International Journal of Advance Research in Science and Engineering
Volume No. 11, Issue No. 05, May 2022
www.ijarse.com
IJARSE
ISSN 2319 - 8354

LOAN DEFAULT FORECASTING USING MACHINE LEARNING

P.Nagendra[1] , G.Sowmya[2] , T.Vamsi Kumar[3] , N.Veera Charan[4] , M.Thanusha[5]

[1]*Assistant Professor ,Dept. of Artificial Intelligence & Data Science , AITS, Rajampet, AP*

[2,3,4,5]*UG Scholars ,Dept. of Computer Science and Engineering AITS, Rajampet, AP*

ABSTRACT

Default appraisal, also known as loan appraisal, is a critical procedure, and banks should assist them in determining whether or not the potential borrower is just a defaulter at such a later stage thus that they might process their application and choose whether or not it will approve it. The conclusions drawn from such calculations will assist banks and insurance companies in reducing their losses and, as a response, increasing the amount of credits available. As a result, it's critical to develop a model that considers the many features of an applicant as well as the results in connection to a relevant applicant. In today's technology-based world, criminals use every means available to borrow the money from their unlawful activities.The rising amount of bad debts resulting from commercial bank loans illustrates the economy's difficulty in problematic banks. We used machine learning and data mining techniques to assess defaulters using a dataset of home loan application information, allowing banks to make good decisions.

Keywords :Loan, credit, prediction, and Data Mining are some of the terms used in this paper.

I.INTRODUCTION

Financial fraud has been recorded on a regular basis in India in recent years. Banking crimes have increased in frequency, complexity, diversity, and cost tremendously as compared to previous methods. As a result, regulators are quite concerned about such risks. The strength and sustainability of a national budget framework aid in determining whether the industry was worth spending in. It provides information on the citizens' health, safety, and living standards. As a result, if the financial system is plagued by high rates of Non Performing Assets, it is a serious problem, as it reflects the financial distress of borrower consumers. Such issues have a significant impact on the Indian economy. The key causes of the increase in stressed assets have been identified as aggressive lending practises, purposeful and conscious defaults, loan fraud cases in some situations, and economic stagnation. According to statistical inputs, 16 out of 60 banks (or 26.5 percentage of the market) are still unable to cover their predicted losses within their current framework. People apply for the loan in big numbers every day for a number of reasons. However, not all of the applications are genuine, and not all of them may be given credit. Analyzing the risk connected with the complainant's demographic data is quite significant.

II.RELATED WORK

[1]. Aditi Kacheria, Nitin Shivakumar, Archana Gupta, and Shreya Sawker [1] developed a model for loan approval authorities that would assist them in judging the authenticity of consumers who had filed for the loan, hence boosting the likelihood of their debts being paid back on time. Their strategy is made up of three parts: K-NN and Binning are used to perform pre-processing. The Nave Probabilistic technique is used to calculate

whether or not to grant a loan to that of a customer. Database Update: Newly discovered data is uploaded to the database for future results.

[2].Archana Gahlaut, Tushar, and Prince Kumar Singh investigated whether data mining techniques can be used to predict and classify a client's credit score (good/bad) in order to lessen the risk of future loan defaults. To create prediction models, algorithms such as Tree Structure, Regression Analysis, Svm Classifiers, Neural Network, Adaptive Booster Model, and Random Forest are employed, with the outcomes of each process shown in graphs. Random Forest was found to be the most promising approach for building a better classification model, with the best accuracy [2].

[3]. Aboobyda Jafar Hamid and Tarig Mohammed Hussain proposed a solution for determining whether or not to offer a loan to a client. Three separate models were created using three different categorization techniques. These algorithms were built using WEKA and therefore are bayesNet, j48, as naiveBayes. The J48 method was determined to be the best based on the findings of these classification techniques since it provides low average error percentage and good accuracy.

[4]. Mrunal Surve, Priya Singh, Sandip Pandit, Pooja Thitme, and Swati Sonawane focused their effort on identifying and analysing the risk associated with commercial bank lending. They employed data mining techniques to assess the risk of lending money. It entails gathering, analysing, and analyzing data from multiple sources in order to compile useful information [4]. They employed the C4.5 classification algorithm to forecast a person's risk percentage when it came to lending money.

