Title

An Explainable Ensemble Learning Framework for Credit Risk Prediction Using Weighted Soft Voting

Authors

Anuj Kumar Singh[1], Priya Verma[2], Arjun Singh[3]

Author Affiliations

Department, Address

Corresponding Author

Anuj Kumar Singh
Email: anujkumarsingh@email.com

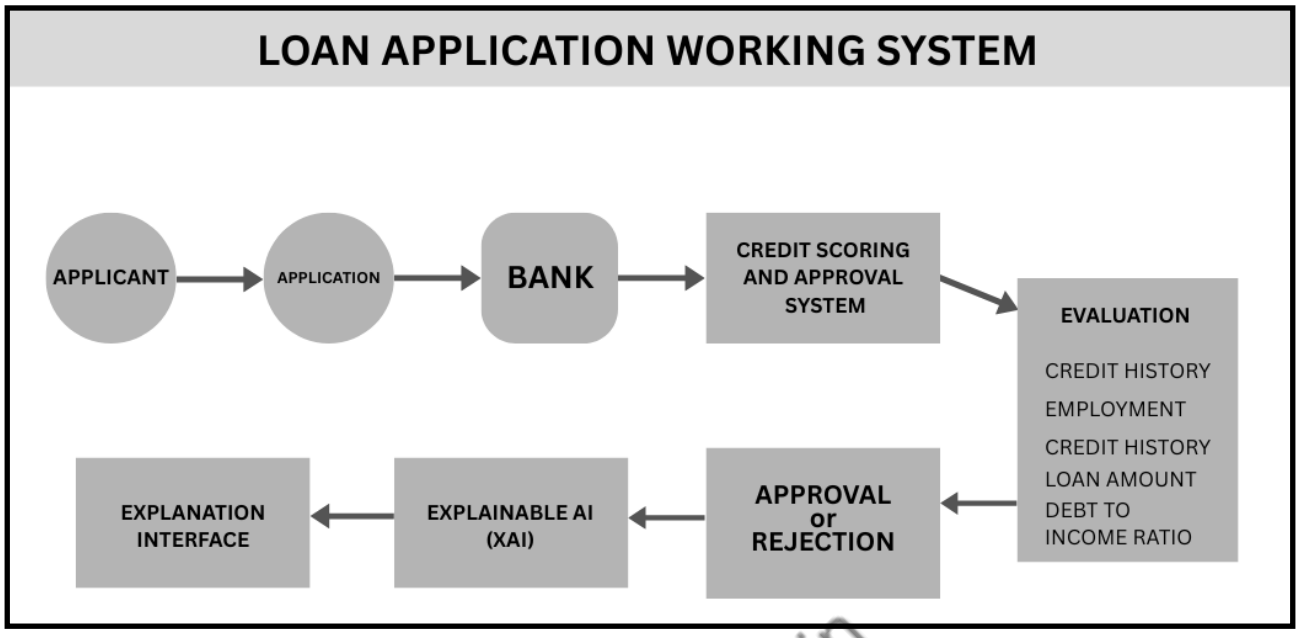
Abstract

Credit risk prediction is a critical function in modern financial system, especially when the reliance on digital lending platforms and automated decision has been increased so much. Financial institutions must evaluate large number of loan applications so that they can minimize default. The models that we haven been used yet is Logistic Regression as it is very simple to use and its interpretability feature. The logical regression model provides clear relationship between input variables of applicant and outcomes, which make it suitable for agencies. However, they still can not determine the complex relationship between the variables and applicant especially when dealing with large data which has even some hidden relations between applicants and their data.
In contrast with this new machine learning models like Random Forest, XGBoost and Neural Networks are more able in prediction. These new models can determine the hidden patterns and relations in data with the applicant, which lead to more accurate prediction as a result accuracy is improved. However these models operate as black box system which means it is difficult to interpret their decision. Here comes the major challenge of financial application where each decision must be transparent and explainable to everyone.
In this study, we proposed an ensemble learning framework that combines all four models using a soft voting mechanism, Our goal is to leverage the strengths of different models while reducing their individual limitations, Additionally after that we integrated Explainable Artificial Intelligence techniques, most specificallY SHAP which will provide clarity of every decision taken.
The framework that we proposed in tested on both synthetic dataset as well as German Credit dataset, The model got an accuracy of 92% as a result it outperforms individual models. SHAP further gives us additional insights which make the model more suitable for real world deployment.

