
Machine Learning Models Evaluation and Feature Importance Analysis on NPL Dataset

Rufael Fekadu *†

School of Electrical and Computer Engineering
Addis Ababa Science and Technology University
Addis Ababa, Ethiopia

Anteneh Getachew *†

School of Electrical and Computer Engineering
Addis Ababa Science and Technology University
Addis Ababa, Ethiopia

Yishak Tadele *†

School of Electrical and Computer Engineering
Addis Ababa Science and Technology University
Addis Ababa, Ethiopia

Nuredin Ali †

Chapa AI Research (ChAIR)
Chapa Financial Technologies
Addis Ababa, Ethiopia

Israel Goytom

Chapa AI Research (ChAIR)
Chapa Financial Technologies
Addis Ababa, Ethiopia

Abstract

Predicting the probability of non-performing loans for individuals has a vital and beneficial role for banks to decrease credit risk and make the right decisions before giving the loan. The trend to make these decisions are based on credit study and in accordance with generally accepted standards, loan payment history, and demographic data of the clients. In this work, we evaluate how different Machine learning models such as Random Forest, Decision tree, KNN, SVM, XGBoost, and AdaBoost perform on the dataset provided by a private bank in Ethiopia. Further, motivated by this evaluation we explore different feature selection methods to state the important features for the bank. Our findings show that SVM achieves the highest F1 score on the KMeans SMOTE over-sampled data. We also found that the most important features are the age of the applicant, years of employment, and total income of the applicant rather than collateral-related features in evaluating credit risk.

*Work done when the authors were a research intern at Chapa.

†Equally contributed to this work.

1 Introduction

Loans are one of the primary sources of income for both private and commercial banks. Large portion of banks' profit directly comes from the interests on loans given. Profitable banks play a vital role in the economic development of each country and a bank with a poor management system is an obstacle to the economic development of the country. The role of banks is spiral in developing countries as there is a growing need for financial institutions.

Although most loans are paid back based on their schedule some others default. Those defaulting loans are known as non-performing loans(NPL). In recent years, evaluating and predicting the non-repayment ability of a customer is one of the most challenging issues in commercial banks. [1] The presence of huge non-performing loans in the banking industry highly affects the safe operation of banks, and may lead to bank failures and even nation-wide financial crisis. Since lending without better evaluation techniques may lead to immediate losses, finding a way to reduce the credit risk of financial institutions has been a major area of concern for many researchers around the globe.

Banks use different criteria to predict the creditworthiness of applicants. To evaluate the creditworthiness of customers most Ethiopian banks focus on loan payment history, demographic data, and collateral quality of the clients. Although this system tends to work well sometimes loans default due to various reasons. In recent years the attention of Ethiopian based banks on their non-performing loans has grown. Different types of initiatives should be made to reduce the amount of non-performing loans to strengthen the financial institutions. Globally, plenty of research has been conducted [2] to determine the creditworthiness of a customer, partitioning the credit groups of good or bad payers.

The rise of AI in finance and beyond has understandably garnered a great deal of attention in recent years [3]. Different machine learning techniques have been evaluated to predict non-performing loans on different banks' datasets.

This work aims to evaluate the performance of various machine learning algorithms on the dataset. We also explore different feature selection mechanisms to select the most important features and state those features to the bank. To the best of our knowledge, there is no extensive research on identifying defaulting loans in Ethiopia, which results in the problem getting worse.

2 Related Works and Background

There have been various works addressing the performance of machine learning models on predicting NPLs in different countries. Banks follow different criteria to predict whether a loan will default or not. Various machine learning models have been evaluated and their performance differs based on the data used.

[4] Applied different machine learning algorithms for loan prediction. Their work concluded that random forest has much better accuracy than other algorithms such as logistic regression, decision tree, and support vector machine. [5] Makes a comparative analysis of different algorithms and concludes that random forest is among the best methods to try for credit risk prediction. [6] After reviewing various machine learning methods available and the equally numerous applications, it is clear that declaring a single best method is impossible. Methods have specific strengths and weaknesses that align with different applications. A comprehensive study is conducted to compare the performance of XGBoost algorithm with logistic regression [7]. Their results show that the XGBoost algorithm works much better.

[8] Established a solid comparison between different classification algorithms. As a result, they found random forest as the best performing model, and Naive Bayes as the worst in terms of accuracy and Area Under the Curve (AUC). [9] Investigated Logistic Regression, Random Forest, Decision Tree, AdaBoost, XGBoost, Artificial Neural Network and Support Vector Machine. They applied SMOTE to overcome the problem of imbalance. Their results shows that XGBoost without SMOTE implementation obtained the best result with respect to the chosen model evaluation metric which is F1-Score.

