**A NOVEL AND EFFECTIVE MULTI-MODEL-BASED DEFAULT RISK ANALYSIS AND PREDICTION IN BUSINESS SECTOR**

**Abstract.** The precise evaluation of credit risk continues to be a crucial component of prudent decision-making and risk management in the banking and lending industries in the ever-changing financial landscape of today. While traditional methods of credit risk analysis, which frequently depend on isolated modeling methodologies, have proven effective, they might not be able to fully represent the complexity of today's financial settings. The need for methods that can provide increased predictive power and adaptability grows as markets change and become more complicated. To address this problem, ensemble approaches have surfaced as a strong contender, offering a framework that combines the predictive power of several models into a coherent whole. This study uses a range of machine learning algorithms, including XGBoost, CatBoost, Decision Tree, Logistic Regression, KNN, and Random Forest to explore the potential in the field of credit risk analysis. By leveraging the unique properties of many algorithms within an ensemble framework, the objective is to improve forecast accuracy while also strengthening the robustness and adaptability of default risk assessment approaches. This introduction discusses how ensemble approaches can revolutionize credit risk analysis and establish the groundwork for a full discussion of them. It also offers insights into practical implementation considerations and empirical validations.

**Keywords:** Machine Learning, Analysis of credit risk, Ensemble learning

**1 INTRODUCTION**

One of the key tenets of finance is credit risk, which is the possibility of losses for investors and financiers due to uncertainties regarding borrowers' capacity to repay loans. Multiple kinds of risks are just a few of the numerous variations that might occur, each with its own set of challenges. Creditworthiness is determined by several criteria, including income, credit history, and economic conditions, in addition to macroeconomic indices like GDP growth and unemployment rates that affect borrowers' ability to repay loans [1].

As different businesses are impacted differently by laws and economic movements, sectoral and industry risks contribute to complexity. Market fluctuations influence the vulnerability of issuers and borrowers, raising credit risk. Diversification, collateralization, credit derivatives like CDS, and sophisticated risk assessment models are examples of risk reduction strategies. Credit scoring and risk assessment have been improved by technological developments, which now include more parameters for better accuracy [2]. The use of neural networks, logistic regression (LR), and other applications can all be useful in credit risk analysis. Research indicates that adding experts' contributions to machine learning models may assist minimize bias, which emphasizes the importance of ethical lending methods. The primary objective of multidisciplinary research is to improve the efficacy of models, notably in spotting anomalous trends and preventing unscrupulous lending [3].

**2 LITERATURE REVIEW**

In reaction to the advent of big data technologies, data accessibility, and processing capacity, banks and lending organizations are revamping their commercial models. Risk nursing, model consistency, and actual loan dispensation are critical mechanisms of transparency and decision-making. The study uses machine and deep learning models to create binary classifiers that use the 10 most important features from each model to estimate the probability of a loan default [8]. Credit risk analysis has become more and more important since the collapse of subprime mortgages in 2007 and the global financial crisis of 2008. This research employs a multitude of models, to perform credit risk analysis on an overwhelming peer-to-peer loan dataset comprising one million observations. SVM is a highly accurate method [13], but surprisingly, decision trees, logistic regression, MLP, PNN, and deep learning produce the most precise outcomes [1]. As different businesses are impacted differently by laws and economic movements, sectoral and industry risks contribute to complexity. Market fluctuations influence the vulnerability of issuers and borrowers, raising credit risk. Diversification, collateralization, credit derivatives like CDS, and sophisticated risk assessment models are examples of risk reduction strategies. Credit scoring and risk assessment have been improved by technological developments, which now include more parameters for better accuracy.

Reapplying machine learning techniques to a dataset that differs from standard bank loans and has a substantially higher number of observations is one of the primary contributions. SVM and decision tree results are consistent with earlier research. Based on the peer-to-peer dataset, MLP and logistic regression perform equally well, adding to the controversy over whether MLP is superior to logistic regression. PNN's poor performance begs the question of how well it can handle data imbalance [2].

The results of this study demonstrate the inadequacies of conventional credit risk analysis techniques: on peer-to-peer loan data, logistic regression and MLP yield equal results, but deep learning performs badly. These results highlight the shortcomings of single models and the need for ensemble approaches to increase credit risk assessment's forecast accuracy and robustness.

