Financial Risk Prediction

Devyansh Chaudhary
devyansh22156@iiitd.ac.in

Dhawal Bansal
dhawal22159@iiitd.ac.in

Dhawal Garg
dhawal22160@iiitd.ac.in

Manas Chhabra
manas22276@iiitd.ac.in

Abstract

Empowering individuals through financial literacy is crucial, as many people may lack the tools or knowledge to accurately assess their financial health. This project aims to provide a user-friendly tool for assessing financial risk, helping individuals take control of their finances. By understanding their financial stability, users can make informed decisions about savings, investments, loans, and spending habits. Such a tool can reduce financial stress—a common issue that affects quality of life—by enabling individuals to identify potential financial problems early and take proactive measures. Ultimately, this promotes financial well-being and enhances overall quality of life.[GitHub Link]

1. Introduction

Financial risk prediction is a critical component in managing both personal and corporate finances. In today's unpredictable economic climate, the ability to assess potential risks before they occur allows individuals and organizations to make more informed decisions about investments, savings, and loans. The goal of this project is to develop a machine learning-based model that can predict financial risks based on various financial indicators. By utilizing historical data and statistical methods, we aim to provide an accurate and scalable solution that empowers users to make proactive financial decisions.

2. Literature Review

In this section, we discuss two research papers that contribute significantly to the field of financial risk prediction using ensemble methods such as Random Forest and XGBoost. These papers explore the robustness and effectiveness of ensemble models in addressing credit risk and financial default prediction problems.

2.1. Dynamic Effectiveness of Random Forest Algorithm in Financial Credit Risk Management

This paper evaluates the Random Forest algorithm's performance in predicting loan defaults in financial credit risk datasets. The authors demonstrate how Random Forest, with its use of bootstrapping and bagging, outperforms traditional decision tree models by reducing variance and improving predictive accuracy. The research was conducted on a dataset with over 32,000 entries, showing the importance of ensemble models in large-scale financial datasets. The study highlights how Random Forest's feature importance mechanism helps in identifying key factors influencing financial risks [5].

2.2. Ensemble Methodology: Innovations in Credit Default Prediction Using LightGBM, XGBoost, and LocalEnsemble

This research focuses on improving credit default prediction by utilizing multiple ensemble learning techniques, including LightGBM and XGBoost. The study emphasizes the power of feature engineering and the combination of diverse models to enhance accuracy. The proposed ensemble approach improves model generalization and reduces overfitting when applied to customer profiles and behavioral data in lending scenarios. The authors show that using a soft voting ensemble strategy further enhances the model's robustness in financial risk assessments [7].

These two studies provide a comprehensive understanding of the application of ensemble models in financial risk prediction. Their findings validate the use of ensemble techniques, such as Random Forest and XGBoost, for improving prediction accuracy and reducing model overfitting.

3. Dataset and Data Preprocessing

The Financial Risk Assessment Dataset includes various financial and demographic attributes such as income, credit score, loan amount, assets value, number of dependents, previous defaults, gender, marital status, and education level. The dataset reflects real-world financial scenarios, including imbalanced classes and missing values. In the data preprocessing stage, the following steps were applied:

- **Handling Missing Data:** Missing values in numerical features like income, credit score, and loan amount were imputed using the median to prevent data skew.

- **Encoding Categorical Variables:** Categorical features (e.g., gender, education, marital status) were encoded numerically using `LabelEncoder`.
- **Addressing Class Imbalance:** SMOTE was applied to balance the 'Risk Rating' classes by oversampling minority categories.
- **Feature Scaling:** Numerical features were normalized using `StandardScaler` to improve model performance.
- **Correlation Analysis:** A correlation heatmap revealed no significant linear relationships among features.

Note: We excluded certain irrelevant features, such as 'State', 'Country', and 'City', for convenience, as they were not categorical in nature.

4. Methodology and Model Details

We employed Decision Tree, Random Forest, and XGBoost models for financial risk prediction, along with a Voting Classifier that combined the outputs of XGBoost, Decision Tree, and Random Forest. The strengths and limitations of each model, along with key hyperparameters, are outlined below.

4.1. Decision Tree

The Decision Tree algorithm splits data based on feature significance, providing high interpretability but prone to overfitting. We used a **CART** variant.
Advantages: Easy to interpret, handles both categorical and numerical data.
Disadvantages: Prone to overfitting.

4.2. Random Forest

Random Forest reduces overfitting by averaging the output of multiple decision trees. It is more accurate than a single tree but computationally expensive.
Advantages: Reduces overfitting, handles many features, provides feature importance.
Disadvantages: Computationally intensive, less interpretable than a single tree.

