**Abstract**

Finance sector in India comprises mainly of commercial banks that generate most of their business through the interest received on loans disbursed to people and organisations. However, credit lending comes with risk of consumer defaulting the loan i.e., inability to pay back the due sum of money agreed upon by both the parties. Ensuring that the borrower will be able to pay back the loan on the proposed terms is the biggest challenge faced by banks in today’s ever-changing world. This paper proposes to leverage the use of Machine Learning (ML) algorithms to predict which clients are more likely to default on their using historical financial data like credit and debit card data, bank account transactions, previous credit history, current bank loans and income of the client.

This study makes use of the Credit Risk Model Stability data available on Kaggle provided by the Home Credit finance provider, founded in 1997. The dataset consists of 32 csv training files and 36 testing files comprising of financial records of over 1.5 million clients, collected from internal and external sources. The research is a comparative analysis between Bayesian learning, Lazy learning, ensemble learning and gradient boosting machine learning algorithms. Gradient boosting algorithms Lightgbm, XGBoost and Adaboost outperform other algorithms and predict with upto 99% accuracy whether a client will default a loan or not. Results of this study can be scaled and applied to a real world dataset and holds the immense potential to revolutionise the financial industry.

*Keywords – credit; oan default; Lightgbm; XGBoost; Adaboost*

**Introduction**

**Literature Survey**

This section discusses in brief about some of the work that has already been done on creating Machine Learning models using various algorithms to help decrease the credit card default and help the banking authorities and financial firms select an eligible candidate with low credit risk.

Bhoomi Patel et al. in paper [1] used different algorithms like categorical boosting and compared four algorithms named Logistic Regression, Gradient Boosting, CatBoost Classifier and Random Forest to predict loan default. In paper [1] CatBoost Classifier outperformed other algorithms. It has an accuracy of 84.045%, whereas the other algorithms were 14.963%, 84.035%, and 83.514% respectively.

Mehul Madaan et al. in paper [2] compared Random Forest and Decision Tree algorithms to predict loan default. The conclusion of paper [2] is that Random Forest with an accuracy of 80% outperformed Decision Tree algorithm that gave an accuracy of 73%. The dataset that they used had biased data.

Huannan Zhang et al. in paper [3] compared Random Forest, Decision Tree and Logistic Regression algorithms to showcase the application of Random Forest Classifier in Loan default forecast. The conclusion of paper [3] is that the Random Forest Algorithm (≈ 86%) exceeds the decision tree (≈80%) and logistic regression classification (≈80%).

Abhishek Agarwal et al. in paper [4] have mentioned about Logistic Regression, Random Forest, Decision Trees, Naïve Bayes and KNN in paper [4]. The main motive of paper [4] is to compare measures between the original dataset before and after applying the Principal Component Analysis. The conclusion of paper [4] is that the accuracy of Logistic Regression was best in both the cases and Decision Tress was not affected much.

Alžbeta Bačová and František Babič in their paper [5] have used Random Forest, AdaBoost and XGBoost for predictive analysis for credit card default. The results of paper [5] showed that the performance of these algorithms was very similar.

Mohammad Ahmad Sheikh et al. in paper [6] have used Principal Component Analysis (PCA) to analyse its importance. The conclusion of paper [6] Is that the model is marginally better after applying PCA.

Lili Lai in paper [7] has compared AdaBoost, XGBoost, Random Forest, KNN and Multi-Layer Perceptron algorithms to predict loan default. The conclusion of paper [7] is that AdaBoost outperformed all the other algorithms followed by XGBoost.

Theoneste Ndayisenga in his paper [8] has mentioned the use of Logistic Regression, Decision Tree, Support Vector Machines, Random Forest, KNN, Gausian Naive Bayes, Gradient Boosting and XG Boost. The result of the analysis of these algorithms shows that Gradient Boosting (≈ 81%) is the best model to predict bank default followed by XG Boost (≈ 80%).

**Methodology**

* **Dataset**

The dataset used in the model is the Home Credit – Credit Risk Model Stability provided by the Home Credit organisation on Kaggle. It consists of 32 training files and 36 testing files. All the files are available in csv and parquet format.

* **Concatenating the files**

Merging all the data from all the necessary files for training using Polars

* **Preprocessing the data**
* **Feature Engineering**
* **Model training**
* **Testing and Evaluation**

**References**

[1] Bhoomi Patel, Harshal Patil, Jovita Hembram and Shree Jaswal **(year)** Loan Default Forecasting using Data Mining

[2] Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain and Preeti Nagrath **(year)** Loan default prediction using decision trees and random forest: A comparative study.

[3] Huannan Zhang, Yilin Bi, Wangdong Jiang1, Chuntian Luo, Shengjia Cao, Peng Guo and Jianjun Zhang **(year)** Application of Random Forest Classifier in Loan Default Forecast

[4] Abhishek Agarwal, Amit Rana, Karan Gupta, Neeta Verma **(year)** A Comparative Study and enhancement of classification techniques using Principal Component Analysis for credit card dataset.

[5] Alžbeta Bačová and František Babič **(year)** Predictive Analytics for Default of Credit Card Clients

[6] Mohammad Ahmad Sheikh, Amit Kumar Goel and Tapas Kumar **(year)** An Approach for Prediction of Loan Approval using Machine Learning Algorithm

[7] Lili Lai **(year)** Loan Default Prediction with Machine Learning Techniques

[8] Theoneste Ndayisenga **(year)** Bank Loan Approval Prediction Using Machine Learning Techniques