**3. PROPOSED SYSTEM AND ARCHITECTURE**

In our initial credit risk study, we used a variety of models, such as Decision Trees, XGBoost, CatBoost, Random Forests, K-nearest neighbor, and Logistic Regression (LR), to evaluate the dataset consisting of 12 columns. Selecting the top three models with the best performance indicators for additional development, we assessed each model according to its capacity to predict credit risk. Using different learning rates, tree depths, and regularization terms, we optimized these models' hyperparameters to improve their predictive potential.



**Figure 1.** The Architecture of the Proposed System

The proposed approach uses a traditional machine learning methodology. Several machine learning methods, including XGBoost, CatBoost, K-nearest neighbor, Random forests, Decision Trees, and Logistic Regression, are used in the process. The aforementioned algorithms are trained using a training dataset, and their performance is subsequently evaluated through testing. This method allows models to be compared in the context of credit risk analysis based on how well they predict particular outcomes, such as failing to repay loans.

**4. RESULT AND ANALYSIS**

Default risk analysis using ensemble models, an ensemble regression modeling of the credit risk analysis of a dataset. This is structured as follows:

1. Exploratory Analysis

2. Data Visualization

3. Data Preprocessing

4. Evaluation of Models

5. Models Optimization and Ensembling

6. Final Remarks

**5.1 Exploratory Analysis**

Essential libraries for data manipulation, visualization, and modeling are imported to support the credit risk analysis task. And upload the dataset. The dataset is composed of 12 columns, A description of the columns is reported in the table here below:

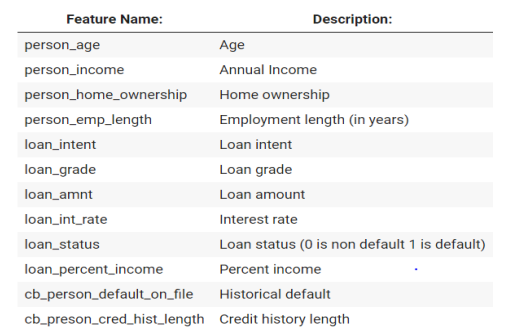
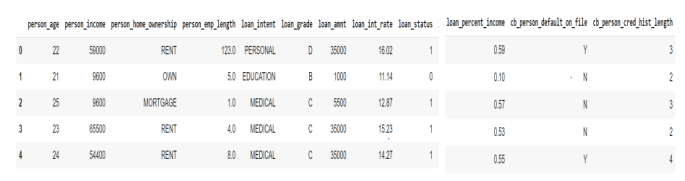


Figure 2. Dataset Information

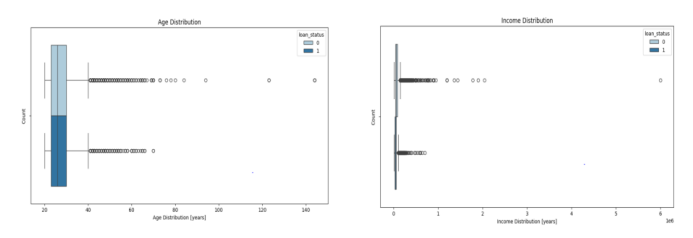
Here are the first 5 rows of the dataset along with the corresponding headers.

Table 1. First 5 rows of data

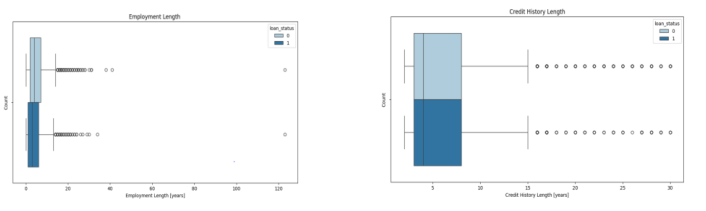


**5.2 Data Visualization**

Data visualization transforms raw data into visual representations like charts and graphs, facilitating the discovery of patterns and trends. It aids in understanding complex datasets, identifying outliers, and exploring relationships between variables. Effective visualization enhances the communication of insights and supports data-driven decision-making across diverse domains. Here below we plot all the features separately in order to show any dependence on the target.