[5]. Using only a simple majority vote approach, Puvvala Ravikumar as well as Vadlamani Ravi created a set of ensemble classifiers. They used SVM, ANFIS, Logistic RBF, Semionline RBF1, Mathematically equivalent RBF, and Semionline RBF2, MLP as component of the ensemble, and created the ensembles by selecting 2, 3, 4, 5, Six classifications at a time from across all 7 classifiers [5].

[6]. Shiju Sathyadevan as well as Surya Gangadharan used a criminal law and computer programming perspective to create a data mining process that really can help solve crime faster. They have concentrated on criminal aspects that occur on a daily basis rather than on the causes of crime. They employed the Nave Bayes algorithm for categorization, and the Apriori algorithm to find frequently occurring criminal patterns. The decision tree idea is used to make predictions.

[7]. Dr. S. Prakasam and K. Chitra Lekha presented a thorough study on data mining techniques and the use of such systems in detecting cyber-crime in real-time applications. It demonstrates how data mining tools aid fraud detection in a variety of industries, including e-commerce, insurance, and health. It also explains how Clustering techniques can be used to detect cyber-crime in the banking sector. This approach data is divided into related clusters, allowing patterns and ordering to be detected. To represent the probability distribution, a Gaussian mixed replica is used. The output of the this novelty filter is then used to make the detection decision [7].

[8]. In their research, K. Chitra Lekha as well as Dr. S. Prakasam use the J48 classification, K-Means clustering approach, and Influenced Experiential classifier to create a model for cyber-crime prediction. The proposed model performs better in terms of forecasting cyber-crime in financial industries. The Influenced Associative Classifier is a well-organized strategy to use the classification approach with Association Rule Mining to improve classification prediction accuracy. The use of K-Means with the J48 approach and the Influenced Association Classifier results in a superior prediction of cyber-crime risks in financial sectors [8].[9]. Jin et al.

employed data mining to predict the risk of a loan application and compared three data mining algorithms: support vector machines, decision trees, & neural networks. They also employed a 10-fold cross-validation procedure in conjunction with a relatively high percent hit ratio to demonstrate the correct prediction. A cumulative lift coefficient analysis was used to assess the quality. The best results were achieved with SVM [9].

III.FIGURES AND TABLES

In this section, we addressed the numerous attributes that influence the outcome, or, to put it another way, the primary variables that influence the default response from the target attribute. We looked at the behaviour of each of our attributes to see if they had an impact on a particular attribute. We also discovered how many attributes the target attribute is dependent on. A heatmap was used to analyse everything. The correlation matrix is represented visually as a heatmap.It aids in the speedy identification and verification of relationships between columns.

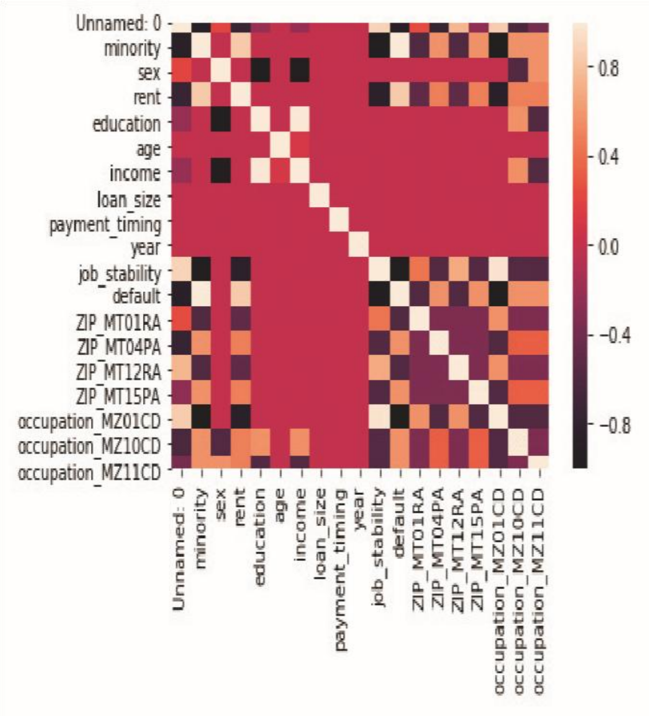


Figure 1: Heatmap of Correlation

The preceding graph aids in the identification of all the qualities that have a positively and negatively impact on target attribute, default. Below are the qualities that have a favourable impact just on default attribute.

