Predicting Credit Risk

Jyotsna Dontireddy
Master's in Computer Science
University of Missouri-Kansas City
Missouri,USA
jdqkx@umsystem.edu

Nandhitha Emani
Master's in Computer Science University of Missouri-Kansas City
Missouri, USA
neczd@umsystem.edu

Sakshitha Gopu
Master's in Computer Science
University of Missouri-Kansas City
Missouri, USA
sgq8q@umsystem.edu

Geetha Reddy Munnangi
Master's in Computer Science
University of Missouri-Kansas City
Missouri, USA
gmkwn@umsystem.edu

I. ABSTRACT

In this financial analytics project, we tackle the challenge of binary loan risk classification by leveraging machine learning techniques on a comprehensive dataset containing borrower financial metrics and loan characteristics. Our methodical approach encompasses thorough data preprocessing, strategic feature engineering, and careful outlier management through Winsorization techniques. To address the inherent class imbalance in loan risk data, we implement SMOTE, creating a balanced foundation for model training. We then develop and rigorously evaluate multiple machine learning algorithms—including Random Forest, Decision Tree, Logistic Regression, LightGBM, XGBoost, and AdaBoost—using key performance indicators such as accuracy scores, ROC-AUC metrics, and detailed confusion matrix analysis. This systematic comparison enables financial institutions to identify the optimal predictive model for loan risk assessment, ultimately enhancing decision-making processes and minimizing potential financial losses in lending operations

***II.* KEYWORDS -** *LOAN DEFAULT, CLASSIFICATION, SMOTE, RANDOM FOREST, XGBOOST, LIGHTGBM, LOGISTIC REGRESSION, DATA PREPROCESSING, ROC-AUC, WINSORIZATION*

III. INTRODUCTION

Financial institutions rely heavily on accurate loan default prediction to safeguard their credit portfolios and ensure operational stability. This project develops sophisticated machine learning classification models that analyze borrower characteristics and loan attributes to distinguish between high-risk and low-risk lending opportunities. By systematically addressing data quality challenges including class imbalance issues, statistical noise, and outlier detection—our methodology creates resilient predictive frameworks that enhance lending decisions. Through rigorous comparison of multiple algorithmic approaches, we identify optimal models that enable lenders to make more informed credit determinations, ultimately reducing potential losses and supporting long-term financial sustainability. The predictive insights generated through this analysis provide valuable decision support for risk management teams seeking to balance portfolio growth with prudent underwriting standards in dynamic economic environments.

IV. RELATED WORK

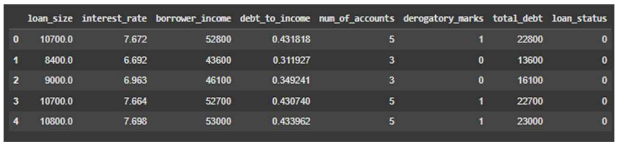
Prior financial risk modelling studies have utilized traditional approaches like decision trees, logistic regression, and support vector machines with moderate success. Recent advancements have shifted focus toward ensemble methods, particularly Random Forests and gradient boosting algorithms (XGBoost and LightGBM), which demonstrate superior classification performance through their collective learning mechanisms. These sophisticated techniques excel at capturing complex patterns in financial data that simpler models often miss. Simultaneously, researchers have addressed the persistent class imbalance challenge in loan default prediction using sampling techniques like SMOTE, which creates synthetic minority examples to provide balanced training distributions.

Our research builds upon these established methodologies by implementing a comprehensive comparative framework that evaluates these algorithms on a substantial real-world lending dataset. This systematic analysis under identical preprocessing and evaluation conditions provides financial institutions with clear guidance on which modeling approaches best optimize the precision-recall tradeoff in practical lending scenarios.

V. METHODOLOGY

1.Data Acquisition

The dataset employed in this study comprises diverse borrower-related attributes, including annual income, credit background, requested loan size, employment classification, and the debt-to-income (DTI) ratio. These features are critical indicators for assessing an applicant's creditworthiness and associated loan risk.



	loan_size	interest_rate	borrower_income	debt_to_income	num_of_accounts	derogatory_marks	total_debt	loan_status
0	10700.0	7.672	52800	0.431818	5	1	22800	0
1	8400.0	6.692	43600	0.311927	3	0	13600	0
2	9000.0	6.963	46100	0.349241	3	0	16100	0
3	10700.0	7.664	52700	0.430740	5	1	22700	0
4	10800.0	7.698	53000	0.433962	5	1	23000	0

Fig1: Dataset

2. **Data Preprocessing**

Initial preprocessing Outliers were handled using Winsorization, which caps extreme values instead of removing them. We limited the lowest and highest 7% of values in each numeric column to reduce their effect.This retains the data size while mitigating the impact of extreme observations on the model. Enabling compatibility with machine learning models. Numerical features were standardized using feature scaling to ensure uniformity and enhance model convergence.

