Loan Eligibility Prediction Using Machine Learning Algorithms

26 December, 2022

1st Khairun Nisa
dept. Computer Science Engineering
BRAC University
19101376

2nd Arko Mazhar
dept. Computer Science Engineering
BRAC University
18201118

3rd Razit Kabir
dept. Computer Science Engineering
BRAC University
22241180

4th Sumaiya Mehjabeen
dept. Computer Science Engineering
BRAC University
19101116

***Abstract*—An accurate credit risk estimation system is necessary for the flawless and profitable operation of any financial organization. In a dynamic economy where the rate of loan defaults is steadily rising, it is becoming more difficult for the regulators of banking firms to accurately analyze credit requests and combat the hazards posed by non - performing loans. This research presents, in light of these occurrences, a proposed model with the concept and applications of machine learning models that can precisely evaluate the risk of credit while granting loans. Also, by recognizing the potential loan defaulters,we can create a better economic banking system which will also make the practices efficient.**

***Index Terms*—Models, Machine learning, Loan Prediction, Prediction, Testing, Training, KNN, Support Vector Machine, Random Forest.**

I. Introduction

Banks play a crucial role in the economy since they provide loans, receive deposits, arrange payments, and assist consumers with risk management. The primary business of banks is lending. The majority of the bank's earnings is derived directly from the interest received on loans. “There are now around 45 million student loan borrowers who collectively owe almost $1.6 trillion worth of debt in the U.S.” only [1] . In Bangladesh, “The loan amount varies from 50000 BDT to 20 Lac BDT, depending on the student's needs. The rate of interest is set at 9% per year” [2]. Even though a bank grants a loan after a regressive process of verification and testimonials, there is no assurance that the selected applicant is the proper applicant. This procedure requires additional time when performed manually. We can predict if a given hopeful is secure or not, and the entire testimonial process is mechanized by machine literacy. Due to the rapid use of online banking and payment digitization, the volume of transactions has expanded dramatically in recent years, necessitating that financial institutions employ more effective fraud protection methods [3]. Otherwise, the situation ”Bad loans” occurs, which occurs when the borrower no longer pays according to the loan's terms.”Due to the global financial crisis and ensuing European sovereign debt crisis led to a substantial increase in the number of nonperforming loans within the European banking system”. Bad loans decrease the profitability of banks and limit their ability to extend fresh credit. Large amounts of poor loans can pose capital adequacy issues for banks and, in the worst case scenario, lead to default [4]. Bangladesh's banking sector is witnessing the accumulation of bad loans at an alarming rate, as borrowers show a growing tendency to default on loans. Experts have attributed the situation to the banking industry's lack of adequate governance [5]. Thus, it is evident that loan default is a very serious problem that not only has a significant negative influence on financial institutions, but may also significantly drag down an entire nation's economy.

Machine learning has gained so much traction in recent years that it is now employed extensively in a wide range of fields, including banking services. The application of Machine Learning by banks to improve customer experience and back-office operations is continually expanding [6]. With machine learning algorithms, it is much easier, more accurate, and faster to spot patterns in loan defaults or credit deceit. Also, the number of loan defaulters and loans that have been written off is at an all-time high. This makes it more important than ever to have a good credit risk assessment model. This research is therefore going toward the application of machine learning models to more accurately predict credit eligibility in order to reduce the number of problematic loans in banking loan service. This study will discuss six supervised learning algorithms, as well as adequate dataset manipulation and data analysis, in order to determine the optimal method for predicting loan-eligible applicants. The classification techniques will include random forest, support vector machine, gradient boosting, and others.. In the paper, a comparison examination of each model is being conducted to determine the optimal model for credit risk evaluation. In the rest of this study, previous work and progress in the following fact are discussed and the proposed model as well as the result analysis are addressed.