TABLE I.MAJOR POSITIVE ATTRIBUTES AFFECTING DEFAULT

Sr. No.	Attribute	Description
1	Minority	Whether a person belongs to a minority group or not
2	Rent	States whether a person pays a rent or no

Also described below are the attributes which have a detrimental impact just on default attribute.

TABLE II.MAJOR NEGATIVE ATTRIBUTES AFFECTING DEFAULT

Sr. No.	Attribute	Description
1	job_stability	Describes whether a person has a stable job or no
2	occupation	Describes category of a person’s job

The following are the traits that are highly connected with one another. These characteristics are interdependent in a beneficial way.

TABLE III.CORRELATED ATTRIBUTES

Sr. No.	Attribute	Description
1	Income	Describes the total annual income of a person
2	Education	Describes the education of a person

A correlation matrix was also used to examine the relationship between income and education. Outliers were also looked for in the scatter plot.

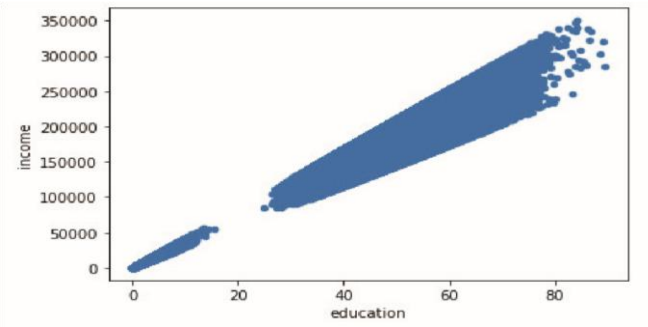
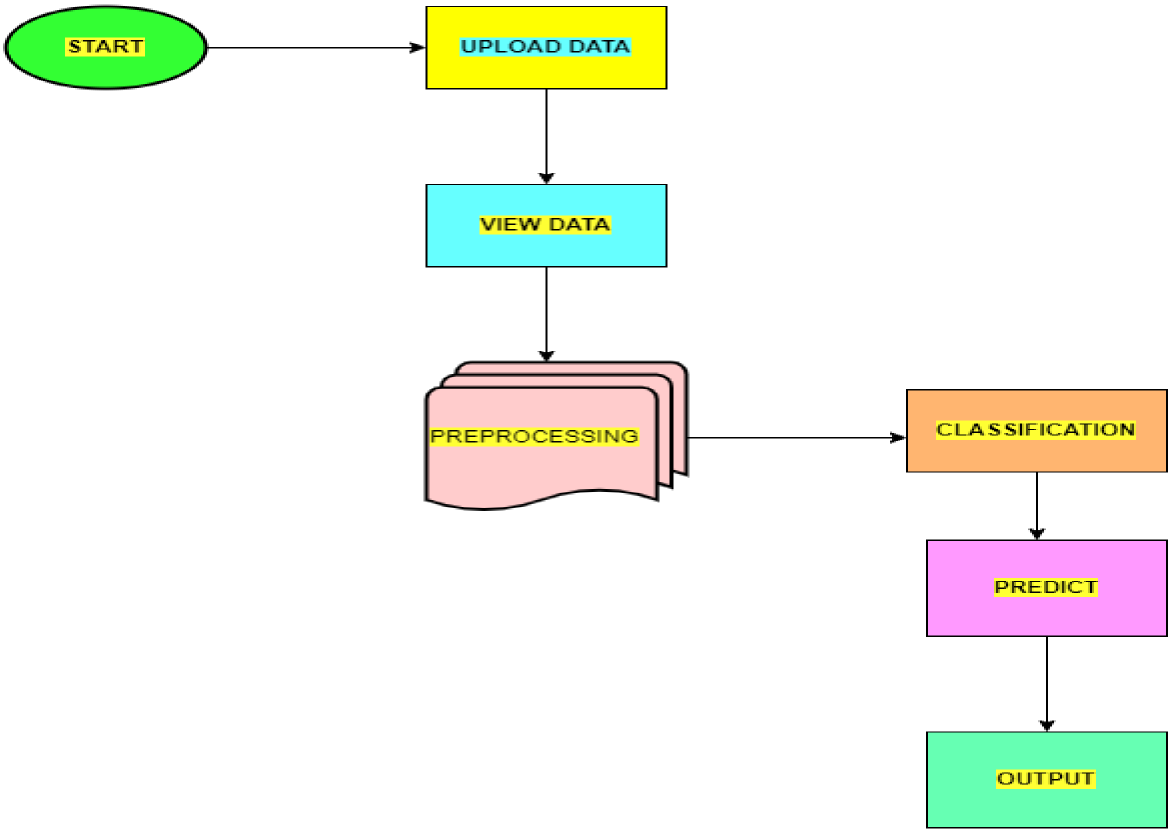


