2020 International Conference for Emerging Technology (INCET)
Belgaum, India. Jun 5-7, 2020

Loan Default Forecasting using Data Mining

Bhoomi Patel[1], Harshal Patil[2], Jovita Hembram[3], Shree Jaswal[4]
[1,2,3,4]Department of Information Technology, St. Francis Institute of Technology, Mumbai, India
[1]bhoomi16patel98@gmail.com, [2]harshalpatil31598@gmail.com, [3]jovita.hembram04@gmail.com, [4]shreejaswal@sfit.ac.in

***Abstract*—Estimation or assessment of default on a debt is a crucial process that should be carried out by banks to help them to assess if a loan applicant can be a defaulter at a later phase so that they process the application and decide whether to approve the loan or not. The conclusion derived from such assessments helps banks and other financial institutions to lessen their losses and eventually increase the number of credits. Hence, it becomes vital to construct a model that will take into account the different aspects of an applicant and derive a result regarding the concerned applicant. All available means to loan the money from their illicit activities are used for criminal activities in today's technology-based realm. The increasing number of bad debts resulting from commercial banks' loans reflects the growing problem of distraught banks within the economic system. We have used data mining algorithms to predict the likely defaulters from a dataset that contains information about home loan applications, thereby helping the banks for making better decisions in the future.**

Keywords—loan, credit, prediction, data mining

I. INTRODUCTION

In the recent past, occurrences of financial fraud have perpetually been reported in India. As compared to traditional ways the frequency, intricacy, diversity, and price of banking frauds have increased exponentially. Consequently, such issues are a grave cause of concern for regulators. The robustness and stability of a country's fiscal structure help ascertain whether the country's economy is worth investing in. It gives an idea about the wellness, security and living standards of its citizens. Thus, if the banking system is troubled with high levels of Non-Performing Assets then it is a critical issue, as it mirrors the financial difficulty of borrower clients. Indian economy undergoes to a great extent from such problems. It has been observed that the primary reasons for the spurt in stressed assets are typically aggressive lending practices, willful and conscious default/loan frauds/corruption in some cases, and economic retardation. Statistical inputs show that 16 out of 60 banks (26.5 percent market share) were unable to cover their expected losses from their current framework. Everyday people apply for loans in large numbers, for a variety of purposes. But not all the applicants are legitimate, and not everyone can be credited. It is extremely important to analyze the risk associated, by reviewing the applicant's demographic data.

Section 2 contains literature survey of related work. Section 3 describes the loan dataset used. Section 4 contains a brief description of data mining algorithms used. Section 5 and 6 contains result and conclusion respectively.

II. LITERATURE SURVEY

Aditi Kacheria, Nidhi Shivakumar, Archana Gupta and Shreya Sawker [1], proposed a model for the loan approval authorities which would help them judge the credibility of customers who have applied for the loan, hence increasing the chances of their loans being paid back in time. Their model comprises of three components:

a) Pre-processing-which is done using K-NN and Binning.

b) Classification - Naïve Bayes algorithm is used to decide whether to sanction loan to a customer.

c) Database Updation: The newly found data is added for further results.

Archana Gahlaut, Tushar, Prince Kumar Singh studied if data mining techniques are helpful for predicting and classifying the client's credit score (good/bad) to reduce the future risks in giving loans to clients who cannot repay. Algorithms like Decision Tree, Linear Regression, Support Vector Machine, Neural Network, Adaptive Boosting Model and Random Forest are used to build predictive models, and the results of each algorithm are represented in graphs. Random Forest proved to be the most promising algorithm with the highest accuracy for constructing a better classification model [2].

Aboobyda Jafar Hamid along with Tarig Mohammed Ahmed proposed a solution, to yield a decision whether to grant a loan to a client [3]. Three different models based on three classification algorithms were constructed. These algorithms are bayesNet, j48, and naiveBayes which were implemented using WEKA. Based on the results of these classification algorithms, J48 algorithm was concluded to be the finest because it produces low mean absolute error and high accuracy.

Mrunal Surve, Priya Shinde, Sandip Pandit, Pooja Thitme and Swati Sonawane in their project, the focus was to identify and analyze the risk in giving a loan of commercial banks. To analyze risk in giving loan they have used data mining techniques. It includes analyzing and processing data from various resources and summarize into valuable information [4]. They have used C4.5 classification algorithm for predicting the risk percentage for an individual to give loans.

Puvvala Ravikumar and Vadlamani Ravi, built a collection of ensemble classifiers using a simple majority voting scheme. As part of the ensemble, they made use of seven classifiers viz., SVM, ANFIS, Linear RBF, Semi-online RBF1, Orthogonal RBF and Semi-online RBF2, MLP. They designed the ensembles by taking 2, 3, 4, 5, 6 classifiers at a time from all 7 classifiers [5].