**Figure 3.** Age Distribution and Income distribution

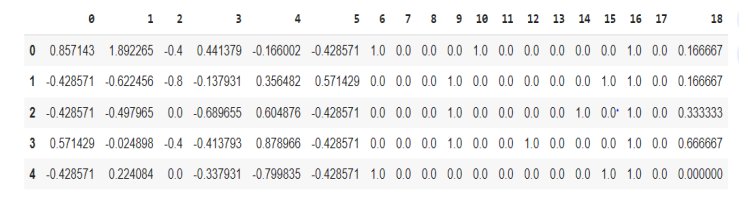


**Figure 4.** Employment Length and Credit History Length

**5.3 Data Cleaning and Preprocessing**

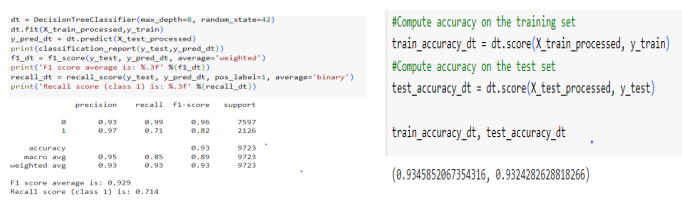
The 4 boxplots above show a significant number of outliers in person\_age, person\_income, person\_emp\_length, and cb\_person\_cred\_hist\_length. In particular, there are a few outliers in person\_age and person\_emp\_length 27 above 122 years. First of all, we check the number of Nans. person\_emp\_length has 887 Nans and loan\_int\_rate has 3094 Nans. We are going to substitute them later on using the average of all samples. We have also found 165 duplicate rows in the dataset.

**Table 2.** Dataset after preprocessing



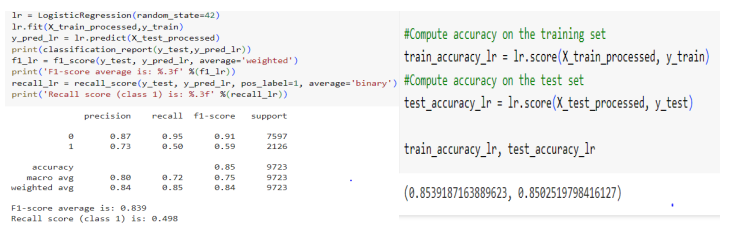
**5.4 Evaluation of Models**

**5.4.1 Decision Tree**



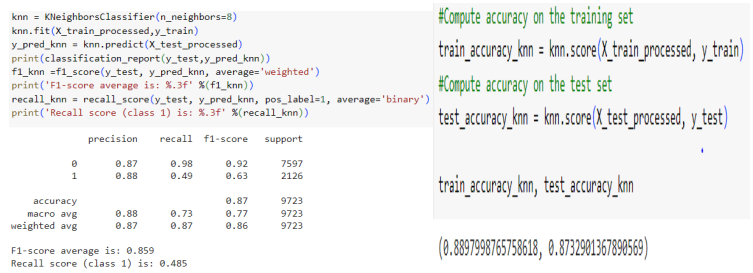
The above Figures depict the accuracy of the Decision Tree model's predictions on both the training and testing datasets. It gives 93% accuracy.

**5.4.2 Logistic Regression**

****

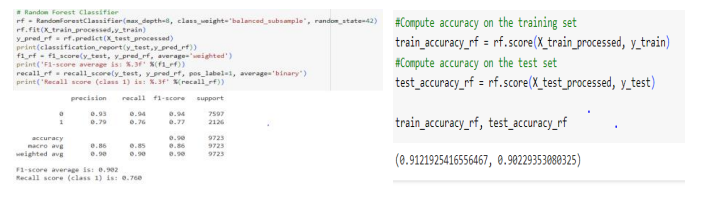
The above Figs depict the accuracy of the Logistic Regression model's predictions on both the training and testing datasets. It gives 85% accuracy.

**5.4.3 KNN Classifier**



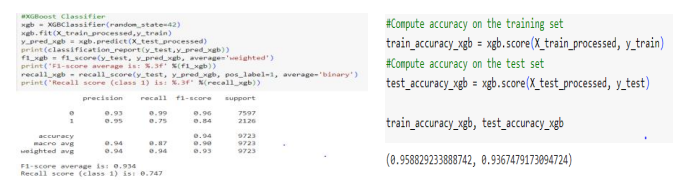
The above Figures depict the accuracy of the K-Nearest Neighbor model's predictions on both the training and testing datasets. It gives 87% accuracy.

