Score Comparison

The figure 2 depicted above explains the f1 score of various models that is given as a bar chart, which clearly describes which model gives better performance in the process of approval of a loan by the bank to the individuals who are applying for a loan.

In the x co-ordinate, different models are placed and in the y co-ordinate f1 score ranges from 0.800 to 1.000, where all the models lies above the point 0.925, that implies all the five models produces better performance and able to predict the correct individual who will be able to repay the loan on the scheduled time period.

V. CONCLUSION

This study concludes that machine learning acts as a life saver in every area including banking sector that eases the manual process of predicting the approval of a loan by checking each and every application on a daily basis, which takes lot of human effort and time. While comparing the four different models the accuracy and f1 score for Random Forest is higher than other models except Voting Classifier that uses all the four models to produce the best model among four using hard voting which calculates the best model based on the majority of votes.

As the dataset used in for this study contains only 13 columns and 4269 entries, the classification is done easily and efficiently, but when handling large datasets, one must concentrate on the feature selection that would make the model more accurate and efficient.

This loan approval prediction helps to avoid selecting wrong people who would lead to loan defaults or otherwise who could not be able to repay the loan. The models used here makes the process of predicting the loan approval by using pre-processed data that contains no null values, which added an advantage in improving the accuracy and f1 scores of all the models.

This study would help the people in the banking sector or researchers or anyone who are interested in knowing whether an individual who is applying for a loan will get approval or not within minutes, with less manual effort and less error.

VI. REFERENCES

[1] R. Nancy Deborah, S. Alwyn Rajiv, A. Vinora, C. Manjula Devi, S. Mohammed Arif and G. S. Mohammed Arif, "An Efficient Loan Approval Status Prediction Using Machine Learning," *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, Mumbai, India, 2023.

[2] C. Prasanth, R. P. Kumar, A. Rangesh, N. Sasmitha and D. B, "Intelligent Loan Eligibility and Approval System based on Random Forest Algorithm using Machine Learning," *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, Uttarakhand, India, 2023.

[3] P. Tumuluru, L. R. Burra, M. Loukya, S. Bhavana, H. M. H. CSaiBaba and N. Sunanda, "Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms," *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Coimbatore, India, 2022.

[4] A. Gupta, V. Pant, S. Kumar and P. K. Bansal, "Bank Loan Prediction System using Machine Learning," *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, Moradabad, India, 2020.

[5] E. Kadam, A. Gupta, S. Jagtap, I. Dubey and G. Tawde, "Loan Approval Prediction System using Logistic Regression and CIBIL Score," *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2023.

[6] P. Nagaraj, K. Nikhil, K. V. S. Sai Ram Santosh Babu, D. H. T. Reddy, R. R. Sekar and T. D. Rajkumar, "Loan Prediction Analysis Using Innumerable Machine Learning Algorithms," *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, Chennai, India, 2023.

[7] K. Bhatt, P. Sharma, M. Verma and K. Agarwal, "Loan Status Prediction in the Banking Sector using Machine Learning," *2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, Ghaziabad, India, 2023.

143

ORIGINALITY REPORT

18%	11%	12%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Cincinnati Student Paper	2%
2	Submitted to Coventry University Student Paper	2%
3	www.ijraset.com Internet Source	2%
4	R Nancy Deborah, S Alwyn Rajiv, A Vinora, C Manjula Devi, S Mohammed Arif, G S Mohammed Arif. "An Efficient Loan Approval Status Prediction Using Machine Learning", 2023 International Conference on Advanced Computing Technologies and Applications (ICACTA), 2023 Publication	1%
5	Submitted to Oxford Brookes University Student Paper	1%
6	Submitted to Liverpool John Moores University Student Paper	1%
7	machinelearninghd.com	

	Internet Source	1%
8	Submitted to University of Teesside Student Paper	1%
9	P. Nagaraj, K. Nikhil, K. V. S. Sai Ram Santosh Babu, D. Hari Tejaswar Reddy, R. Raja Sekar, T. Dhiliphan Rajkumar. "Loan Prediction Analysis Using Innumerable Machine Learning Algorithms", 2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), 2023 Publication	1%
10	Submitted to University of West London Student Paper	1%
11	ijisrt.com Internet Source	1%
12	C. Prasanth, R. Praveen Kumar, A. Rangesh, N. Sasmitha, Dhiyanesh B. "Intelligent Loan Eligibility and Approval System based on Random Forest Algorithm using Machine Learning", 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), 2023 Publication	1%
13	Emilie Nault, Peter Moonen, Emmanuel Rey, Marilyne Andersen. "Predictive models for assessing the passive solar and daylight	<1%

	potential of neighborhood designs: A comparative proof-of-concept study", Building and Environment, 2017 Publication	
14	github.com Internet Source	<1 %
15	Submitted to National College of Ireland Student Paper	<1 %
16	sifisheriessciences.com Internet Source	<1 %
17	webthesis.biblio.polito.it Internet Source	<1 %
18	www.igi-global.com Internet Source	<1 %
19	digitalcommons.fiu.edu Internet Source	<1 %
20	mp.jvolsu.com Internet Source	<1 %
21	Parkavi A., Kaushiki Shaha, Purva Rajodiya, Samruddha S. "Advanced Agro Management Using Machine Learning and IoT", 2023 IEEE North Karnataka Subsection Flagship International Conference (NKCon), 2023 Publication	<1 %
22	ia902500.us.archive.org Internet Source	<1 %

23	towardsdatascience.com Internet Source	<1 %
24	www.mdpi.com Internet Source	<1 %
25	arcabc.ca Internet Source	<1 %
26	Dr. Sudha K, T.P. Anish, C Balakrishnan, Dr. Srikanth Lakumarapu, Dr. P.J. Beslin Pajila, R Siva Subramanian. "Leveraging Machine Learning for Customer Intelligence: An Experimental Analysis Learning Classifiers", Procedia Computer Science, 2023 Publication	<1 %
27	Maya Astriyani, Wiga Maulana Baihaqi, Chyntia Raras Ajeng Widiawati. "Development of Chatbot Features for Stunting Education Using Artificial Neural Network Algorithm", 2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2023 Publication	<1 %
28	Ton Duc Thang University Publication	<1 %

Exclude quotes	Off	Exclude matches	Off

Exclude bibliography Off