The work by [10] used K-means SMOTE to solve the imbalance problem and feature importance scores generated by random forest are fed into BP neural networks as an initial weight. The work concludes that the improved version of smote algorithm (k-means SMOTE) effectively solves the imbalance problem and the introduced technique improves the prediction performance of the model

to a certain extent. [11] Evaluated various machine learning algorithms to predict default loans and concluded that Neural Network model was the best model with higher accuracy and low average square error also Random Forest model better resulted than Logistic Regression model.

3 Dataset and Challenges

3.1 Dataset Description

We used a dataset from one of the largest Ethiopian Private Banks, Bank of Abyssinia. There were two separate files, applicants record and loan record. The application record contains information about the applicants. It has 438557 records with 18 features. The loan record contains the monthly repayment status of each applicant. It has 1048575 records with 3 features. We merged both files based on the Customer ID, and got 36,326 records. The dataset has class imbalance, null entries, invalid features, and duplicate records.

3.2 Exploratory Data Analysis(EDA)

In order to get a better understanding of the dataset, we provided some explanatory plots. The distribution of the target class is shown in Figure 1. There is a high correlation between CNT_FAM_MEMBERS and CNT_CHILDREN. Since both features are not correlated to the target class STATUS. CNT_CHILDREN is dropped randomly, As [12] a high correlation between features may lead to a biased feature importance ranking and unstable models.

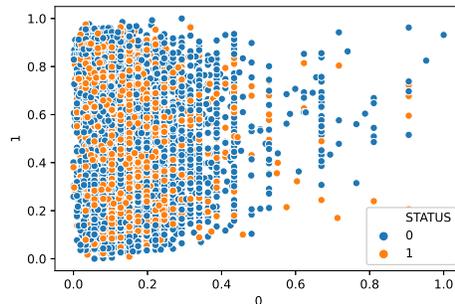


Figure 1: Shows the distribution of the target class.

4 Experiments

4.1 Data Pre-processing

Since the dataset has null values, invalid entries, and duplicates some pre-processing techniques are introduced before it is used for training and testing.

Null /Empty Entries Incomplete data is an unavoidable problem in dealing with most real-world data sources [13]. In this dataset, there is a noticeable amount of null entries. Therefore, features with 30% of empty entries are removed.

Invalid features and Duplicate records Invalid features are columns in which their information has no meaning for the objective. This includes columns that contain only a single value, or columns that indicate whether a customer provided personal information or not. Hence, the features are dropped.

From the 36,326 records, we found 25,268 duplicate entries on the dataset. The duplicates were dropped as the presence of duplicates in the learning sample does indeed pose a problem[14].

Categorization and outlier removal An outlier is an unlikely observation in a dataset and may have one of many causes [15]. In this case, we removed outlier values of numeric features.

The credit status of the applicants for each month was recorded as c: paid off that month, x: no loan for the month, 0: (1-29) days pass overdue, 1: (30-59) days pass overdue, 2: (60-89) days pass overdue, 3: (90-119) days pass overdue, 4: (120-149) days pass overdue and 5: more than 150 days pass overdue. We categorized c, x, and 0 as good loans and 1, 2, 3, 4, and 5 as bad loans.

Normalization and encoding The features came in different formats. Eg., string, unbounded integers, floating numbers, Boolean values. This poses a challenge to work with machine learning algorithms. We encoded the categorical features to be represented by integer values. One hot encoding is then used for features that are with more than two categories.

A series distribution of certain data could affect the performance of machine learning [16]. To get an equal contribution of the features, the numeric features are normalized. After pre-processing, the dataset reduced to 11058 records and 13 features.

The dataset was categorized into training and testing: 80% of the dataset for training, 20% for testing. The imbalance treatment is applied on the training set only.

4.2 Imbalance Treatment

Class imbalance is a problem that exists when the distribution of the target feature has a high variation: one of the target classes is found highly distributed while the other is scarcely distributed. In the dataset, the ratio of individuals who have entered NPL is very low compared to the total loan. 78.5% of the data are performing loans while 21.5% are non performing loans. To solve this problem, the following two oversampling techniques are implemented to generate synthetic samples that favour the minority class.

K Means-SMOTE

K Means-SMOTE [17] - works in three steps, the first is to group the input samples into different clusters. The second is to identify a cluster with the low distribution of minority class, assign more synthetic samples, and then finally over-sample each filtered cluster using SMOTE [18]. Using this technique we synthesized minority class samples and added them to the original training data.