**5.4.4 Random Forest Classifier**

****

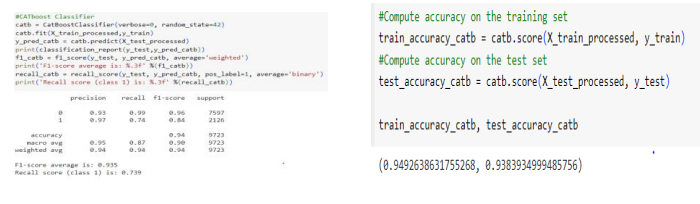
The above Figs depict the accuracy of the Random Forest Classifier model's predictions on both the training and testing datasets. It gives 90% accuracy.

**5.4.5 XG Boost Classifier**

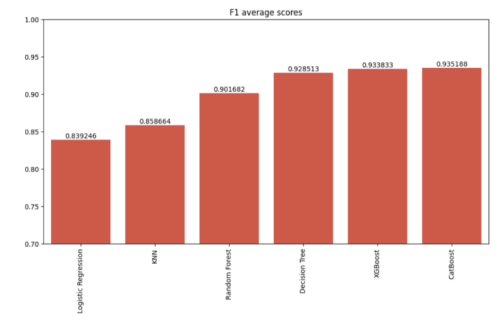
****

The above Figs depict the accuracy of the XG Boost Classifier model's predictions on both the training and testing datasets. It gives 94% accuracy.

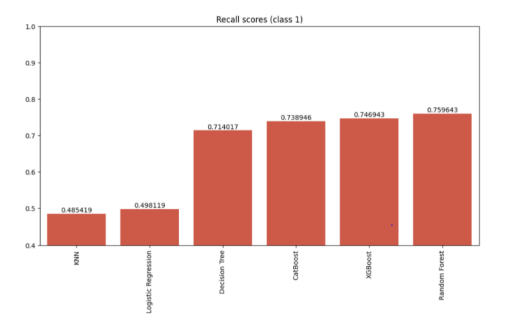
**5.4.6 CatBoost Classifier**

****

The above Figs depict the accuracy of the CatBoost Classifier model's predictions on both the training and testing datasets. It gives 94% accuracy.



**Figure 5.** F1 Scores Before Optimization



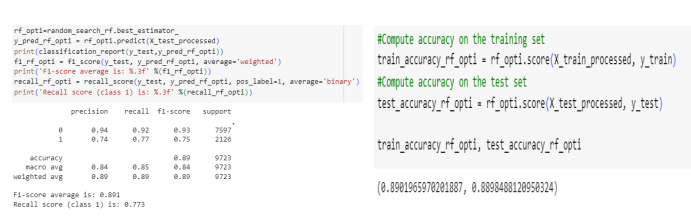
**Figure 6.** Recall Scores Before Optimization

Thus, XG Boost, Random Forest, and CatBoost algorithms have demonstrated superior accuracy based on the above analysis.

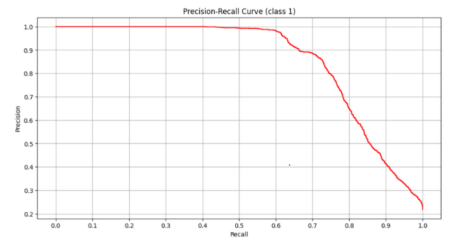
**5.5 Model Optimization and Ensemble**

Here below we decide to perform a hyperparameter tuning with a Randomized Search on the three best-performing models trained so far i.e. Random Forest, XGBoost, and CatBoost. At the end of the optimization, we train also a voting classifier with both soft and hard options. All three models are optimized based on the recall scoring since we are looking for a model that is able to predict default customers with a low number of false negatives. Indeed we care less about precision than recall since, in this particular problem, it is key to identify correctly the biggest number of customers with high default risk.

