Bank Loan Approval Prediction Using Data Science Technique (ML)

Authors: Ahmed Rashid , Adwait Babu, Abhishek Yadav , Ajith R

Abstract

Banks are making major part of profits through loans. Loan approval is a very important process for banking organizations. It is very difficult to predict the possibility of payment of loan by the customers because there is an increasing rate of loan defaults and the banking authorities are finding it more difficult to correctly access loan requests and tackle the risks of people defaulting on loans. In the recent years, many researchers have worked on prediction of loan approval systems. Machine learning technique is very useful in predicting outcomes for large amount of data. In this paper, four algorithms are used such as Random Forest algorithm, Decision Tree algorithm, Naive Bayes algorithm, Logistic Regression algorithm to predict the loan approval of customers. All the four algorithms are going to be used on the same dataset and going to find the algorithm with maximum accuracy to deploy the model. Henceforth, we develop bank loan prediction system using machine learning techniques, so that the system automatically selects the eligible candidates to approve the loan.

Introduction

I. INTRODUCTION

A loan is the major source of income for the banking sector of financial risk for banks. Large portions of a bank’s assets directly come from the interest earned on loans given. The activity of lending loans carry great risks including the inability of borrower to pay back the loan by the stipulated time. It is referred as “credit risk”. A candidate’s worthiness for loan approval or rejection was based on a numerical score called ”credit score”. Therefore, the goal of this paper is to discuss the application of different Machine Learning approach which accurately identifies whom to lend loan to and help banks identify the loan defaulters for much-reduced credit risk.

