IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.047 ⁜ Vol. 10, Issue 5, May 2022

DOI: 10.17148/IJIREEICE.2022.10574

BANK LOAN CREDIT RISK ANALYSIS

Ranjitha J [1], Yamini A M[2], Navyashree N R[3],Vidya BM[4]

Assistant Professor, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India[1]

Student, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India[2,3,4]

Abstract: The previous decade has seen a significant increase in data collection, particularly in the financial sector. Banks are in fact, one of the most prolific generators of big data. No other organisation has gathered as much data as the bank on its clients. The importance of gathering and interpreting this data cannot be overstated characteristic for decision making, especially in the financial sector. One of the most significant and common decisions that banks has to make, is approval of loan. The problem is figuring out low to construct a effective, powerful, competent, and ethical exoloration of personal data, in addition to make loan applicant proposals more relevant and personalised. Machine learning is a promising option for dealing with these issues. As a, result several machine learning techniques have been presented in recent years to solve the loan approval problem. Artificial Intelligence has grown steadily in recent years as modern computers processing abilities and ability to learn on their own have improved. When a few parameters from the ether (dynamic) are measured, a large amount of data is collected. The amount of information available is just too enormous for humans to encode explicitly. Machines that learn this information can produce more precise results/predictions at specific times. The environment evolves over time. Machines that can adapt to their surroundings would eliminate the need for ongoing redesign. The banking system has been updated thanks to automated technologies, bots and computers. Because the amount of data generated over time is so large, automation tools and computer programmes are in high demand. We created an ML model of prediction using both classification and regression methods in this project. The formula approach is used to create a linear regression model from scratch. To fit The dataset, classification algorithms such as Support Vector Machine (SVM), Random Forest Classifier, and KNN algorithms are used. To understand the pattern of projected data, comparisons must be done throughout implementation. Regression procedures, such as linear regressions (built from scratch), will improve the assignments efficiency (categorical)

Keywords: Bank Loan Credit Risk, SVM, KNN, Random Forest , Machine learning.

I. INTRODUCTION

This chapter gives a quick review of the project, the problem-solving approach that was utilized, as well as the project's outcomes and future scope. The chance of a loss arising from a creditor's inability to return a loan or meet any other contractual commitments to the investor is known as credit risk. It usually refers to the possibility that a lender may not be able to collect the outstanding amount plus premium, resulting in a loss of revenue and higher collection expenses. Unwanted cash may be issued in order to produce additional funds to offset credit risk. Despite the fact that no one can foretell who will default on their responsibilities. There is a risk that when lenders or banks issue mortgages, credit cards, Visas, or other sorts of credit or loans, the borrower will not repay the debt. Similarly, if a company continues to expand credit to a client, there is a potential that the client would not pay their invoices. Credit risk also includes the possibility that an insurer may refuse to pay when requested or that an insurance co will be unable to pay a claim. Credit risks are calculated using the debtor's general ability to repay an advance as shown by the loan's precise terms.

Loan approval, from the perspective of data analysis, is a binary classification problem in which a set of loan candidates data is reviewed and labelled as "high" or "low" risk. Binary classification is a simple type of classification in which a set of data is divided into two classes depending on some criteria. This type of categorization is most commonly employed when we wish to predict a specific outcome with only two possible values. Spam identification, medical diagnosis, credit card fraud detection, and, in our case, loan approval are just a few examples. Binary categorization is a relatively basic problem, despite its simplicity. Support Vector Machines, Decision Trees, K-nearest neighbour, Bayesian Classification, and Logistic Regression are some of the paradigms used to learn binary classifiers.

II. LITERATURE SURVEY

[1] "A survey on big data market : pricing, trading and protection"

The subject of big data trading is explored in this study. To be more explicit, the first step was to analyse existing big data research and identify the big data lifecycle for data trading, which included data gathering, data analytics, data pricing, data trading, and data protection. Then they looked at previous research on big data pricing. They explained the

DOI: 10.17148/IJIREEICE.2022.10574

relevance of data pricing, classified distinct market structures, data pricing methods, and data pricing models, and then enumerated the benefits and drawbacks of each category. They discussed critical difficulties related with data trading and potential solutions for the data trading process. They dug deeper into auction tactics, examining various systems, trading platforms, and associated complications. Lastly, as the final stage of the big data lifecycle, they looked into data protection. They classified contemporary copyright protection solutions and discussed the difficulties of copyright protection for massive data. It should be noted that the primary goal of this survey is to provide a comprehensive understanding of big data trading. In order to stimulate big data research and development, they emphasised the variety of topics connected to data pricing, data trading, and data protection, as well as areas that remain unsolved.