**5.5.1 Random Forest-Hyperparameters Optimization**

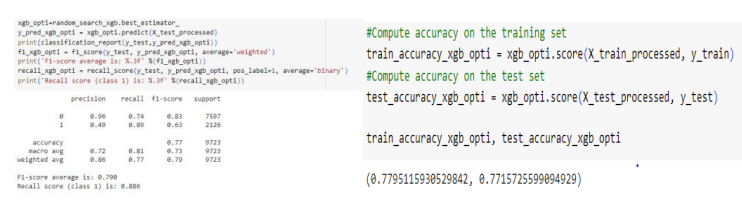
****

The above code depicts the accuracy of the Random Forest Classifier model's predictions on both the training and testing datasets after the optimization of the model. It gives 89% accuracy.

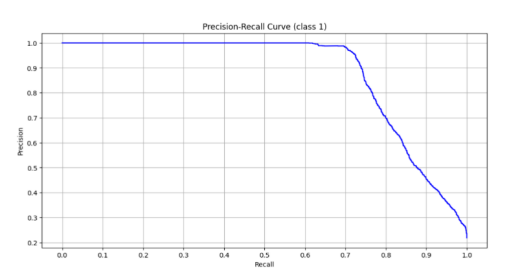


**Figure 7.** Precision-Recall Curve of Random Forest Classifier

**5.5.2 XG Boost-Hyperparameters Optimization**

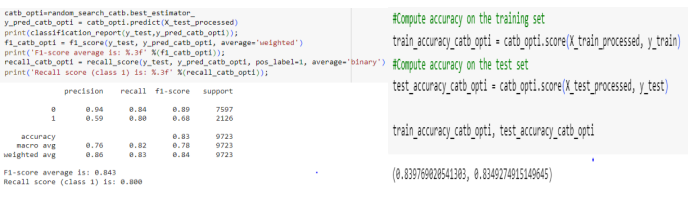
****

The above Figures depict the accuracy of the XG Boost Classifier model's predictions on both the training and testing datasets after the optimization of the model. It gives 77% accuracy.

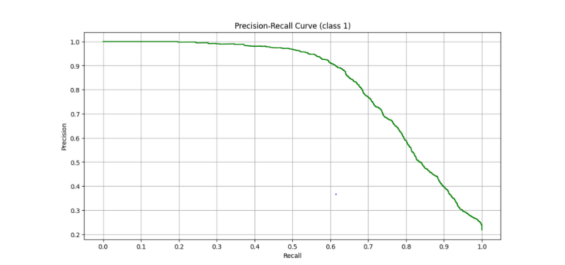


**Figure 8.** Precision-Recall Curve of XG Boost

**5.5.3 CatBoost-Hyperparameters Optimization**

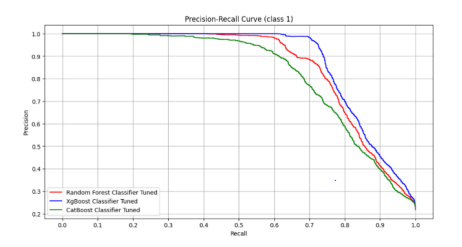
****

The above Figs depict the accuracy of the CatBoost Classifier model's predictions on both the training and testing datasets after the optimization of the model. It gives 83% accuracy.



**Figure 9.** Precision-Recall Curve of CatBoost

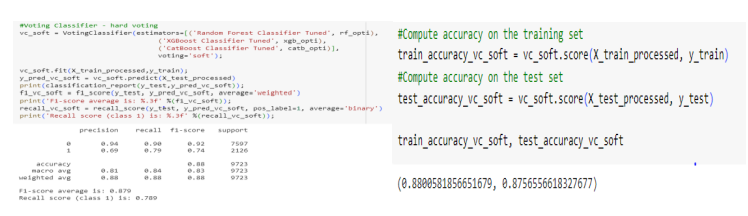
Here below we compare the precision recall curves of the three tuned ensemble models. We can see that XGBoost is a clear winner with a higher Area Under the Curve.



**Figure 10.** Comparision of curves of Random Forest,XG Boost,CatBoost

**5.5.4 Soft Voting**

In soft voting, each base model provides a probability or confidence score for each class (for classification) or a predicted value (for regression). The final prediction is determined by averaging the probabilities or predicted values across all base models, and then selecting the class with the highest average probability or the average predicted value.