II. Literature Review

Firstly,This study [7] focuses on proper loan requests and minimizing credit default risk. They found a rise in loan defaults. Modern machine learning algorithms play a crucial role in analyzing loan default risk, which may boost employment possibilities, financial stability, and profitability. First, the researchers utilized dummy variables to categorize people as loan-repaid (1) or not-repaid (2). (Labeled as 0). Skewed data creates fake minority class instances. SVM with RFECV is the best loan risk assessment model with 99.9% accuracy.

After that,The purpose of this work [8] is to offer a methodology for accurately calculating a borrower's credit score in order to make responsible lending decisions. To Gather Information,. The authors have utilized a publicly available data collection of lending club repository loans for their analysis. The author claims that the dataset contains information such as the borrowers' demographics, income levels, and repayment patterns. The authors have built a KNN algorithm credit rating model using Euclidean distance. The authors employed just 30% of the total data for validation, whereas 70% were used for model training. The plan was to sort the clients into groups according to their debt standing, income, and property ownership to weed out the most severe situations.

Bank workers check the details [9] of candidates manually and give the loan to eligible aspirants. Checking the details of all aspirants takes a lot of time. The ML model for prognosticating the credit threat of a bank. The adverse impact of loan prepayment rates on banks is a major issue, and banks are looking for further effective ways to handle loan blessing processes. Then we estimated the loan dereliction vaccination of Lending Club and Singapore Private Bank using R.F, L.R, G.B, KNN etc machine literacy models.

In an old research [14],the author describes Logistic Regression, a machine learning tool that uses predictive and probabilistic methodologies to solve the problem of loan acceptance prediction. She deduced from her prior work that a baseline criterion is required to determine whether a loan applicant is approved or not. Data Munging (Pre processing of data) and Standardizing the data notion are used to manage extreme values. Data are then fitted into the logistic regression model after all of these data training methods have been completed. The author doesn't declare any practical testing of the data whereas she only describes the procedures of building the model. The author concludes by describing the limitations of the logistic regression model, which include the need for a large sample size for parameter prediction; the inability to produce continuous outputs; and the requirement of independent variables for estimation.

In a recent paper [13] the authors use Naive Bayes, Random Forest, unpruned C4.5 decision tree, and pruned C4.5 decision tree, as well as Information Gain (IG), Genetic Algorithm (GA), and Particle Swarm Optimization (PSO) to increase classification efficiency and prediction accuracy (PSO). First, the writers focused on measuring borrowers' risk level to decide loan acceptance and interest rates. Support Vector Machine, two NN classifier approaches, the Levenberg-Marquardt algorithm (LM) with PLs filter, and others increased accuracy, performance, and training mistakes from prior studies. After preprocessing and feature extraction, Synthetic Minority Oversampling (SMOTE) is used for imbalanced data; Feature selection; Genetic algorithm-based feature selection and PSO elimination are implemented for unnecessary and noisy features. The author assessed four data mining classification algorithms. Also, Feature selection enhances loan eligibility forecasts, although the variation in improvement between three models was not spectacular - by the authors.

From here we can see the difference in times of using ML models, recently using the classifier and boosting algorithms enhanced the accuracy in outcomes.

III. Data Collection

The report's data came mostly from three publicly accessible data sets: lending club repository loan data, Singapore Banking Dataset, and Kaggle Partial Bank Loan Dataset. The data that was gathered includes loans that were authorized as well as loans that were denied approval. These include the lender's gender, marital status, details linked to dependents, academic qualification, monthly salary, loan amount, and area of residence. The data also includes loan payment history and demographic information about the borrower's financial situation. The standard client category and the non-default customer group are both included in the data set. After all was said and done, a total of 11 attributes and 6000 cases were chosen for the evaluation.

IV. Data Preprocessing:

Data preprocessing is essential for enhancing data quality overall. Quality decisions must be supported by quality data. Therefore, data preprocessing is required to improve accuracy and minimize errors.