Introduction

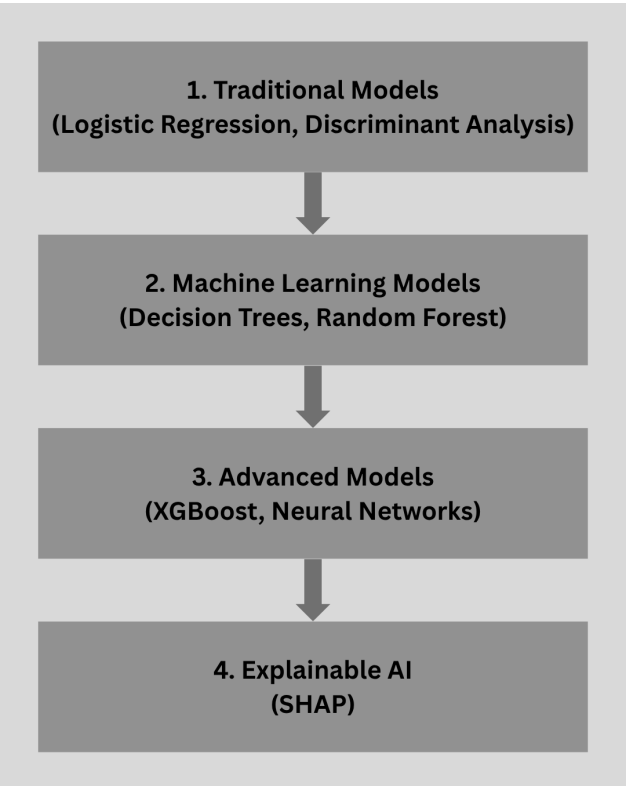
The financial sector has gone to very much transformation over the past decade, the advancements in technologies and the data has been increased so much. One of the most important thing which is affected by this transformation is risk assessment in credit system. Nowadays institutions have to process thousands of loan applications daily, and with so much increase in number of applications

manual evaluation of them is almost impossible. As a result these type of ML models have become very essential.



The prediction involves how a borrower will default if he got a loan. This decision has direct impact on the profit or institutions as well as their financial stability. Correct prediction reduce losses for the institutions and even improve decision making whereas inaccurate models lead to various financial risks.

The old credit scoring models were dependent on statistical techniques such as logical regression and discriminant analysis. However the financial data is becoming more and more complicated and commonly contains dynamic interaction between variables. As a result, simple linear model may fail to catch all the required relationships.These new ML models provide very flexible modeling



abilities and they perform very good in credit prediction. Composite learning approaches combines multiple algorithms to produce more trustworthy prediction than any single model.
Although even after having these advantages, machine learning models often lack explainability.In financial environments where transparency and accountability are the two most important things, this deficiency can limit acceptance.
The goal of this research is to make an explainable AI framework that combines ensemble machine learning models with the SHAP-based explanation which will improve accuracy as well as transparency in credit risk prediction. The main goal is to build a system which is both accurate and interpretable.
The remainder of this paper is organized as follows: Section 2 presents the problem statement, Section 3 reviews related literature, Section 4 describes the proposed methodology, Section 5 discusses datasets and evaluation metrics, Section 6 presents experimental results, and Section 7 concludes the study.

Problem Statement

There are so much advancements in ML but still credit risk prediction system faces lot of challenge. The most important problem is the trade of between accuracy and interpretability. The models which perform very good like XGBoost and Neural Network provide more accuracy than rest but these models lack transparency. Where on the other hands the old models are very much interpretable but they may perform very badly on complex databases.

Model	Accuracy	Interpretability
Logistic Regression	Medium	High
Random Forest	High	Medium
XGBoost	Very High	Low
Neural Network	Very High	Very Low

Literature Review

The credit risk prediction has been studied in a large scale in financial analytics and machine learning researches. The earlier credit scoring systems were primarily dependent on those statistical models such as logistic regression because they were very much understandable and were very easy to implement. Decision trees and discriminant analysis were also applies as they helped in categorizing the borrower in different risk categories.

Now as computational power advances and the availability of bigger financial datasets, researchers have been started exploring machine learning models for credit risk management. Random Forest is one of the ML model which became a popular ensemble learning technique because it combines multiple decision trees that were trained on random parts of data which help in generalization and it reduces overtraining. Gradient Boosting algorithm such as XGBoost has improve predictive performance by repeatedly minimizing the errors of model and optimizing classification accuracy of the model.