[2] " An Investigation of Credit Card Default Prediction in Imbalanced Datasets"
They have proposed something in this regard. In different sectors, machine learning approaches have been used in conjunction with the usage of unbalanced methodologies. The goal of this research is to use supervised learning techniques to forecast the client's behaviour when it comes to paying off their credit card amount. Because an uneven dataset is important for improving the model's performance in classification challenges, various normalisation strategies were utilised to balance the dataset. They began by employing exploratory data analysis approaches, such as data normalisation, to study the datasets. They began with the GBDT model and then compared the results to those from standard machine learning methods. The GBDT model has a greater prediction accuracy rate than typical machine learning-based models. On a Taiwan client's credit dataset, the GBDT approach had the best accuracy of 88.7% when using the K-means SMOTE re sampling method. The findings acquired from the Taiwan client's credit dataset were notably better than the results obtained from the other datasets used in this study.

[3]"Azure ML based Analysis and Prediction Loan Borrowers Credit Worthy"
Because religious budgeting foundations offer higher-interest loans, a peer-to-peer lending model is increasingly popular. The focus of this research is on whether or not borrowers will default, which is critical for a P2P loaning stage. This study provides a comprehensive, structured empirical examination of the most common techniques to addressing the problem of class imbalance and big data. They thoroughly investigated and ensured that no leakage in dataset records occurred while working on the dataset preparation. They showed and introduced their studies utilizing the Lending Club data set. They ran various tests to better understand the clean dataset and ensure that it was ready for their model. One of their goals in this project is to employ machines.

[4]"XGBoost Model and its application to personal Credit Evaluation"
The XGB approach is used in this research to create a credit evaluation system based on big data, which is then applied to personal credit evaluation, and the model's performance is assessed using several metrics. They explored the theoretical modeling of the credit classification problem using the XGB algorithm in this research, and then they applied XGB to the personal loan situation utilizing open data from the Lending Club Platform. The empirical research confirmed the clear advantages of XGB in feature selection and classification performance when compared to the performance of Logistic Regression, Decision Tree, Random Forests, and GBDT. Setting up a credit evaluation model utilising XGB based on multi-classification or regression problems could be an intriguing problem for future research.

[5] "CCCNet: An Attention Based Deep Learning Framework for Categorized Counting of Crowd in Different Body States"
They introduce classified crowd counting, a new type of crowd counting that counts the number of persons sitting and standing in a picture, in this work. We propose CCCNet, a three-phase deep learning architecture that incorporates both detection-based categorized density maps and global crowd density maps using attention mechanisms to effectively count the number of people sitting and standing in an image to solve the categorized crowd counting problem. Extensive testing on photos with widely variable human densities and cross-scene contexts demonstrates CCCNet's usefulness and superiority over competing systems. For seated and standing crowd counts, CCCNet has an MAE of 4.15 and 4.80, respectively, and an RMSE of 7.96 and 8.59.

[6] "Analyzing Data Mining Techniques on Bank Customers for Credit Score"
In this study, multiple data mining models are tested on a bank's data collection, and a model is chosen for the bank based on the results. The bank can readily forecast which customers would be advantageous to them and which customers will be defaulters or non-performing assets using that model. For prediction, three different models are used, with the decision tree performing best on the data set, with accuracy better than KNN and SVM.