To produce a data mining procedure Shiju Sathyadevan and Surya Gangadharan. S [6], have used an approach between criminal justice and computer science that can help solve crimes faster. They have focused mainly on crime factors that occur each day, rather than focusing on causes of crime occurrence. For classification they used Naïve Bayes

978-1-7281-6221-8/20/$31.00 ©2020 IEEE

algorithm, frequently occurring crime patterns were detected using Apriori algorithm. For prediction, the decision tree concept is considered.

K. Chitra Lekha and Dr. S. Prakasam proposed a broad study on techniques of data mining and the responsibility of such systems on the detection of cyber-crimes in real-time applications. It explains how data mining techniques help multiple sectors like E-commerce, insurance, health for fraud detection. Further, it explains how cyber-crime can be detected in the banking sector using Clustering techniques. This technique helps to split the data into related clusters which helps patterns and orders become detectable. Here, a Gaussian mixture replica is applied to model the probability density function. The detection decision is then concluded from the output of this novelty filter [7].

K. Chitra Lekha and Dr. S. Prakasam in their paper produce the model for cyber-crime prediction with J48 classifier, K-Means clustering technique and Influenced Associative classifier. For predicting cyber-crime in banking sectors, the proposed model gives a better prediction outcome. Influenced Associative Classifier provides a well-organized way to utilize the classification method with Association Rule Mining, which enhances the prediction accuracy for classification. The implementation of K-Means with J48 technique and Influenced Association Classifier provides a better prediction outcome over the hazards of cyber-crime in banking sectors [8].

Jin et al. used data mining technique to forecast the risk associated with a loan application and also compared different data mining models: support vector machine, decision trees, and neural networks. They also used 10-fold cross-validation method in combination with the large value of average percent hit ratio to show the appropriate prediction. To evaluate the quality, cumulative lift curve analysis was performed. Best results were obtained using SVM [9].

III. Data Description

A. Data Source

We obtained home loan data set from Kaggle [10]. The dataset consists of various variables such as minority, sex, ZIP, rent, education, income, loan size, payment timing year, job stability and occupation.

B. Data Description

The dataset has 64000 tuples and 14 attributes. 1 out of the 14 attributes is the target attribute viz. default. The dataset is split into training data and testing data having shapes of (480000,14) and (160000,14) respectively.

IV. Experiments and Results

A. Major Attributes

In this section, we discussed the various attributes that affects the result, or in other words, the major variables that have an impact on the behaviour of the target attribute i.e. default. We analysed each of our attribute's behaviour and whether they have an impact on the target attribute. We also found the number of attributes on which the target attribute depends on. All of this was analysed using a heatmap. A heatmap is a visual representation of the correlation matrix. It helps to quickly identify and check correlations amongst the columns.

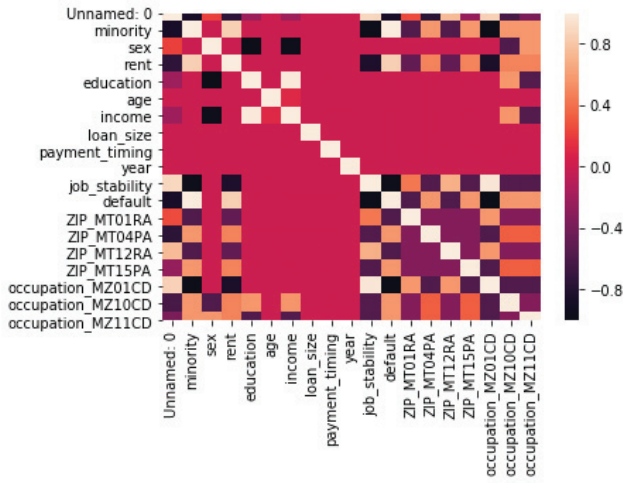

Fig. 1. Correlation Heatmap

The above plot helps to identify all the attributes that have a positive and a negative impact on the target attribute i.e. default. The attributes that have a positive impact on the default attribute are described below.

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

Further, the attributes that have a negative impact on the default attribute are described below.

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

Also, the attributes that are highly correlated to each other are as follows. These attributes are positively dependent on each other.

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

Further, a scatter plot was plotted to check the dependencies of income on education. Also, the scatter plot was analysed for any outliers.