4.3. XGBoost

XGBoost [2] improves accuracy by building trees sequentially to correct errors. It includes regularization to control overfitting and supports parallel training.
Advantages: High accuracy, regularization (L1/L2), fast training.
Disadvantages: Complex, performance depends on hyperparameter tuning.

4.4. Voting Classifier

The Voting Classifier [3] aggregates the predictions from multiple models (XGBoost, Decision Tree, and Random Forest) to enhance accuracy using soft voting.
Advantages: Improved accuracy, robust to individual model weaknesses.
Disadvantages: Increased complexity, harder to interpret.

4.5. Multi-Layer Perceptron

The MLP [6] is a type of neural network capable of capturing non-linear relationships in data through its layered architecture. It works well with complex datasets but requires careful tuning and sufficient data.
Advantages: Captures non-linear patterns, flexible architecture, supports multiple outputs.
Disadvantages: Sensitive to hyperparameter tuning, computationally intensive, prone to overfitting in small data sets.

4.6. Support Vector Machine

SVM [4] uses margin maximization to create decision boundaries and is effective in high-dimensional spaces. It is robust with well-separated classes but may struggle with noisy data or overlapping classes.
Advantages: Effective in high-dimensional spaces, robust to overfitting with proper kernel choice.
Disadvantages: Computationally expensive for large datasets, requires careful kernel and parameter selection, sensitive to outliers.

The following table outlines the key hyperparameters used for each model.

Model	**Hyperparameters**
Decision Tree (dt_model)	max_depth=9, min_samples_split=2, random_state=42
Random Forest (rf_model)	n_estimators=1000, max_depth=500, min_samples_split=5, random_state=42
XGBoost (xgb_model)	n_estimators=1550, max_depth=3, random_state=42, learning_rate=0.01
Voting Classifier	rf_model, dt_model, xgb_model
Multi-Layer Perceptron	hidden_layers=(150,100,50,), solver=adam, learning_rate=adaptive, max_iter=1000, random_state=42, alpha=0.001, activation_function=tanh
Support Vector Machine	kernel=rbf, C=1, gamma=0.1

5. Results and Analysis

5.1. Overview of Model Performance

Table 1 summarizes the performance of the Decision Tree, Random Forest, XGBoost, and Voting Classifier models based on key metrics: precision, recall, and F1-score.

Model	Precision	Recall	F1-Score	Accuracy
Decision Tree	0.44	0.54	0.46	53.80%
Random Forest	0.48	0.57	0.48	56.60%
XGBoost	0.48	0.58	0.47	58.03%
Voting Classifier	0.45	0.57	0.46	56.63%
MLP	0.46	0.45	0.45	44.60%
SVM	0.45	0.44	0.44	43.90%

Table 1. Classification Report for Financial Risk Assessment Models

5.2. Model-Specific Analysis

Decision Tree: The Decision Tree model [1] achieved a test accuracy of 53.80%. While it provided quick and interpretable decisions, the model struggled with generalization, especially in predicting low-risk and high-risk borrowers, likely due to overfitting.

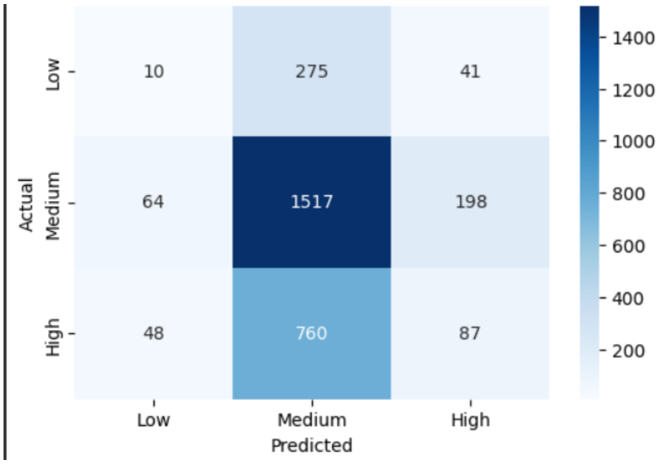

Figure 1. Confusion Matrix for Decision Tree

Random Forest: The Random Forest model performed better, with a test accuracy of 56.60%. It reduced overfitting through ensemble learning, providing more reliable predictions across categories. However, it still faced difficulty in accurately predicting low- and high-risk categories, which points to the impact of class imbalance.