II. LITERATURE SURVEY

A recent development of machine learning techniques and data mining has led to an interest of implementing these techniques in various fields [17]. The banking sector is no exclusion and the increasing requirements towards financial institutions to have robust risk management has led to an interest of developing current methods of risk estimation. Potentially, the implementation of machine learning techniques could lead to better quantification of the financial risks that banks are exposed to. Within the credit risk area, there has been a continuous development of the Basel accords, which provides frameworks for supervisory standards and risk management techniques as a guideline for banks to manage and quantify their risks. From Basel II, two approaches are presented for quantifying the minimum capital requirement such as the standardized approach and the internal ratings based approach (IRB) [16]. There are different risk measures banks consider in order to estimate the potential loss they may carry in future. One of these measures is the expected loss (EL) a bank would carry in case of a defaulted customer. One of the components involved in ELestimation is the probability if a certain customer will default or not. Customers in default means that they did not meet their contractual obligations and potentially might not be able to repay their loans [18]. Thus, there is an interest of acquiring a model that can predict defaulted customers. A technique that is widely used for estimating the probability of client default is Logistic Regression [19]. In this thesis, a set of machine learning methods will be investigated and studied in order to test if they can challenge the traditionally applied techniques. A prediction is a statement about what someone thinks will happen in the future. People make predictions all the time. Some are very serious and are based on scientific calculations, but many are just guesses. Prediction helps us in many things to guess what will happen after some time or after a year or after ten years. Predictive analytics is a branch of advanced analytics that uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions. “Adyan Nur Alfiyatin, Hilman Taufiq [14] and their friends work on the house price prediction. They use regression analysis and Particle Swarm Optimization (PSO) to predict house price”. One other similar work on the Mohamed El Mohadab, Belaid Bouikhalene [15] and Said Safi to predict the rank for scientific research paper using supervised learning. Kumar Arun, Garg Ishan and Kaur Sanmeet [13] work on bank loan prediction on how to bank approve a loan. They proposed a model with the help of SVM and Neural networks like machine learning algorithms. This literature review helps us carry out our work and propose a reliable bank loan prediction model. Manjeet et al (2018) [24] there are seven types of variables that may influence consumer loan default; consumer’s annual income, debt-income ratio, occupation, home ownership, work duration and whether or not consumer possesses IJARSCT ISSN (Online) 2581-9429 International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) Volume 5, Issue 1, May 2021 Copyright to IJARSCT DOI: 10.48175/IJARSCT-1168 451 www.ijarsct.co.in Impact Factor: 4.819 a saving/checking account. In a work by Steenackers [26] and Goovaerts, the key factors that may influence loan default are borrower’s age, location, resident/work duration, owner of phone, monthly income, loan duration, whether or not applicant works in a public sector, house ownership and loan numbers. Another study by Ali Bangher pour [27] on a large dataset within the period of 2001-2006 indicated that loan age was the most important factor when predicting loan default while market loan-to-value was the most effective factor for mortgage loan applications. In addition to identifying factors that may influence loaned fault, there is also a need to build robust and effective machine learning models that can help capture important patterns in credit data. The choice of model so great importance as the chosen model plays a crucial role in determining accuracy, precision and efficiency of a prediction system. Numerous models have been used for loan default prediction and although there is no one optimal model, some models definitely do better than others. In 2019, Vimala and Sharmili [1] proposed a loan prediction model using and Support Vector Machines(SVM)methods. Naïve Bayes, an independent speculation approach, encompasses probability theory regarding the data classification. On the other hand, SVM uses statistical learning model for classification of predictions. Dataset from UCI repository with 21 attributes was adopted to evaluate the proposed method. Experimentations concluded that, rather than individual performances of classifiers (NB and SVM), the integration of NB and SVM resulted in an efficient classification of loan prediction. In 2019, Jency, Sumathi and Shiva Sri [2] proposed an Exploratory Data Analysis(EDA) regarding the loan prediction procedure based on the client's nature and their requirements. The major factors concentrated during the data analysis were annual income versus loan purpose, customer 's trust, loan tenure versus delinquent months, loan tenure versus credit category, loan tenure versus number of years in the current job, and chances for loan repayment versus the house ownership. Finally, the outcome of the present work was to infer the constraints on the customer who are applying for the loan followed by the prediction regarding the repayment. Further, results showed that, the customers were interested more on availing short-tenure loans rather than long-tenure loans. In 2019, Supriya, Pavani, Saisushma, Vimala Kumari and Vikas [3] presented a ML based loan prediction model. Themodulesin the present approach were data collection and pre-processing, applying the ML models, training followed by testing the data. During the pre-processing stage, the detection and removal of outliers and imputation removal processing were carried out. In the present method, SVM, DT, KNN and gradient boosting models were employed to predict the possibilities of current status regarding the loan approval process. The conventional 80:20 rule was adopted to split the dataset into training and testing processes. Experimentation concluded that, DT has significantly higher loan prediction accuracy than the other models. In 2017, Goyal and Kaur [4] presented a loan prediction model using several Machine Learning (ML) algorithms. The dataset with features, namely, gender, marital status, education, number of dependents, employment status, income, co applicant's income, loan amount, loan tenure, credit history, existing loan status, and property area, are used for determining the loan eligibility regarding the loan sanctioning process. Various ML models adopted in the present method includes, Linear model, Decision Tree (DT), Neural Network (NN), Random Forest (RF), SVM, Extreme learning machines, Model tree, Multivariate Adaptive Regression Splines, Bagged Cart Model, NB and TGA. When evaluated these models using Environment in five runs, TGA resulted in better loan forecasting performance than the other methods. In 2016, Aboobyda Jafar Hamid and Tarig Mohammed Ahmed [5] presented a loan risk prediction model based on the data mining techniques, such as Decision Tree (J48), Naïve Bayes (NB) and BayseNet approaches. The procedure followed was training set preparation, building the model, Applying the model and finally. Evaluating the accuracy. This approach was implemented using Weka Tool and considered a dataset with eight attributes, namely, gender, job, age, credit amount, credit history, purpose, housing, and class. Evaluating these models on the dataset, experimental results concluded that, J48 based loan prediction approach resulted in better accuracy than the other methods. In 2016, Kacheria, Shivakumar, Sawkar and Gupta [6] suggested a loan sanctioning prediction procedure based on NB approach integrated with K-Nearest Neighbor (KNN) and binning algorithms. The seven parameters considered were income, age, profession, existing loan with its tenure, amount and approval status. The sub-processes include, Preprocessing (handling the missing values with KNN and data refinement using binning algorithm), Classification using IJARSCT ISSN (Online) 2581-9429 International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) Volume 5, Issue 1, May 2021 Copyright to IJARSCT DOI: 10.48175/IJARSCT-1168 452 www.ijarsct.co.in Impact Factor: 4.819 NB approach and Updating the dataset frequently results in appropriate improvement in the loan prediction process. Experimentation put-forth the conclusion that, integration of KNN and binning algorithm with NB resulted in improved prediction of loan sanctioning process. In 2016, Goyal and Kaur [7] suggested an ensemble technique based loan prediction procedure for the customers. The sub processes in the present method includes, data collection, filtering the data, feature extraction, applying the model, and finally analysis the results. The various loan prediction procedures implemented in the present method were Random Forest (RF), SVM and Tree model with Genetic Algorithm (TGA). The parameters considered for evaluating the models were accuracy, Gini Coefficient, Area Under Curve (AUC), Receiver Operating Curve (ROC), Kolmogorov - Smirnov (KS) Chart, Minimum Cost - Weighted Error Rate, Minimum Error Rate, and K-Fold Cross Validation parameters. Experimentation outcome concluded that the integration of three methods (RF, SVM and TGA) resulted in improved loan - prediction results rather than individual method ‘s prediction. In 2006, Sudhamathy [8] suggested a risk analysis method in sanctioning a loan for the customers using R package. The various modules include data selection, pre-processing, feature extraction and selection, building the model, prediction followed by the evaluation. The dataset used for evaluation in this method was adopted from UCI repository. To fine tune the prediction accuracy, the pre-processing operation includes the following sub-processes: detection, ranking and removal of outliers, removal of imputation, and balancing of dataset by proportional bifurcation regarding testing and training process. Further, feature selection process improves the prediction accuracy. When evaluated, the DT model resulted in 94.3% prediction accuracy. The process of analyzing data from different perspectives and extracting useful knowledge from it. Tithe core of knowledge discovery process. The various steps involved in extracting knowledge from raw data. Different data mining techniques include classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression etc. Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to classification technique. This approach frequently employs Decision tree based classification Algorithm. In classification, a training set is used to build the model as the classifier which can classify the data items into its appropriate classes. A test set is used to validate the model.