A. Data Cleaning

As part of data preprocessing, data cleansing is performed to clean the data by filling in missing values, smoothing noisy data, resolving inconsistency, and removing outliers. There are particular parameters that determine the prediction, thus the columns that have no direct or indirect impact on classifier methods are eliminated from the datasets. The missing or null values are then identified and substituted. These are then filled with the mode and median values. As there are numerous columns with categorical values, these must be mapped to be accepted as input in datasets and by models. For removing outliers, we employ a clustering technique that groups similar data points. Outliers or inconsistent data are the tuples that are not contained within the cluster. We prepared our data in this manner for future exploratory analysis.

B. *Exploratory Data Analysis*

In exploratory data analysis, one of the most significant components is to understand the level of your analysis. This is determined by the number of variables or columns that we have in our research. Depending on that the level of analysis is divided into three different analysis techniques. They are: Univariate analysis, Bi-variate analysis and Multivariate analysis.

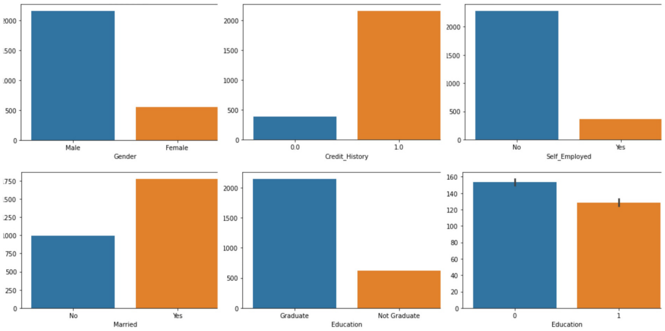

Fig. 1. Univariate analysis for parameters

This is one of the first and basic steps that we perform in exploratory data analysis. This helps us to understand more about the dataset and ensures us to go on the right path whenever we are implementing the machine learning models. As we have splitted our dataset before into train and test datasets, we mapped the N to 0 and Y to 1, and implemented Univariate analysis. When the data contain only one variable and don't deal with a cause or an effect relationship, a univariate analysis technique is used.

We have made a number of observations, such that male applicants outnumber female applicants and that there is a noticeable distinction between self-employed applicants and others. Afterwards, we implemented bivariate analysis which is slightly more analytic than univariate analysis, hence we used it for those data variables that need comparison between them. We have conducted it by correlation coefficients and regression. Thus the bivariate analysis measures the correlation between two variables.

Lastly the multivariate analysis which is the most complex form of statistical analysis and used only when there are more than two variables in a dataset. Here we have done a correlation matrix.