Figure 2 shows a scatter graph of income by educational attainment.

Both the qualities, wealth and education, are directly dependant on each other, as shown in the graph above. There are also a few outliers in the sample, with values that are greater than the average for each of two factors.

IV.ARCHITECTURE



V.METHODOLOGY AND ALGORITHMS

1. Upload as well as View Data:

Create a CSV file dataset, then upload and view the loan dataset.

2. Pre-processing:

Delete any files that are empty or duplicated.

3. **Normalize:**

All files should be normalised.

4. Logistic Regression:

Logistic regression is just a statistical model uses the logistic function to represent a binary variables in the most basic form, however there are many more advanced extensions.

5. Random Forest:

Random forests, also known as random decision forests, are just an unsupervised learning, regression, and other tasks that works by training a large number of decision trees and then outputting a class that is the method of the classes (categorization) or the mean/average prediction.

6.Gradient boosting :

It is just a machine learning method for classification and regression problems that generates a prediction model is a representation of an ensemble with weak prediction models, often decision trees.

7.Graph:

The graph's points frequently represent a relationship among two or more elements.

ALGORITHMS

This section defines all of the different data mining models that were employed, as well as the results that we obtained when applying each of them. For each model, the accuracy value obtained is also shown. The following is a list of the algorithms that were used:

a. Logistic Regression

b. . Gradient Boosting

c. CatBoost Classifier

d.Random Forest

1) Logistic Regression: Regression Analysis is a categorical classification technique that assigns observations to one of several classes. It is used to classify data points into binary categories. Categorical classification means that an output can be classified into one of two groups (1 or 0). By altering its output with the logistic sigmoid function, logistic regression provides a probability value.

2) Gradient Boosting: Gradient Boosting is a machine learning strategy for classification and regression issues. As in form of an ensemble many weak prediction models, it creates a model for prediction. The low classification algorithm is performed to changed versions of the data in such a sequential way in the boosting algorithm, resulting in a sequence of weak classifiers. Gradient Boosting is a type of ensemble learning that combines numerous weak decision trees to produce a powerful classifier. These decision trees are combined to produce a powerful gradient boosting model.

3) CatBoost Classifier: Boosting is just a mix of the words "Category" and "Boosting," and it is an unbiased boosting using categorical features. CatBoost's library works well with a variety of data types, including text, audio, image, and historical data. It uses multiple statistical methods to automatically handle categorical values. It's a gradient boosting approach for decision trees. Catboost provides two important algorithmic progress: order boosting and a category feature processing approach. To avoid the prediction shift caused by a specific sort of target leakage found across all gradient boosting technique implementations, each of these strategies employs random permutations of both the training dataset.

4) Random Forest: Rf is a set of basic tree predictors that can respond to a collection of predictor values when presented with a list of them. This looks to be a class membership in the classification issue, as it associates, as well as classifies, a set of independent predictor values with one of the categories contained in the dependent variable. The tree result in regression would be an estimate of both the dependent variable based on the predictors.