Conditional Tabular GAN(CTGAN)

Works based on the principle of the CGAN model to synthesize new samples that are optimized for tabular data which accounts for continuous and discrete features[19].

4.3 Important Features Selection

It's important to identify a set of significant features relevant to determine the credit risk of the loanee. The following methods are applied to get the most important features from the dataset.

Random Forest feature selection calculates feature importance based on node impurities at each decision tree and takes a mean of all decision tree feature importance to get the final feature importance. According to [20], RF feature selection method are extremely useful and efficient in selecting important features.

Extreme Gradient Boosting(XGBoost) calculates feature importance after the boosted decision tree is built. [21] found that XGBoost preserves key features necessary for prediction.

In this work, we have used RF and XGBoost feature selection methods to obtain the feature importance score. The feature importance scores generated by random forest are shown in figure 2a. While the scores generated by XGBoost are shown in 2b. The top three important features are selected as there is a high variation compared to the score of the fourth feature. The three important features obtained from both RF and XGBoost are the same. To cross-validate the results of random forest and XGBoost, we implemented Recursive feature elimination using random forest to provide the three most important features which resulted in the same features. The most important features based on the above experiments are age years, total amount of income, and years of employment.

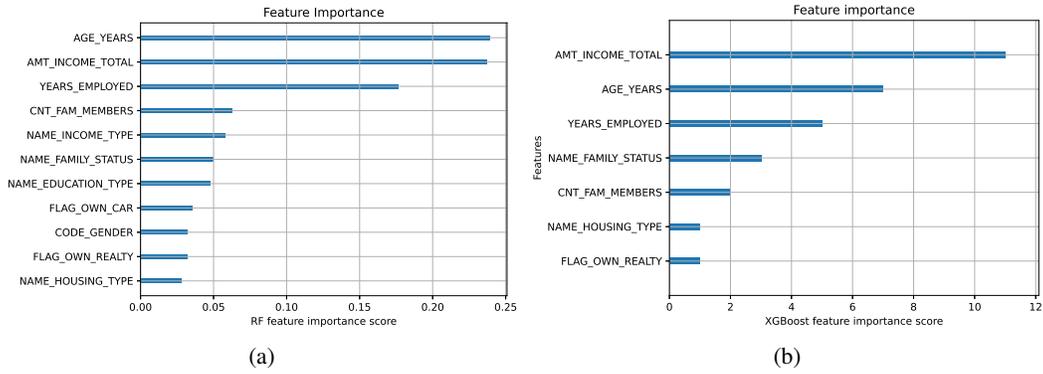


Figure 2: Feature Importance scores

5 Results

In order to evaluate the performance of the models, appropriate metrics should be selected. For the purpose of identifying the non-performing loans we identify positive classes (i.e. high precision) while minimizing those the model is miss-classifying as positive(i.e. high sensitivity). In this case, F1-score, which is the weighted average of Precision and sensitivity, is used as an evaluation metric as it conveys the balance between the precision and the recall. The models are evaluated based on their $F1$ score of the minority class which is 1.

The results show that the use of K Means-SMOTE over-sampling is effective compared to the unbalanced dataset, under-sampled dataset and a dataset over-sampled using CTGAN. Support vector machine achieves the highest performance.

Table 1: Models performance on KMeans SMOTE over-sampled data

Model	Precision(1/0)	Recall(1/0)	F1-score(1/0)
Random Forest Classifier	0.2392/0.7806	0.1907/0.8261	0.2122/0.8261
Logistic Regression	0.216/0.7737	0.3185/0.6684	0.2574/0.6684
Decision Tree Classifier	0.1795/0.7664	0.1602/0.79	0.1693/0.79
SVM Classifier	0.2253/0.7782	0.3185/0.6859	0.2639/0.6859
K Neighbors Classifier	0.1961/0.7669	0.2434/0.7138	0.2172/0.7138
XGBoost	0.1391/0.771	0.0426/ 0.9244	0.0652/ 0.9244
AdaBoost	0.2304/0.7791	0.215/0.7941	0.2225/0.7941
Voting Classifier	0.1699/0.7684	0.1075/0.8493	0.1317/0.8493