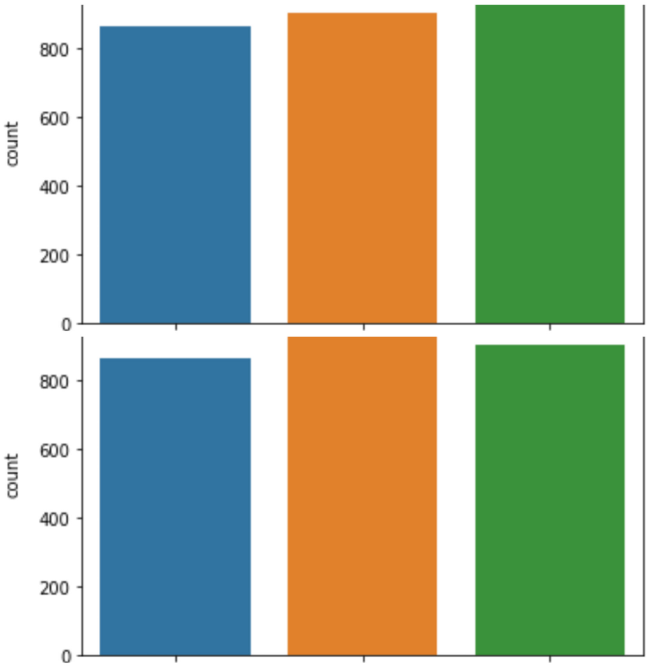

Fig. 2. Bivariate analysis for parameters

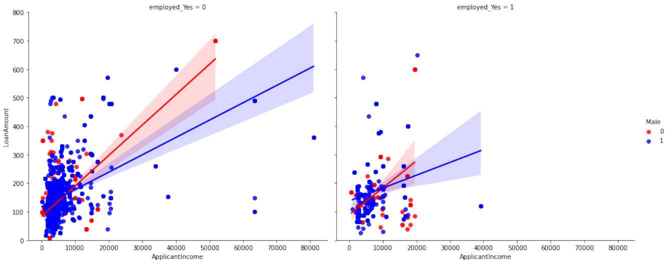

Fig. 3. Visualizing the Correlations and Relations

V. Machine Learning Classification Models

In our data set we have been used six machine learning classification models for the loan prediction.

They are-

1. Random Forest
2. Logistic Regression
3. K-Nearest Neighbors
4. Gradient Boosting Classifiers
5. Naive Bayes
6. Support Vector Machine

Why do we choose them?

A. *Random Forest*

Instead of depending just on one decision tree, the random forest gathers the data from all of the trees and makes predictions about the future based on the majority of them.

B. Logistic regression

A classification algorithm is logistic regression. With the aid of several independent variables, it is used to forecast a binary result (1/0, True/False, Yes/No). We utilize logistic regression since there are many logistic variables in our dataset.

C. K-Nearest Neighbors

In a banking system, KNN may be used to determine if a person is qualified for a loan or whether they have qualities with defaulters. The K-NN method presupposes a resemblance between the new case/data and the cases already in existence. Different KNN classifiers are tried and assessed to assess the credit risk.

D. Gradient Boosting Classifiers

The effectiveness of gradient boosting techniques has been demonstrated in a vast range of applications, making them potent instruments for extracting precise predictions from data sets. This model is used because it combines several weak learners into a single powerful model.

E. Naive Bayes

Multi-class prediction issues can be solved with naive Bayes. Since our data set contains several classes, this is one of the better models to use to assess the dataset's correctness.

F. Support Vector Machine

With a grid search method for improved prediction and cross-validation for much more trustworthy findings, SVM was employed in this situation to anticipate the outcomes. We utilize SVM when the distribution of the data is irregular. Because the data in our dataset were erratic, we used this model.

VI. Result

Here, as we have examined 6 types of models and reviewed in greater depth, we found that in the Random Forest model there is the best accuracy we got. The larger number of trees in the forest prevents higher accuracy and overfitting. Because the random forest combines numerous trees to forecast the class of the dataset, some decision trees may predict the correct output while others may not. The other five models gained comparatively lower accuracy than the Random Forest model. The Random Forest model got an accuracy of 93.397359%. On the other hand, the K- Nearest Neighbour and Gradient Boosting Classifier model also got accuracies of 86.434574% and 85.594238% respectively which are comparatively lower than the accuracy of Random Forest model. The accuracies of Logistic Regression and Naive Bayes are so close to each other which are respectively 78.751501% and 76.830732%. On the other hand, the SVM model got the lowest accuracy among all, which is 69.747899%.