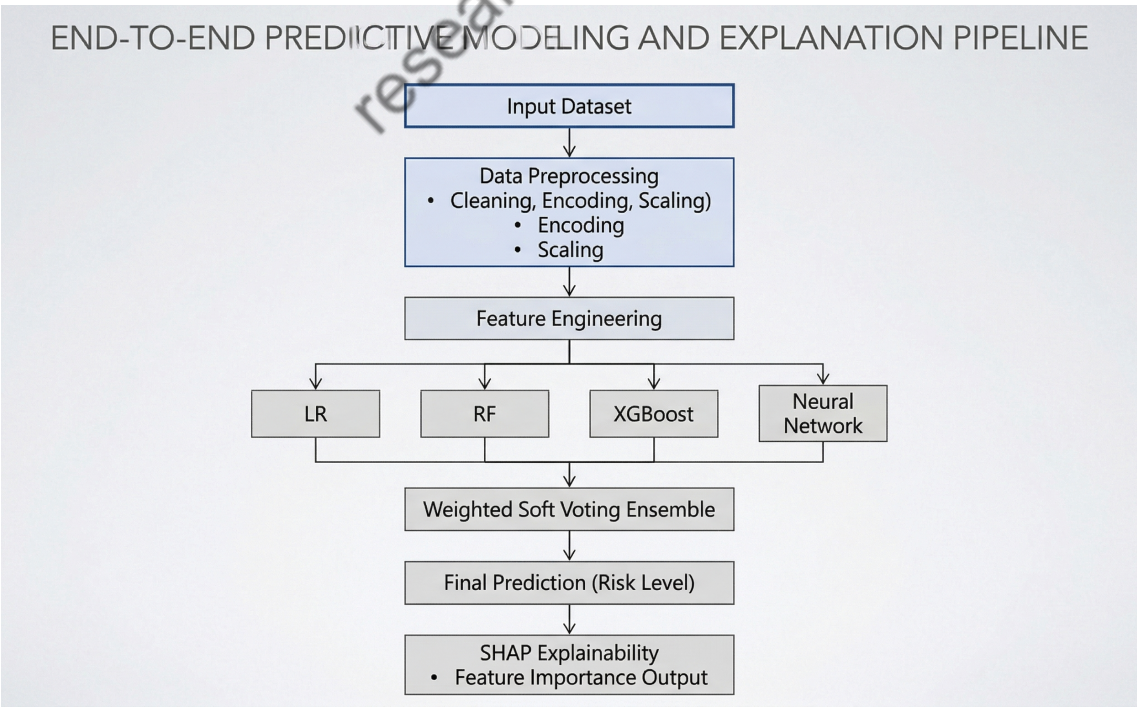
Neural Networks have also been applied to these tasks because they can determine the relation between the borrower attributes and default risk for the application.
Multi Layer perceptron architectures have been used widely to predict loan defaults and even for identifying financial risk patterns. While the models play a vital role in improving the accuracy of prediction, the commonly suffer from transparency and understandability.

To deal with these limitations of ML models, researches have introduced some explainability techniques such as LIME and SHAP. SHAP uses Shapely values which are derived from cooperative game theory to measure how each feature plays role in a prediction. This technique gives both global and local explanation. Recent researches highlights that the explainable AI is very important in financial systems where algorithmic decisions should remain transparent and accountable.

Methodology

The framework that we are suggesting will combine four different machine learning algorithms together to predict borrower credit risk. These four models are Logistic Regression, Random Forest, XGBoost and a Multi-Layer Perceptron Neural Network. Each model analyzes attributes of the borrower and they autonomously predicts that if the borrower represents a high or low credit risk.

Logistic Regression is being used a starting model because it has been used in huge amount in tradition credit scoring systems. Random Forest helps in improving the prediction accuracy as it combines multiple decision trees that were trained on random parts of the dataset. XGBoost applies gradient boosting techniques which successively improves the model prediction as it minimizes classification errors of the model. The relationships between the probability of risk and borrower attributes is determined by neural network.