[7] "Loan Default Prediction Model Improvement through Comprehensive Pre processing and Features Selection" m
To create loan default prediction models, this work used Nave Bayes, Decision Tree, and Random Forest classifiers. This work also examined three feature selection algorithms: Information Gain, Genetic Algorithm, and Particle Swarm Optimization, using multiple data pretreatment techniques. Preprocessing approaches significantly improved the

DOI: 10.17148/IJIREEICE.2022.10574

minority class prediction. The degree of improvement differed amongst the various classifiers. Applying feature selection algorithms enhanced the model as well, albeit there was little difference in improvement between the three strategies utilized. It can be inferred that data preparation is a critical stage in the development of a classification model since it has a significant impact on model accuracy. When working with a huge dataset, using features selection techniques is also very important; not only does it improve accuracy, but it also saves time

[8] "Improving Credit Risk Prediction in Online Peer-to Peer Lending Using Feature selectionwith Deep learning" They have presented an effective strategy based on feature selection and deep learning in this work. With the best accuracy, the approach determined the top n features. The accuracy of the LDA classifier employing the chosen features is higher than that of other approaches. With fewer features, the performance of P2P lending services can be enhanced. The use of a GPU for processing reduces the amount of time it takes to complete the task. As a result, the workload of staff credit assessment may be decreased because they will not be required to enter a large number of attributes into a database during the validation process. Their method is effective in credit risk analysis, according to the experimental data.

[9] "Data Analytics and ML: Bank Transactions over a Long Period of Time" They argue that a bank that is proactive in business in the twenty-first century has a lot of day-to-day transactions. To draw inferences, data analytics have to be performed on the data – both historical and current trends. The purpose was to construct or improve the machine learning model and compare accuracy. To analyse and derive findings, a python code was built and executed in the Jupyter platform. To fit the dataset, classification algorithms such as Support Vector Machine (SVM), Random Forest Classifier, and KNN algorithms are used. To understand the pattern of projected data, comparisons must be done throughout implementation. The accuracy of the Random Forest Classifier model is 99.9%, the accuracy of the Logistic regression model is 99.75 percent, the accuracy of the KNN model is 80.89 percent at K=3 and 5, and the accuracy of the SVM model is 75.87 percent. We may deduce that the Random Forest Classifier and Logistic Regression models fit this dataset the best. Because this data behaves well with the Linear regression algorithm, Linear regression is modelled from scratch without the use of libraries, resulting in higher accuracy (91.155%) and F1 score.

[10] **.** "An Approach for Predictionof Loan Approval using Machine Learning Algorithm", Proceedings of the International"
Cleaning and processing of data, imputation of missing values, experimental analysis of data set, model construction, and testing on test data are all steps in the prediction process. The best case accuracy found on the original data set is 0.811 on Data set. After analysing the data, the following conclusions were drawn: those applicants with the lowest credit scores will be denied a loan since they have a higher risk of defaulting on the loan. Most of the time, applicants with a high income and requests for a smaller loan are more likely to be approved, which makes sense because they are more likely to repay their debts. Other factors, such as gender and marital status, appear to be unrelated.

III. MACHINE LEARNING ALGORITHMS TO PREDICT BANK LOAN CREDIT RISK ANALYSIS

Here, are few machine learning algorithms we used to predict bank loan credit risk

A. K NEAREST NEGHBIOUR

It's a classification-oriented supervised machine learning algorithm. It determines the distance between the test data and the input and makes the appropriate prediction. The distance between data points is calculated by kNN, as we saw earlier. The Euclidean Distance formula is used for this.
The Euclidean Distance Formula

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

In machine learning, the above formula takes into account n number of dimensions, or features. The data point closest to the test point is believed to be from the same class as the test point. Because the method above works in n dimensions, it may be used to any characteristics.

DOI: 10.17148/IJIREEICE.2022.10574

B. SUPPORT VECTOR MACHINE

A supervised machine learning technique known as a support vector machine (SVM) can be utilized for regression and classification. SVM regression is a probabilistic technique that uses a set of mathematical functions. A kernel is a set of functions that transforms data inputs into the format that is intended. When dealing with non-linear regression problems, SVM converts the input vector(x) to an n-dimensional space called a feature space since it solves regression problems using a linear function (z). Non-linear mapping techniques are used to finish the mapping after applying linear regression to space.

C. LOGISTIC REGRESSION

The categorical dependent variable is predicted using logistic regression utilising a set of independent variables. A categorical dependent variable's output is predicted using logistic regression. As an outcome, the result must be a discrete or categorical value. It can be Yes or No, 0 or 1, true or false, and so on, but instead of giving exact values like 0 and 1, it delivers stochastic values that are somewhere between 0 and 1. Except for how they are employed, Logistic Regression is very similar to Linear Regression. Regression problems are solved using linear regression, and classification problems are solved using logistic regression.