Table 2: Models performance on CTGAN over-sampled data

Model	Precision(1/0)	Recall(1/0)	F1-score(1/0)
Random Forest Classifier	0.2143/0.7767	0.0426/0.9552	0.0711/0.9552
Logistic Regression	0.2115/0.7723	0.284/0.6963	0.2424/0.6963
Decision Tree Classifier	0.1449/0.7746	0.0203/0.9657	0.0356/0.9657
SVM Classifier	0.1989/0.769	0.2252/0.74	0.2112/0.74
K Neighbors Classifier	0.1862/0.7608	0.2576/0.6771	0.2162/0.6771
XGBoost	0.1111/0.7748	0.0101/ 0.9767	0.0186/ 0.9767
AdaBoost	0.0/0.7771	0.0/1.0	0.0/1.0
Voting Classifier	0.1216/0.7736	0.0183/0.9622	0.0317/0.9622

6 Conclusion

In this work, we evaluated the performance of various machine learning algorithms on predicting non-performing loans. We introduced different imbalance treatment techniques like CTGAN and

kmeans-SMOTE on our dataset. The model's performance was evaluated on unbalanced, under-sampled, over-sampled (using CTGAN and Kmeans-SMOTE) dataset.

The results shows that the kmeans-SMOTE over-sampling technique improved the models performance. Support Vector Machine on the kmeans-SMOTE over-sampled data achieved the highest performance.

We also selected the most important features using different feature selection methods on the dataset. The results show that Age years, total amount of income, and years employed are the most determining features. This indicates that banks should give attention to those features in addition to the current working system. For future works, we will work on enhanced feature selection methods. We encourage researchers to explore more towards such methodologies to predict defaulting loans.

7 Data Availability

The dataset used and analysed during the current study is not publicly available due to the bank's principle, but can be available from the corresponding author on reasonable request.

References

- [1] Bishnu Kumar Adhikary. Nonperforming loans in the banking sector of bangladesh: Realities and challenges. page 21.
- [2] Feng-Chia Li, Peng-Kai Wang, and Gwo-En Wang. Comparison of the primitive classifiers with extreme learning machine in credit scoring. pages 685–688.
- [3] Donald R. Heath. Prediction machines: the simple economics of artificial intelligence. 21(3):163–166. Publisher: Routledge _eprint: <https://doi.org/10.1080/15228053.2019.1673511>.
- [4] Lin Zhu, Dafeng Qiu, Daji Ergu, Cai Ying, and Kuiyi Liu. A study on predicting loan default based on the random forest algorithm. 162:503–513.
- [5] Nazeeh Ghatasheh. Business analytics using random forest trees for credit risk prediction: A comparison study. 72:19–30.
- [6] Joseph Breeden. *A Survey of Machine Learning in Credit Risk*.
- [7] Yu Li. Credit risk prediction based on machine learning methods. In *2019 14th International Conference on Computer Science Education (ICCSE)*, pages 1011–1013. ISSN: 2473-9464.
- [8] Mohammad Pradhan, Sima Akter, and Ahmed Marouf. Performance evaluation of traditional classifiers on prediction of credit recovery. pages 541–551.
- [9] Daria Granström and J. Abrahamsson. Loan default prediction using supervised machine learning algorithms.
- [10] Ying Chen and Ruirui Zhang. Research on credit card default prediction based on k-means SMOTE and BP neural network. 2021:e6618841. Publisher: Hindawi.
- [11] Başak Gültekin and Betül Erdoğan Şakar. Variable importance analysis in default prediction using machine learning techniques:. In *Proceedings of the 7th International Conference on Data Science, Technology and Applications*, pages 56–62. SCITEPRESS - Science and Technology Publications.
- [12] Laura Toloşi and Thomas Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. 27(14):1986–1994.
- [13] Sotiris Kotsiantis, Dimitris Kanellopoulos, and P. Pintelas. Data preprocessing for supervised learning. 1:111–117.
- [14] Aleksander Kołcz, Abdur Org, Chowdhury, and Joshua Alspector. Data duplication: an imbalance problem?
- [15] Jason Brownlee. Data preparation for machine learning: Data cleaning, feature selection, and data transforms in python. pages 54 – 65. 2020.

- [16] Jun-Mo Jo. Effectiveness of normalization pre-processing of big data to the machine learning performance. 14(3):547–552. Publisher: Korea Institute of Electronic Communication Science.
- [17] Felix Last, Georgios Douzas, and Fernando Bacao. Oversampling for imbalanced learning based on k-means and SMOTE. 465:1–20.
- [18] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. 16:321–357.
- [19] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN.
- [20] Rung-Ching Chen, Christine Dewi, Su-Wen Huang, and Rezzy Eko Caraka. Selecting critical features for data classification based on machine learning methods. 7(1):52.
- [21] Cheng Chen, Qingmei Zhang, Bin Yu, Zhaomin Yu, Patrick J. Lawrence, Qin Ma, and Yan Zhang. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. 123:103899.