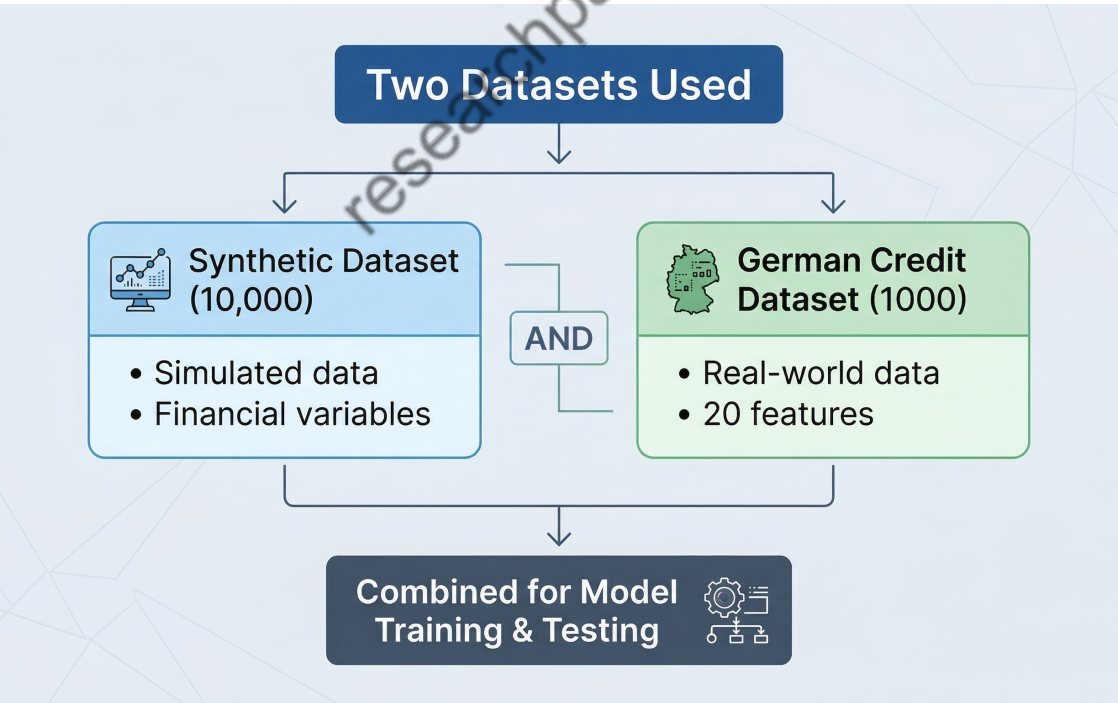
Before training any model, we need to treat the data where numerical variables are standardized to ensure stable scaling and categorical features are encoded using suitable encoding techniques. The dataset is then divided into two parts training and testing subsets which helps in evaluating model performance objectively.

Then we will combine predictions from all four model using collective voting strategy. The final risk is determined by the majority prediction given by models individually. This kind of approach will improve stability and reduces the influence of any single algorithm on final decision.

SHAP values are being used to obtain model transparency. SHAP calculates how much every features is contributing in determining the increase or decrease in predicted risk level, and it allows analysts to identify the most important factors which influences credit decisions.

Dataset Description

To know how good explainable artificial intelligence is, we will be using two different datasets in this study. Using these two datasets will help us in ensuring that the model we developed, it perform good under different conditions and scenarios and it is not limited to a single type of data. We generated one dataset artificially to represent modern financial scenarios of market, while we took second dataset from a public repository which almost everyone uses for credit risk research. These two different datasets helps our system to be tested in both controlled and practical environments. This improves the reliability on the experimental results and helps us to demonstrate that the framework can be used in different types of financial data.



A synthetic dataset with 10,000 has been used. We made the synthetic data by imitating real financial conditions. Each record represents a loan applicant and his some attributes. By generating 10,000 samples this dataset helps the ML model to learn patterns.

The synthetic data contains a number of attributes related to the loan applications. These attributes include loan amount, loan duration and loan purpose as these represents all the characteristics of the loan request. These dataset include a lot more information about the applicant like annual income, applicants monthly debt, estimated home value. These will help in determining the financial capacity of the value.

The second dataset that we are using in this research is a German Credit dataset, which is obtained from the UCI ML Repository. This dataset has been used most widely in similar researches. It contains 1000 records of loan applicants, every applicant has 20 different attribute. These attributes includes variables like credit history, whether he is employed or not, loan purpose and lot other things. Each record is then labelled as good or bad for credit risk and with the help of that ML models can differentiate between risky and non risky borrowers.