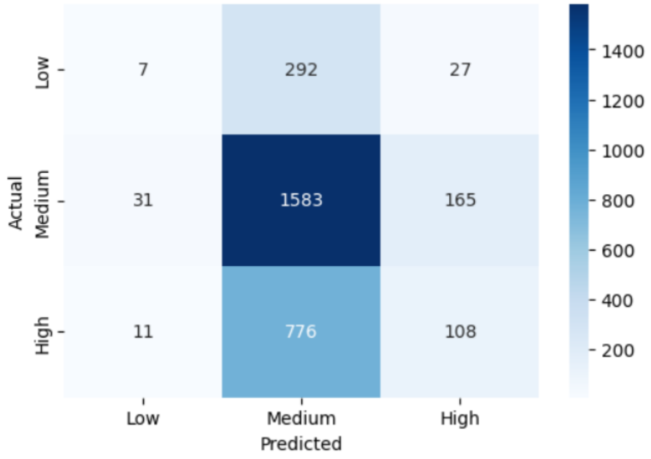

Figure 2. Confusion Matrix for Random Forest

XGBoost Model: XGBoost [2] slightly outperformed Random Forest, with an accuracy of 58.03%. It excelled in classifying medium-risk borrowers but struggled with low- and high-risk categories. XGBoost's gradient boosting approach and scalability make it a strong candidate for further optimization to handle class imbalance.

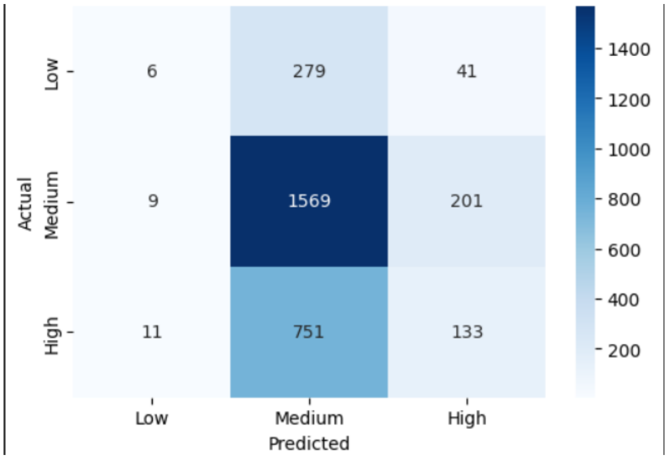

Figure 3. Confusion Matrix for XGBoost

Class	Precision	Recall	F1-Score
0	0.17	0.01	0.02
1	0.59	0.95	0.73
2	0.36	0.06	0.10
Macro Avg	**0.38**	**0.34**	**0.28**
Weighted Avg	**0.48**	**0.58**	**0.47**

Classification Report for XGBoost.

Voting Classifier Model: The Voting Classifier [3], which combines Decision Tree, Random Forest, and XGBoost, achieved a test accuracy of 57.13%. Despite this, the overall performance was constrained by its lower precision and recall in distinguishing low- and high-risk borrowers, reflecting the challenges presented by the imbalanced dataset.

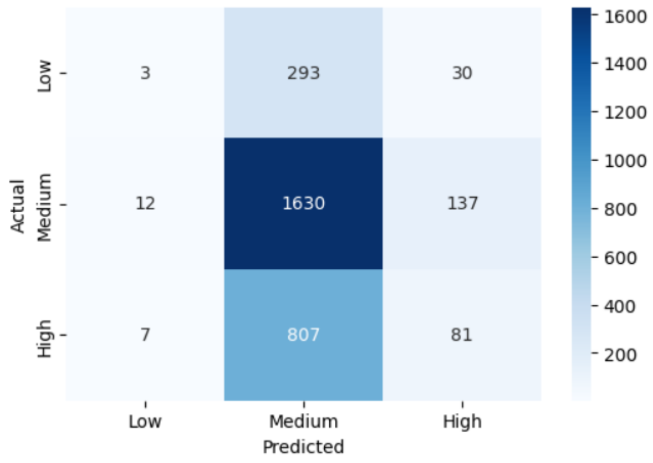

Figure 4. Confusion Matrix for Voting Classifier

Multi-Layer Perceptron: The Multi-Layer Perceptron (MLP) [6] model achieved a precision of 0.46, recall of 0.45, and an F1-score of 0.45, with an accuracy of 44.60%. These results indicate that while the MLP model is moderately balanced in terms of precision and recall, its overall performance lags compared to other models like Random Forest and XGBoost.