D. RANDOM FOREST

Random Forest is a classifier that combines a number of decision trees on different subsets of a dataset and averages the results to increase the dataset's predicted accuracy. Instead than relying on a single decision tree, the random forest collects the forecasts from each tree and predicts the final output based on the majority votes of predictions. The higher the number of trees in the forest, the more precise it is and the problem of over-fitting is avoided
.

CONCLUSION

A bank that is engaged in business from the twenty-first century has a lot of day-to-day transactions. To draw inferences, data analytics have to be performed on the data – both historical and current trends. The purpose was to construct or improve the machine learning model and compare accuracy. To analyse and derive findings, a python code was built and executed on the Jupiter platform. To fit the dataset, classification algorithms such as Support Vector Machine (SVM), Random Forest Classifier, and KNN algorithms are used. To understand the pattern of projected data, comparisons must be done throughout implementation. efficiency of the Random Forest Classifier model is 99.9%, the efficiency of the Logistic regression model is 99.75 percent, the efficiency of the KNN model is 80.89 percent at K=3 and 5, and the efficiency of the SVM model is 75.87 percent. We may deduce that the Random Forest Classifier and Logistic Regression models works best for the dataset.

ACKNOWLEDGEMENT

We would like to take this opportunity to thank our guide **Mrs. Ranjitha J**, for giving us all the help and guidance needed. We are grateful for her kind support and valuable suggestions.

REFERENCES

[1] Fan Liang, Wei Yu, DOU AN, QINGYU YANG, XINWEN FU, AND WEI ZHA "A Survey on Big Data Market: Pricing, Trading and Protection", Digital Object Identifier 10.1109/ACCESS.2018.2806881 Corresponding author: Wei Yu (wyu@towson.edu)

[2] Talha Mahboob Alam, Kamran Shaukat, Ibrahim A. Hameed, (Senior Member, IEEE), Suhuai Luo, Muhammad Umer Sarwar, Shakir Shabbir, Jiaming Li, and Matloob Khushi, "An Investigation of Credit Card Default Prediction in the Imbalanced Datasets", Digital Object Identifier 10.1109/ACCESS.2020.3033784

[3] Khaldoon Alshouiliy, Ali AlGhamdi, and Dharma P Agrawal, "AzureML Based Analysis and Prediction Loan Borrowers Creditworthy", DOI 10.1109/ICICT50521.2020.00053

[4] Hua Lia, Yumeng Cao, Siwen Li, Jianbin Zhao and Yutong Sun, "XGBoost Model and its Application to Personal Credit Evaluation", : DOI 10.1109/MIS.2020.2972533, IEEE Intelligent Systems

[5] Sarkar Snigdha Sarathi Das, Syed Md. Mukit Rashid, and Mohammed Eunus Ali, "CCCNet: An Attention Based Deep Learning Framework for Categorized Counting of Crowd in Different Body States", 978-1-7281-6926-2/20/$31.00 ©2020 IEEE

DOI: 10.17148/IJIREEICE.2022.10574

[6] Durgesh Kumar Singh and Noopur Goel, “Analysing Data Mining Techniques on Bank Customers for Credit Score”, 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. June 4-5, 2020

[7] Ahmad Al-qerem, Mays Alhasan and Ghazi Al-Naymat, “Loan Default Prediction Model Improvement through Comprehensive Preprocessing and Features Selection”, 2019 International Arab Conference on Information Technology (ACIT), 978-1-7281-3010- 1/19/$31.00 ©2019 IEEE

[8] Van-Sang Ha, Dang-Nhac Lu, Gyoo Seok Choi, Ha-Nam Nguyen, Byeongnam Yoon, “Improving Credit Risk Prediction in Online Peer-toPeer (P2P) Lending Using Feature selection with Deep learning”, International Conference on Advanced Communications Technology(ICACT)

[9] Kavya K, Manish Y M, Priyanka P A, and Savinay Shukla, “Data Analytics and ML: Bank Transactions over a Long Period of Time”, IJARComputer and Communication Engineering Vol. 9, Issue 3, March 2020, DOI 10.17148/IJARCCE.2020.9331

[10] Mohammad Ahmad Sheikh, Amit Kumar Goel, and Tapas Kumar, “An Approach for Prediction of Loan Approval using Machine Learning Algorithm”, Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020) IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4.