Comparing Machine Learning Alogrithm For Credit Risk Analysis

Parab Sanika
Student ID: 20210987
School of Computing
Dublin City University
Dublin, Ireland
sanika.parab4@mail.dcu.ie

Hemmadi Anjali
Student ID: 20211101
School of Computing
Dublin City University
Dublin, Ireland
anjali.hemmadi2@mail.dcu.ie

Vaidya Kunal
Student ID: 20211188
School of Computing
Dublin City University
Dublin, Ireland
kunal.vaidya2@mail.dcu.ie

Nair Rohit
Student ID: 20210378
School of Computing
Dublin City University
Dublin, Ireland
rohit.nair2@mail.dcu.ie

***Abstract*— Lending money is an integral part of the financial sector. When a customer applies for a loan from a bank or some other financial institution, they check the customer's credit history to ensure they can repay the loan, this process is often called creditworthiness. There are different techniques that are currently used to evaluate the same for the applicants. Some banks and financial institutions have developed models that assist them in their decision-making process. In this project we are comparing different machine learning approaches by sampling the dataset, calculating their accuracy and combining them using the weighted average of their accuracy to determine if the applicant should be granted the loan or not.**

Keywords— logistic regression, random forest, XGBoost, data mining, credit risk analysis

I. Introduction

The banking sector heavily relies on a good credit evaluation system for granting loans. When a customer applies for a loan, the bank checks if the customer can pay back the loan amount. The ability of a customer to pay back the loan amount is called their creditworthiness. Calculating the creditworthiness, in a very simplified manner, would involve checking their credit score, past repayment details, verifying their income, assets and net worth. A higher credit score is indicative of the fact that the applicant is more likely to pay back the loan. The applicants are broadly categorised into two classes, the ones that are likely to repay the loan and the ones that are likely to default.

Financial organisations incur huge losses by lending to someone that is likely to default. This is where credit risk analysis comes in handy. Banks and other financial institutions use it as a tool to evaluate the risk involved in lending a loan to a customer, ensuring least amount of risk and thereby reducing potential losses. As credit risk analysis plays a vital role in this industry, even a slight increase in the accuracy helps avoiding the losses incurred.

This project explores different approaches to address the problem mentioned above. There are multiple approaches that have been implemented previously [2]. The dataset being used has information regarding the past loan grants, customers who defaulted on their payments, information about their employment etc. The aim is to use this dataset to identify patterns which would suggest if a person were likely to default on their payment. This information can be used to deny the loan, increase the interest rates or reduce the loan amount that would be sanctioned. We use logistic regression, random forest and XGBoost algorithm for our project. We trained the model using the training dataset individually, calculated the accuracy of each model to determine how they are performing and then combined them by using their accuracy as a weighted average in the majority vote.

The code used for this project has been uploaded to the GitHub repository [18].

II. Literature Review

Credit risk analysis is very important and discussed subject in the field of banking and finance. Besides that, after the recent surge in data science and several other advances in machine learning, credit risk analysis has gained even more importance. Many significant discoveries have been made in this area, which will serve as a springboard for future research and studies. As shown in [4],[5], and [6], Logistic Regression (LR) models are a feasible solution for credit risk assessment and are used as a standard approach in credit risk analysis. The LR model outperformed the ANN(Artificial Neural Network) model in finding loan defaulters in [5] and [6]. Models based on Random Forest are also used in credit risk assessment. The random forest model has been implemented and proven to be a better approach than the standard approach for credit risk analysis in [8], [12], and [13]. In [7], the authors have implemented multiple models which belong to the class of nonparametric regression approach like logistic regression, kNN(k-nearest neighbours), bNN(bagged k-nearest neighbours), and Random Forest for a comparative analysis which results in the Random Forest model outperforming all the other models. The papers [9] and [11] show the application of the Extreme Gradient Boosting model in identifying defaulters. A comparative analysis was made between the top three models used for credit risk analysis i.e., Logistic Regression, Random Forest, and Extreme Gradient Boost models in [10]. The results clearly show that if applied with proper tuning the Extreme Gradient Boosting model outperforms the other two models in [7].

As a result, previous related works show that much progress has been made in the field of credit risk assessment and that much more can be achieved. According to the reports, a proper dimensionality reduction process, parameter tuning, noise handling, and imbalanced dataset handling are all essential for developing a credit risk assessment model. Computational cost and time complexity should also be considered. This paper puts forward a comparative analysis between the Logistic Regression, Random Forest, and Extreme Gradient Boost model.

A grid search with cross-validation can be used to tune parameters and manage noise, and the issue of an imbalanced dataset will be effectively addressed. Different supervised classification algorithms will be compared using different dimensionality reduction methods, and the best combination for credit risk assessment will be chosen. Overall, our goal is to present a model that addresses all the aforementioned issues and establish an accurate model for evaluating credit risk in the banking industry.

III. Methodology

This project uses the KDD Process (Knowledge Discovery in Database) for Data Mining. It provides a structured process to extract knowledge from data by using apropriate data mining algorithms along with any pre-processing that may be required. Data mining is the process of using specific algorithms to extract patterns and information from the data based on what the end goal is [3].