VI.RESULT

This section compares and contrasts all of the models that have been made. Accuracy, precision, and the f1-score are used to evaluate these models.

MODEL PERFORMANCE EVALUATION

Sr. No	Models Applied (Algorithms)	Accuracy	Precision	F1-score

1	Regression Logistic	0.14963	0.49	0.00
2	Gradient Boosting	0.84035	0.85	0.91
3	CatBoost Classifier	0.84045	0.85	0.91
4	Random Forest	0.83514	0.86	0.89

All values obtained for such various metrics again from various models are represented in the table above. We chose to quantify the model's performance using accuracy because the f1score as well as precision values of almost all the models are identical, with the exception of the logistic regression model. It demonstrates that Logistic Regression is just less accurate than other models. Random Forest's accuracy is also lower than that of Gradient Boosting as well as CatBoost Classifier. As a result, we may conclude that Gradient Boosting as well as CatBoost Classifier are effective in predicting our dataset.

VII.CONCLUSION

Various algorithms were used to forecast loan defaulters in this article. Logistic Regression, Random Forests, Gradient Boosting, and CatBoost Classifier were used to achieve the best results. In comparison to Logistic Regression, the Gradient Boosting technique produces better or equal results. Multiple models are used in the gradient boosting procedure. It's unusual to throw out a variable because the variables' interpretations aren't always clear. On the other hand, even if it reduces overall accuracy of the model and prediction power, removing variables during fitting logistic regression is a common technique. CatBoost converts categorical values to numerical values using a range of statistics based on categorical features and numerical and categorical attributes.With respect to the given dataset, the CatBoost classifier as well as Gradient Boost provide nearly comparable accuracy. Furthermore, these models could be used to make better judgements about loan applications, potentially saving a financial institution from massive losses.

REFERENCES

Kacheria, A., Shivakumar, N., Sawkar, S. and Gupta, A. (2016). Loan Sanctioning Prediction System. [online] Ijsce.org.

[1] http://www.ijsce.org/wpcontent/uploads/papers/v6i4/D2904096416.p df

[2] A. Gahlaut, Tushar, and P. K. Singh, "*Prediction analysis of risky credit using Data mining classification models*," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT),2017.

[3] Hamid, Aboobyda & Ahmed, Tarig. (2016). "*Developing Prediction Model of Loan Risk in Banks Using Data Mining*". Machine Learning and Applications: An International Journal. 3. 1-9. 10.5121/mlaij.2016.3101.

[4] Mrunal Surve, Pooja Thitme, Priya Shinde, Swati Sonawane, and Sandip Pandit. "*Data mining techniques to analyze risk giving loan(bank)*" Internation Journal Of Advance Research And Innovative Ideas In Education Volume 2 Issue 1 2016 Page 485-490

[5] P. Ravikumar and V. Ravi, "*Bankruptcy Prediction in Banks by an Ensemble Classifier*," 2006 IEEE International Conference on IndustrialTechnology, Mumbai, 2006, pp. 2032-2036.

[6] S. Sathyadevan, D. M. S and S. G. S., "*Crime analysis and prediction using data mining*," 2014 First International Conference on Networks & Soft Computing (ICNSC2014), Guntur, 2014, pp. 406-412

[7] Lekha, K. and Prakasam, D. (2018). https://www.researchgate.net/publication/326147494

[8] K. C. Lekha and S. Prakasam, "*Data mining techniques in detecting and predicting cyber crimes in banking sector*," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS),Chennai, 2017, pp. 1639-1643.

[9] 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.

[10] Jannes Klaas, "*Loan Default Model Trap*,"

[11] https://www.kaggle.com/jannesklaas/model-trap

[12] Kavitha, K. (n.d.). "Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6(2), pp. 162–166, 2016.

[13] Somayyeh, Z., & M. Abdolkarim. (n.d.). "Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining A Case Study: Branches of Mellat Bank of Iran", Jurnal UMP Social Sciences and Technology Management, vol. 3(2), pp. 307–316, 2015.

[14] M. Sudhakar, & C.V.K. Reddy. (n.d.). "Two Step Credit Risk Assessment Model For Retail Bank Loan Applications Using Decision Tree Data Mining Technique.