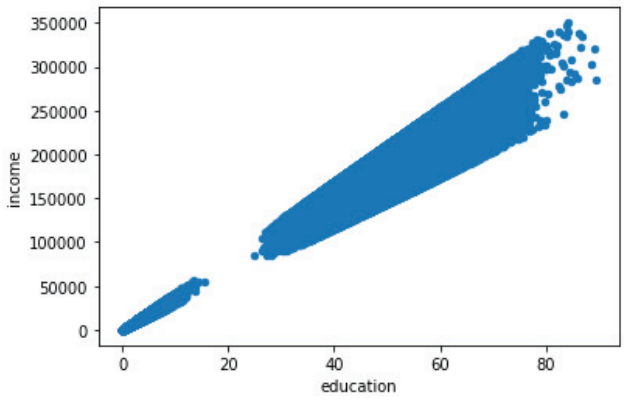

Fig. 2. Scatter Plot of income by education

The above plot shows that both the attributes, income and education are linearly dependent on each other. Also, a few outliers are present in the dataset which have values higher than the normal for each of the two attributes.

B. Models Used

All the data mining classification models used are defined in this portion and the results which we got after applying each of them are listed sequentially. The accuracy value achieved for each model is also presented. The list of algorithms used are mentioned below:

a. Logistic Regression
b. Gradient Boosting
c. CatBoost Classifier
d. Random Forest

1) Logistic Regression: Logistic Regression performs categorical classification by allocating observations to a distinct set of classes. It can be used for the binary classification of data-points. Categorical classification is such that an output belongs to either of the two classes (1 or 0). Logistic Regression returns a probability value by modifying its output using the logistic sigmoid function.

*2) Gradient Boosting:*A machine learning approach for classification as well as regression problems is Gradient Boosting. It constructs a model for prediction in the form of an ensemble of weak prediction models. In boosting algorithm the weak classification algorithm is repeatedly applied to modified versions of the data in a sequential manner, thereby generating a series of weak classifiers. Gradient Boosting is an ensemble learning method that is a compilation of several weak decision trees which results in a powerful classifier. These decision trees combine together to form a strong model of gradient boosting.

3) CatBoost Classifier: CatBoost is unbiased boosting with categorical features and is a combination of words "Category" and "Boosting". The library of CatBoost works fine with various categories of data, such as text, audio, image along with historical data. It handles categorical values automatically using various statistical methods. It is an algorithm for gradient boosting on decision trees. Catboost introduces two critical algorithmic advances: ordered boosting and an algorithm to process categorical features. Each of these techniques is using random permutations of the training dataset to overcome the prediction shift because of a special type of target leakage which is present in all gradient boosting algorithm implementations.

4) Random Forest: Random Forest comprises of ensemble of simple tree predictors which are capable of giving a response when given a set of predictor values. For classification problem, this result appears to be a class membership, which associates, or classifies, a set of independent predictor value with one of the categories present in the dependent variable. Whereas in Regression, the tree result is an estimate of the dependent variable given the predictors.

V. Results

This section shows a comparative study of all the models that were built. These models are evaluated through accuracy, precision, and f1-score.

Performance Evaluation of Models

Sr. No.	Models Used (Algorithms)	Accuracy	Precision	F1-score
1	Logistic Regression	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	0.91

Above table is representing the values obtained for the various metrics from the different models. Since the f1-score and precision values of most of the models are similar excluding logistic regression model, we choose to measure the performance of the model using accuracy. It shows that the accuracy of Logistic Regression is less than other models. Also, the accuracy of Random Forest is comparatively low than Gradient Boosting and CatBoost Classifier. Therefore, we can infer that Gradient Boosting and CatBoost Classifier is doing prediction well for our dataset.

VI. Conclusion

In this paper, various algorithms were implemented to predict loan defaulters. Optimum results were obtained using Logistic Regression, Random Forest, Gradient Boosting and CatBoost Classifier. Gradient Boosting process gives better or equivalent results in contrast with Logistic Regression. Gradient boosting is a process consisting of multiple models. It is unusual to discard a variable as the interpretation of the variables is not straight. On the other hand, it is an accepted practice to eliminate variables while fitting logistic regression, even if it minimizes the overall model accuracy and prediction power. CatBoost alters categorical values into numbers using a variety of statistics on combinations of categorical features and combinations of numerical and

categorical features. CatBoost classifier and Gradient Boost provide almost equal accuracy with respect to the referred dataset. Further, these models can be used to make better decisions on loan applicants and save any financial institution from undergoing huge losses.

VII. Future Work

Here in this paper, we have only considered home loan prediction, a system could be made for predicting defaulters of other loans as well.

Also, whether the non-defaulter would turn out to be a fraudster or not could be predicted.

References

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

[2] http://www.ijsce.org/wpcontent/uploads/papers/v6i4/D2904096416.pdf

[3] 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.

[4] 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.

[5] 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

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

[7] 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

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

[9] 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.

[10] 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.

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

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