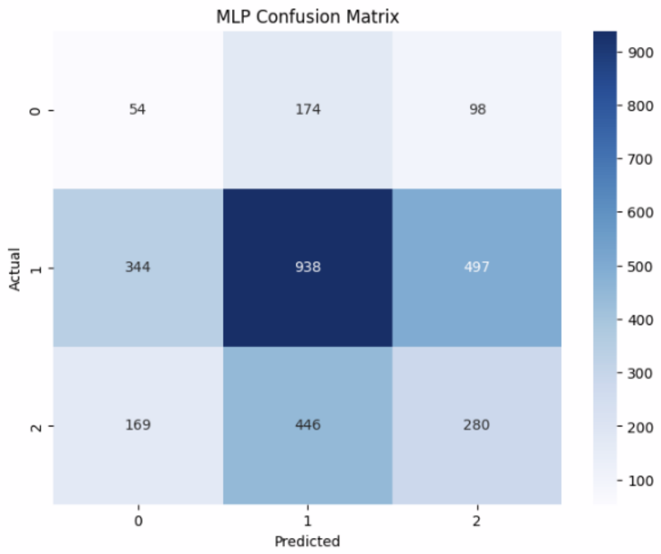

Figure 5. Confusion Matrix for Multi-Layer Perceptron

Support Vector Machine: The Support Vector Machine (SVM) [4] model showed a precision of 0.45, recall of 0.44, and an F1-score of 0.44, with an accuracy of 43.90%. Despite its theoretical strengths in handling high-dimensional data, the SVM model underperformed in this context, likely due to the complexity or distribution of the dataset.

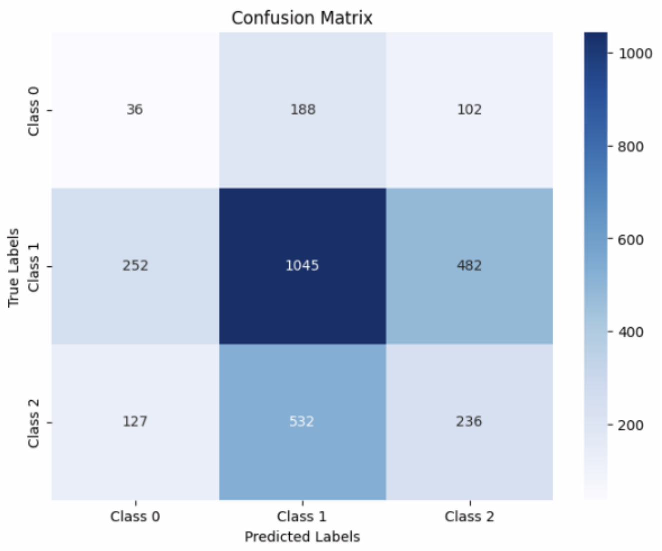

Figure 6. Confusion Matrix for Support Vector Machine

6. Conclusion

The performance analysis of various models reveals that XGBoost outperformed other algorithms with the highest accuracy of 56.93% and balanced precision, recall, and F1-score, effectively handling the complexities of financial risk prediction. Confusion matrices highlighted challenges in classifying low- and high-risk categories, particularly for the Voting Classifier, despite SMOTE's partial mitigation. Feature importance from Random Forest identified income and credit score as key predictors. A Flask-based website was developed to provide real-time financial risk predictions, showcasing the model's practical applicability.

6.1. Team Contributions

Each team member made the following contributions:

- **Devyansh Chaudhary:** Data Pre-Processing and Visualization, Model Selection
- **Dhawal Bansal:** Model Evaluation & Interface development
- **Dhawal Garg:** Model Evaluation & Interface development
- **Manas Chhabra:** Model Selection, Normalization and Results Analysis

References

[1] Bahzad Charbuty and Adnan Abdulazeez. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01):20–28, 2021.

[2] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[3] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[4] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[5] Afolashade Oluwakemi Kuyoro, Olufunmilola Adunni Ogunyolu, Thomas Gbadebo Ayanwola, and Folasade Yetunde Ayankoya. Dynamic effectiveness of random forest algorithm in financial credit risk management for improving output accuracy and loan classification prediction. *Ingénierie des systèmes d'information*, 27(5):815, 2022.

[6] Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7):579–588, 2009.

[7] Mengran Zhu, Ye Zhang, Yulu Gong, Kaijuan Xing, Xu Yan, and Jintong Song. Ensemble methodology: Innovations in credit default prediction using lightgbm, xgboost, and localensemble. *arXiv preprint arXiv:2402.17979*, 2024.