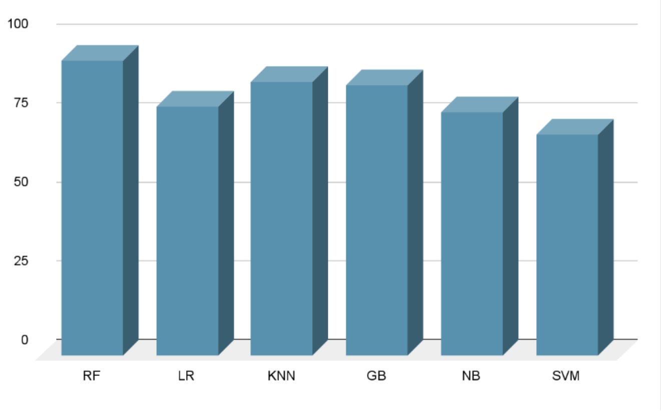

Fig. 4. Model Comparison

VII. Conclusion

In order to eliminate human interference and boost productivity, the rapidly expanding IT sector of today needs to develop new technology and upgrade existing technology. Anyone looking to apply for a loan or the banking system will use this model. It will be very beneficial for managing banks. It is abundantly obvious from the data analysis that it lessens all fraud committed at the time of loan acceptance. Everyone values their time highly, thus by doing this, not only the bank but also the applicant's wait time will be shortened. Here data collection, exploratory data analysis, data preprocessing, model building, and model testing are the analytical processes involved in building this system. Our model has produced impressive results using data from Lending Club and Singapore Private Bank, which can help banks and other institutions to analyze the loan risk of borrowers and help financial institutions continue to operate in a transparent and lucrative manner.

References

[1] 65 Student Loan Statistics: 2022 Data, Trends Predictions, = "https://research.com/education/student-loan-statistics", year = 2022, note = "[Online; accessed 20 Dec-2022]"

[2] Education Loan In Bangladesh, = "https://bangladeshpost.net/posts/education-loan-in-bangladesh-70483", year = 2021, note = "[Online; accessed 11 Oct 2021]"

[3] Machine learning in banking: 8 use cases and implementation guidelines, = "https://www.itransition.com/machine-learning/banking", year = 2022, note = "[accessed July 29, 2022]"

[4] How bad loans affect banks and financial stability, = "https://www.riksbank.se/en-gb/press-and-published/notices-and-press-releases/notices/2019/how-bad-loans-affect-banks-and-financial-stability/", year = 2019, note = "[accessed March, 2019]"

[5] Default loans plague banking sector, = "https://archive.dhakatribune.com/business/banks/2018/03/22/default-loan-plagues-banking-sector", year = 2018, note = "[accessed March 22nd, 2018]"

[6] How Machine Learning is Used in Finance and Banking, = "https://exadel.com/news/how-machine-learning-is-used-in-finance-and-banking/", year = 2022, note = "[accessed July 6, 2022]"

[7] Shoumo, S. Z. H., Dhruba, M. I. M., Hossain, S., Ghani, N. H., Arif, H., Islam, S. (2019, October). Application of machine learning in credit risk assessment: a prelude to smart banking. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 2023-2028). IEEE.

[8] Gomathy, C.K. (2021, December) THE LOAN PREDICTION USING MACHINE LEARNING

[9] Orji Ugochukwu and Ugwuishiwu, hikodili H. (2022. june) Machine Learning Models for Predicting Bank Loan Eligibility, pp. 1–6. Available at: doi.org/10.1109/NIGERCON54645.2022.9803172

[10] Arutjothi, G., Senthamarai, C. (2017, December). Prediction of loan status in commercial bank using machine learning classifier. In 2017 International Conference on Intelligent Sustainable Systems (ICISS) (pp. 416-419). IEEE.

[11] Gupta, A., Pant, V., Kumar, S., Bansal, P. K. (2020, December). Bank Loan Prediction System using Machine Learning. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 423-426). IEEE.

[12] Gautam, K., Singh, A. P., Tyagi, K., Kumar, M. S. (2020). Loan Prediction using Decision Tree and Random Forest. International Research Journal of Engineering and Technology (IRJET), 7(08).

[13] Al-Qerem, A., Al-Naymat, G., Alhasan, M. (2019, December). Loan default prediction model improvement through comprehensive preprocessing and features selection. In 2019 International Arab Conference on Information Technology (ACIT) (pp. 235-240). IEEE.

[14] Vaidya, A. (2017, July). Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.