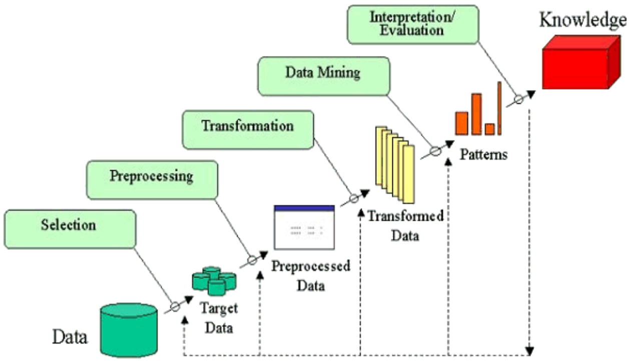

Fig 1. Outline of the Steps of the KDD Process [1]

The steps involved in the KDD process is as follows:

A. *Domain Understanding*

Financial institutions receive many loan applications daily. Before the loan is sanctioned, the organisation checks the applicants credit history to determine if the extent to which the applicant is suitable to receive the loan. The aim of this research project is to calculate risk associated with loans using anonymised past loan payment information of the customers. It attempts to identify the factors that determine if a loan is granted or not and use that to determine the creditworthiness of future applicants.

B. *Data Selection*

The dataset chosen for this project is of Lending Club company which is available on Kaggle. It has past loan application information, their income, length of employment, details if they have defaulted on their loan payments, the zip code of the loan application and other relevant information. It has 27 columns and 395753 records.

C. *Data Cleaning, Pre-processing and Transformation*

The columns loan amount and instalment had high correlation. Grade and sub-grade had values from A-G and A1-A5, B1-B5, etc. The number of rows in F and G grade and sub-grade were not significant enough to impact the predictions as they had far too little records compared to the dataset. For home ownership, a categorical variable, all categories that had records less than 10000 were combined and renamed as others. Similarly, for the column purpose, the same approach was followed. Term has only 2 values, 36 months and 60 months, so they were converted into 0 and 1 respectively. The employment length was replaced with the numeric value. Total account and mort account correlate with each other. The missing values in mort account can be filled with total account. The missing values of mort account are filled with the mean values of mort account that correspond to the groups made using total account.

The column loan status is our target variable, having values “Fully Paid” and “Charged Off”. Here fully paid means the loan amount has been fully repaid by the borrower and charged off means the borrower has defaulted on their loan payment. The values “Fully Paid” and “Charged Off” have been converted to 0 and 1 respectively. The dependent variables were compared with the target variable to get the distribution of the target variable, across all the dependent variables. This is helpful in imputing missing values if any. By this process we discovered that the target variable had homogeneous distribution across employment length which meant employment length does not provide us with any new information to help us with the classification.

The data had a class imbalance issue with the majority class having 318143 records and the minority class having 77610 records. Initially, the dataset was split into test and train which was used to train the three models. To improve the accuracy of the model, we under sampled the dataset to have a 1:1 ratio between the classes. One sample would lead to a biased model, so rather than using one under sampled datasets, 3 under sampled datasets were created. Each model would train 3 times on each individual dataset respectively. The aggregated results of these 3 models would give us a better model which would generalise well on the whole dataset.

Categorical variables cannot be used directly to train the model. They have to be represented in numerical form such that the underlying information and the nature of the categorical variable remains the same. The data that does not have any order were identified and their unique values were counted (for zipcode and sub-grade). Though zipcode looks like numeric data, it is not a value which has any order. Zipcode is just one way to name a place and hence it is treated as a categorical variable. We have created dummies for zipcode to check if it has any effect on the outcome.

D. *Data Mining*

Three different models were chosen which are as follows:

a. *Logistic regression*: Logistic regression is one of the most used algorithms for classification. It uses a logistic function to classify dependent variables. It is efficient to train and does not make any assumptions about the distribution of classes. It tends to overfit if the number of observations is lesser than the features. One major drawback of it is that the boundaries it constructs are linear. It assumes a linear relationship between dependent and independent variable and hence the nature of the dataset matters a lot while using this algorithm.

b. *Random Forest:* This algorithm was first introduced in 1995 by Tin Kam Ho [16]. It is an ensemble learning method for classification and regression [17]. It is an extension of decision tree and solves the overfitting problem often associated with random forest. It can run efficiently over large or imbalanced datasets. But it often takes longer time to train the model. Though it is not black box it is tough to know what is going on.

c. *Extreme Gradient Boost:* Extreme Gradient Boost, also known as XGBoost, is a class ensemble machine learning algorithm. It can be used for classification or regression problems. It is excellent at handling outliers. This, like random forest, can handle large datasets and is not prone to overfitting. But tuning it becomes a bit difficult due to the amount of hyperparameters involved.