III. EXISTING SYSTEM

Anomaly detection relies on individuals’ behaviour profiling and works by detecting any deviation from the norm. When it is used for online banking fraud detection, it suffers from three disadvantages. First, for an individual, the historical behaviour data are often too limited for profiling his/her behaviour pattern. Second, because of the heterogeneous nature of transaction data, there is no uniform treatment to various attribute values, which will become a potential barrier for development of the model and for further usage. Third, the transaction data are highly skewed, and it becomes a challenge for utilizing the label information effectively. Anomaly detection often suffers from poor generalization ability and a very high false alarm rate. We argue that individuals’ limited historical data for behaviour profiling and fraud data’s highly skewed nature could account for this defect. Since it is straightforward to use information from other similar individuals, similarity measurement itself becomes a great challenge due to heterogeneous nature of attribute values.

A. Disadvantages

1. They had proposed a mathematical model and machine learning algorithms were not used.
2. Class Imbalance problem was not addressed and the proper measure were not taken.

B. Proposed System

In our proposed system, we combine datasets from different sources to form a generalized dataset and use four machine learning algorithms such as Random forest, Logistic regression, Decision tree and Naive bayes algorithm on the same dataset .The dataset we collected for predicting given data is split into training set and test set in the ratio of 7:3. The data model which was created using Machine learning algorithms are applied on training set and based on maximum test result from the four algorithms, the test set prediction is done using the algorithm that has maximum performance. After that, we deploy the model using Flask Framework.

C. Advantages

1. Performance and accuracy of the algorithms can be calculated and compared.
2. Class imbalance can be dealt with machine learning approaches.

Conclusion

The analysis starts from data cleaning and processing missing value, exploratory analysis and finally model building and evaluation of the model. The best accuracy on public test set is when we get higher accuracy score and other performance metrics which will be found out. This paper can help to predict the approval of bank loan or not for a candidate.

References

[1] Arun Kumar, Ishan Garg, and Sanmeer Kaur, \"Loan Approval Prediction Using Machine Learning Approach,\" 2018. [2] K. Hanumantha Rao, G. Srinivas, A. Damodhar, and M. Vikas Krishna at International Journal of Computer Science and Telecommunications published an article titled \"Implementation of Anomaly Detection Technique Using Machine Learning Algorithms\" (Volume2, Issue3, June 2011). [3] G. Arutjothi and C. Senthamarai, \"Prediction of loan status in commercial banks using machine learning classifier,\" International Conference on Intelligent Sustainable Systems (ICISS), 2017. [4] \"AzureML based analysis and prediction of loan applicants creditworthy,\" by Alshouiliy K, Alghamdi A, and Agrawal D P I n 2020, Third International conference on information and computer technologies. [5] \"Developing prediction model of loan risk in banks using data mining Machine Learning and Applications,\" Hamid A J and Ahmed T M, 2016. [6] M. Li, A. Mickel, and S. Taylor \"Should this loan be approved or denied?\" published a paper in the Journal of Statistics Education in 2018. [7] A. Vinayagamoorthy, M. Somasundaram, and C. Sankar, \"Impact of Personal Loans Offered by Banks and Non-Banking Financial Companies in Coimbatore City,\" 2012. [8] M. Cary Collins, Ph.D., and Frank M. Guess, Ph.D., MIT\'s Information Quality Conference, 2000, \"Improving information quality in loan approval processes for fair lending and fair pricing.\" [9] Arun Kumar, Ishan Garg, and Sanmeet Kaur, \"Loan approval

prediction based on machine learning approach,\" National Conference on Recent Trends in Computer Science and Information Technology, 2016. [10] Sivasree M S and Rekha Sunny T, \"Loan Credibility Prediction System Using Decision Tree Algorithm,\" International Journal of Engineering Research & Technology (IJERT), Vol. 4 Issue 09, September-2015. [11] Ji?í Doležal, Ji?í Šnajdr, Jaroslav Belás, Zuzana Vincúrová, "Model of the loan process in the context of unrealized income and loss prevention", Journal of International Studies, Vol. 8, No 1, 2015, pp. 91-106. DOI: 10.14254/2071-8330.2015.