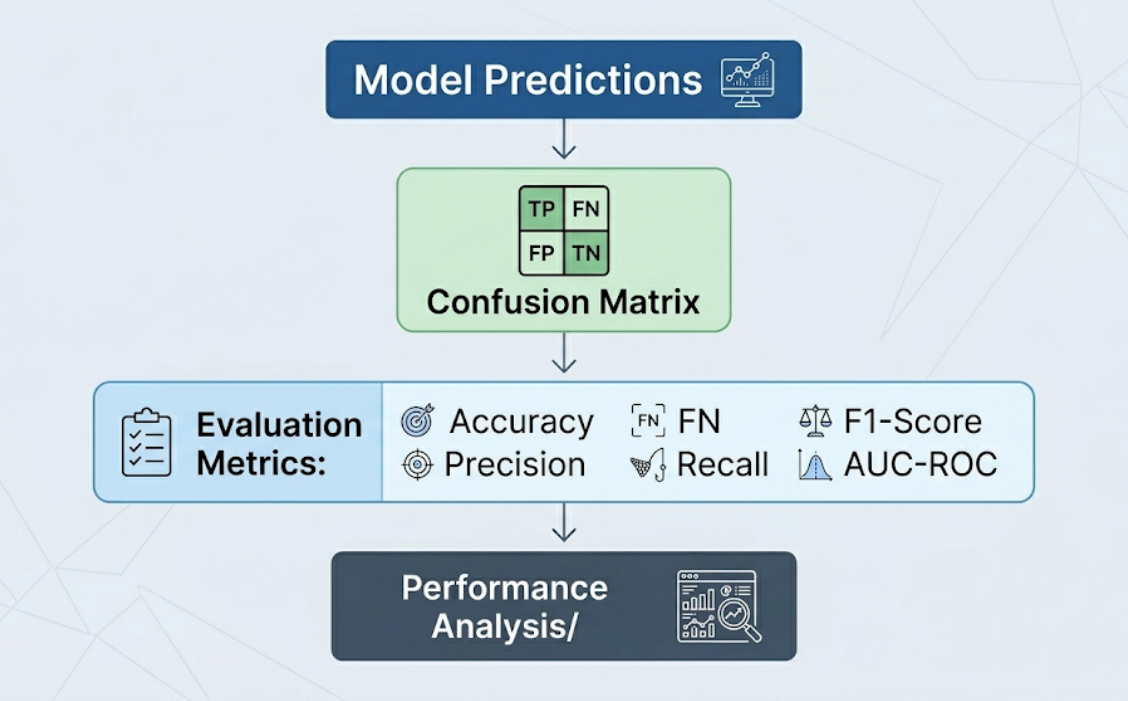
Using two different dataset will provide us several advantages for finding the result. The synthetic data has a large number of sample and it includes modern variables where the German Credit dataset is a standardized benchmark which allow to compare with previous research studies. By using the models on these datasets we'll get to know that our proposed framework performs good for different data and feature sets.

Feature	Synthetic Dataset	German Credit Dataset
Type	Artificial	Real-world
Samples	10,000	1,000
Source	Generated	UCI Repository
Features	Financial attributes (income, loan, etc.)	20 attributes
Purpose	Simulate modern conditions	Benchmark comparison

Evaluation Metrics

To check how effective our system is, we have used some performance metrics. These helped us in measuring how accurately the ML models classifies the borrowers in different risk categories. In financial risk prediction, measuring overall accuracy is not sufficient as if any data gets misclassified into some other category the consequences will be serious. Like if a high risk borrower is classified into low risk and if his loan gets approved then it'll result in a default. Therefore we are using multiple evaluation metrics so that we can evaluate very precisely.
The first metric we used in our study is accuracy, which indicates how many predictions made by the model are correct out of all the cases considered. Accuracy is worked out by taking the number of correct predictions and comparing it to the total number of observations in the dataset. Accuracy gives a straightforward understanding of how well the model is performing and is often used as a baseline metric.

Precision is the second metric that we are using in evaluation. It show the number of correctly predicted positive cases among all cases. Precision tell us how reliable model is when it identifies borrowers as high risk. High precision means most borrowers belong to high risk category. It plays important role in reducing false alarms and ensuring that risk warnings which are generated can be trusted.