Each model was run on 3 different under sampled data. This means, effectively 9 different models were run. Additionally, hyperparameter tuning was performed which gives the best possible model for the given dataset. Logistic regression is one of those models which helps in interpreting the data. Using that we were able to determine how each feature affects the outcome. Logistic Regressing is good for interpretability. Random Forest captures nonlinear relationships well. XGBoost, just like Random Forest, captures nonlinear relationships but does that more efficiently.

E. Interpretation / Evaluation

Table I Performance Of All Models

Model	Accuracy	Precision	Recall 1	Recall 0
Logistic Regression	80.46 %	81 % (0) 80 % (1)	81 %	80 %
Random Forest	79.56 %	78 % (0) 81 % (1)	77 %	82 %
XGBoost	80.8 %	81 % (0) 80 % (1)	81 %	80 %

Table 1 shows the performance scores of each model. The performance of the Random Forest, XGBoost and Logistic Regression are similar. The nature of the data is linear hence Random Forest and XGBoost do not outperform Logistic Regression. Logistic Regression was used to find how each feature affects the outcome. The coefficient values of zipcodes, dummies for which were created, are as follows:

Table III Zipcode Coefficients

Zipcode	Coefficients
05113	-3.5817
11650	3.7942
22690	0.0989
29597	-3.6544
30723	0.1184
48052	0.1957
70466	0.1725
86630	3.7388
93700	3.7951

Table III shows the coefficient values of zipcodes. This shows how different areas have different effects on the loan application. There 9 models can be combined using their respective accuracies as coefficient for the weighted average. A combination of these weighted averages will be used to predict the outcome of the target variable.

IV. Conclusion

There is a lot of data being generated by the minute. The goal of this project was capitalising on this huge source of data and trying to predict the outcome by combining multiple algorithms. Any decision, if not backed with data will not yield good results. We have tried to address this issue by using a weighted average of different models.

Logistic Regression could not produce consistent results despite being close to the other two models in terms of accuracy. Random forest and XGBoost were able to produce consistent results across all runs. However, XGBoost takes considerably lesser time to train compared to Random Forest.

References

[1] "KDD Process/Overview", Www2.cs.uregina.ca. [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html.

[2] T. Pandey, A. Jagadev, S. Mohapatra and S. Dehuri, "Credit risk analysis using machine learning classifiers", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017. Available: https://ieeexplore.ieee.org/document/8389769.

[3] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, no. 3, pp. 37-54, 1996.

[4] H. A. Bekhet and S. F. K. Eletter, "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach," Review of Development Finance, 24-Apr-2014.

[5] G. V. Attigeri, M. M. M. Pai, and R. M. Pai, "Credit Risk Assessment Using Machine Learning Algorithms," Latest TOC RSS, 01-Apr-2017.

[6] Salehi, M. and Mansoury, A., 2011, "An evaluation of Iranian banking system credit risk: Neural network and logistic regression approach", International Journal of Physical Sciences, 6(25), pp.6082-6090.

[7] J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler, "Consumer credit risk: Individual probability estimates using machine learning," Expert Systems with Applications, 21-Mar-2013.

[8] L. Tang, F. Cai, and Y. Ouyang, "Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China," Technological Forecasting and Social Change, 17-Mar-2018.

[9] Y.-C. Chang, K.-H. Chang, and G.-J. Wu, "Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions," Applied Soft Computing, 05-Oct-2018.

[10] P. Carmona, F. Climent, and A. Momparler, "Predicting failure in the U.S. banking sector: An extreme gradient boosting approach," International Review of Economics & Finance, 26-Mar-2018.

[11] F. Climent, A. Momparler, and P. Carmona, "Anticipating bank distress in the Eurozone: An Extreme Gradient Boosting approach," Journal of Business Research, 15-Nov-2018.

[12] M. S. Uddin, G. Chi, M. A. M. A. Janabi, and T. Habib, "Leveraging random forest in micro‐enterprises credit risk modelling for accuracy and interpretability," Wiley Online Library, 30-Nov-2020.

[13] N. Ghatasheh, "Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study," International Journal of Advanced Science and Technology, vol. 72, pp. 19–30, 2014.

[14] Vinod Kumar L, Natarajan S, Keerthana S, Chinmayi K M and Lakshmi N, "Credit Risk Analysis in Peer-to-Peer Lending System," 2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA), Singapore, 2016, pp. 193-196, doi: 10.1109/ICKEA.2016.7803017.

[15] Jerome H. Friedman "Greedy function approximation: A gradient boostingmachine.," The Annals of Statistics, Ann. Statist. 29(5), 1189-1232, (October 2001)

[16] T. K. Ho, "Random Decision Forests," *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pp. 278-282, August 1995.

[17] M. Utari, B. Warsito and R. Kusumaningrum, "Implementation of Data Mining for Drop-Out Prediction using Random Forest Method", *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 2020. Available: 10.1109/icoict49345.2020.9166276.

[18] Nair, R., Hemmadi, A., Vaidya, K., Parab, S., 2021. *Data Mining: Credit Risk Analysis*. [online] GitHub. Available at: https://github.com/nrohit78/Data-Mining_Credit-Risk-Analysis.