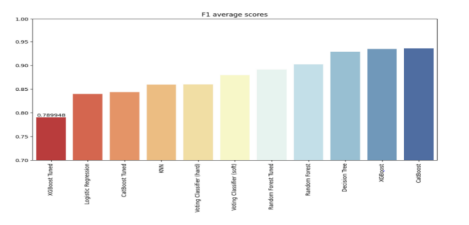


**5.5.5 Hard Voting**

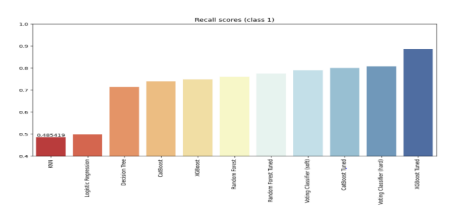
In hard voting, each base model (also called a "weak learner" or "base classifier/regressor") makes its own prediction. The final prediction is determined by a majority vote. That is, the class that receives the most votes from the individual models is chosen as the final prediction.



**5.6 Final Remark**

****

**Figure 11.** F1 Scores after Optimization



**Figure 12.** Recall Scores after Optimization

The optimized XGBoost model is the clear winner scoring the best recall score on class 1 (82%) and a high f1 number (88%).

**5 CONCLUSIONS**

We thoroughly assessed eleven machine learning models in our credit risk study, giving comparisons based on F1 score and recall metrics priority. After optimization, the best model was the XGBoost model, which had an amazing recall rate of 82.5% for high-risk cases. It showed exceptional capacity to detect true positive cases, exhibiting a precise balance between precision and recall with a weighted average F1 score of 88%. Among all the models examined, the XGBoost model had the best precision-recall curve. Although XGBoost demonstrated superior recall, the Voting Classifier (soft) also demonstrated great performance, making it a strong substitute for situations that call for a more balanced model.

**REFERENCES**

[1]. S. Kalaycı, M. Kawasaki, and S. Arslan, "Credit risk analysis using machine learning algorithms," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404353

[2]. Emmanuel, I., Sun, Y. & Wang, Z. A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. J Big Data 11, 23 (2024)

[3]. Bitetto, A., Cerchiello, P., Filomeni, S., Tanda, A., & Tarantino, B. (2023). Machine learning and credit risk: Empirical evidence from small- and mid-sized businesses. \*Socio-Economic Planning Sciences\*, 90, 101746. doi: 10.1016/j.seps.2023.101746.

[4]. P. M. Addo, D.Guegan and B. Hassani, Credit Risk Analysis Using Machine and Deep Learning Models,Risk, 6(2), 2018, p. 38

[5]. J. Lavanya, M. Ramesh, J. Sravan Kumar, G. Rajaramesh and Subhani Shaik,” Hate Speech Detection Using Decision Tree Algorithm”, Journal of Advances in Mathematics and Computer Science, Volume 38, Issue 8, Page 66-75, June-2023.

[6]. Sujan Reddy, Ms. Renu Sri and Subhani Shaik,” Sentimental Analysis using Logistic Regression”, International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.

[7]. Subhani Shaik and Dr. Uppu Ravibabu, "Detection and Classification of Power Quality Disturbances Using curvelet Transform and Support Vector Machines", in the 5th IEEE International Conference on Information Communication and Embedded System (ICICES-2016) at S.A Engineering college, Chennai, India on 25th -26th, February 2016.

[8]. KP Surya Teja, Vigneswara Reddy and Subhani Shaik,” Flight Delay Prediction Using Machine Learning Algorithm XGBoost”, Jour of Adv Research in Dynamical & Control Systems, Vol. 11, No. 5, 2019.

[9]. Subhani Shaik and Dr. Uppu Ravibabu “Classification of EMG Signal Analysis based on Curvelet Transform and Random Forest tree Method” Paper selected for Journal of Theoretical and Applied Information Technology (JATIT), Vol. 95, December 2017.

[10]. Dr. R. Vijaya Kumar Reddy, Dr. Shaik Subhani, Dr. G. Rajesh Chandra, Dr. B. Srinivasa Rao,” Breast Cancer Prediction using Classification Techniques”, International Journal of Emerging Trends in Engineering Research, Vol. 8, No.9,2020.