We used another very important performance metric which is Recall also known as sensitivity. It measures only the positive cases which models identified correctly. In our context, it shows how effectively out model identifies borrowers who are at very high risk. A high recall value means that our system has detected most risky applicants, and it helps the financial institution to avoid approving loans for borrowers who are very risky to be default.

The F1-score is also used to evaluate the performance of the proposed models. It is calculated as harmonic mean of precision and recall, which means it provides balanced measure of two above metrics. It is very useful where we need to balance between identifying risky borrowers and avoiding incorrect classifications of borrowers. In our system both false positive and negative have some consequences, so this score is trusted for that and it is more reliable performance indicator than accuracy alone.

In addition to the previously mentioned metrics, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was used to evaluate model performance A model with higher AUC value is very effective in separating risky applicants and safe applicants.

By using these different kind of metrics, we are providing a very broad analysis of our model performance. Here the first metric accuracy measures correctness of the model, precision tells how much we can rely on positive predictions, and recall detects risky borrowers and at the end F1-score balances those two metrics. Then AUC-ROC metric again checks the model's ability to differentiate between different kind of risk categories. Together all the metrics shows completely how our frameworks performs in predicting financial risks.

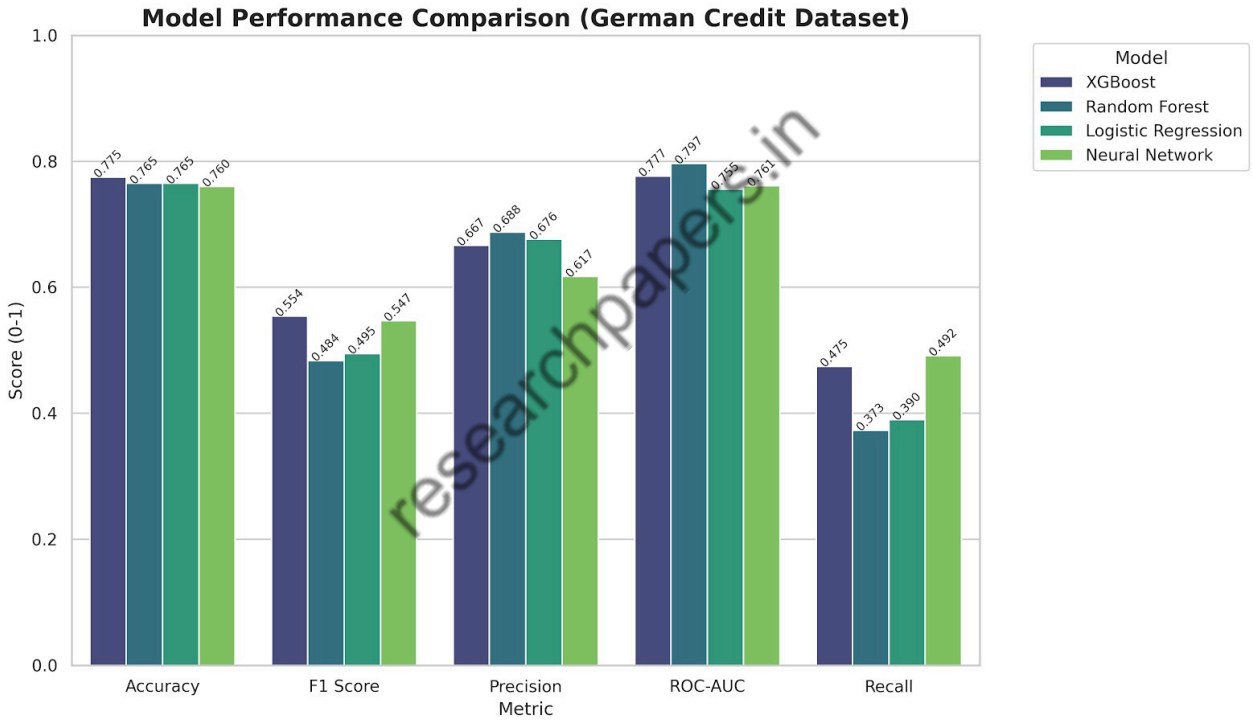
Experimental Results

After preparation of the datasets and defining these evaluation metrics we conducted experiments in order to evaluate the performance of our model. Four machine learning algorithms were implemented in this study: Logistic Regression, Random Forest, XGBoost, and a Multi-Layer

Perceptron Neural Network. We trained every model based on the datasets we had and later we tested it on another separate database in order to know how accurately it can predict. The purpose of using multiple algorithms was to compare their strengths and weaknesses in credit risk prediction and to observe how ensemble decision making can improve reliability.

Before training these models we processed the dataset to make sure that the data we are inputting is suitable for ML Algorithms. The numerical variables were later normalized in order to obtain consistent growth across different features. The attributes such as loan purpose and employment categories were later converted into numerical representation so that ML models could process them effectively. After this preprocessing of data we divided it into training and testing sets so that the model can be trained on unseen data too.

The logistic regression model is acting as baseline for all models in this experiment. It is used very widely in traditional credit scoring system due to its simplicity and its transparency. Logistic Regression model estimates the default probability of borrower by analyzing the relationship



between target classification and the input variables. Although it provides transparent decision but its performance will be limited if there comes a complex relationship between borrower and variables. In our experiment Logistic Regression got reasonable accuracy but perform lower as compare to other advanced ML models.

Model Performance Summary Table

Accuracy	F1 Score	Precision	ROC-AUC	Recall	model
0.775	0.5544554455445545	0.6666666666666666	0.7767760548142806	0.4745762711864407	XGBoost
0.765	0.4835164835164835	0.6875	0.7965500661137155	0.3728813559322034	Random Forest
0.765	0.4946236559139785	0.6764705882352942	0.7554994590695996	0.3898305084745763	Logistic Regression
0.76	0.5471698113207547	0.6170212765957447	0.7610289698281043	0.4915254237288136	Neural Network

Now the random forest model it had improved the prediction performance as compares to Regression one. Random forest builds multiple decision trees and combines all the productions and produce final classification. Every tree is trained on a random part of data and it helps in reducing overfitting and improving generalization. In credit risk prediction system it successfully captures all the interaction between borrower attributes like borrowers income level, credit history, and debt ratio. As a final result this model produce higher accuracy and very strong recall values while identifying high risk borrowers.

Among the models we've evaluated yet, XGBoost gives the highest overall prediction performance, It is a gradient boosting algorithm, it builds decision trees sequentially and even minimize the prediction errors in each case. In the experiment we conducted, XGBoost was able to capture relations between borrower attributes and default risk. Its performance was strong while differentiating between high risk and low risk applicants and as a result we got high accuracy and good AUC-ROC scores.

Model Performance Summary Table (Enterprise)

Accuracy	F1 Score	Precision	ROC-AUC	Recall	model
0.881	0.8033057851239669	0.8307692307692308	0.953145018181818	0.7776	XGBoost
0.883	0.8020304568527918	0.8509874326750448	0.9491648	0.7584	Random Forest
0.798	0.6474694589877836	0.7120921305182342	0.8637928727272727	0.5936	Logistic Regression
0.831	0.7201986754966887	0.7461406518010292	0.8995816727272727	0.696	Neural Network

The neural network has performed very well in our experiment. They are capable of modeling relationships and they can help in identifying complex patterns in data. Overall we got the result which shows that the model we are proposing is very effective for risk prediction. These all models combining together produced reliable results and on the other hand SHAP maintained the transparency as it explained everything.

Discussion

The results from our experiments show clear patterns of credit risk prediction by ML models. Another thing which is important is that how this model performs better than the traditional ones. While Logistic Regression help in understanding us the relation in a very easy way but it misses complex patterns in data which is often hidden whereas other models like Random Forest and XGBoost performs better in that case as they can identify the hidden relations too. And this feature of XGBoost and Random Forest provides us more accurate prediction.

Another important thing is dependency on one model is not always reliable and not suggested too. As every model has its own limitations. For example Logistic Regression is easy to use but it needs a lot of resources while neural network can capture complex patterns hidden in data but it lack transparency. So when we combine multiple models we get benefit of strength of all models. It helps us to create a more balanced prediction system which is very much dependable too.

This collective model decision system helped in improving the reliability on system. When multiple models show same result, then we can trust the result more.Especially in finance where one wrong decision will lead to huge loss.

So, by combining all prediction our framework ensures that we are not biased on prediction of any single model. Explainable AI improves the framework more by making predictions easy to understand. And on the other hand SHAP analysis we get to know which factors influenced the prediction.

Reference

1.Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

3. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765–4774.

4. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence*, 3, 26.

5. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: Definitions, methods, and applications. *Proceedings of the National Academy of Sciences*, 116(44), 22071–22080.