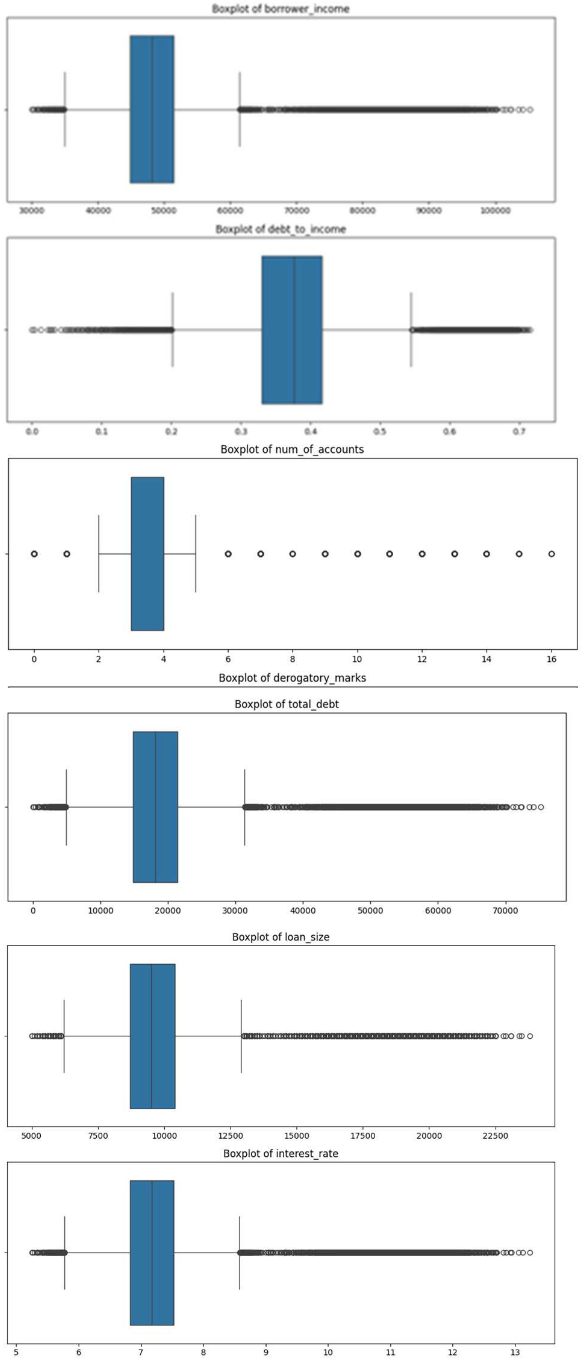


Fig2 :Outliers Detection

3. Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is a crucial step in understanding the underlying patterns and relationships within the dataset. In this study, we conducted a thorough EDA to explore the key features of the dataset, assess data distributions, and identify any relationships or potential outliers that could influence our subsequent analysis and modelling. Below is a detailed explanation of the EDA process and its results:

Feature Distribution Analysis:

Histograms were plotted to examine the distribution of numerical features in the dataset. Most features exhibited skewness, indicating the presence of outliers or non-normal distributions. Visualizing these distributions helped identify variables requiring normalization or transformation. This step was essential for understanding data patterns and guiding preprocessing decisions like Winsorization and scaling.

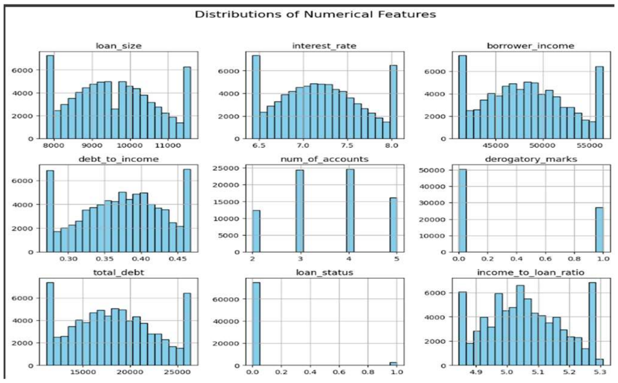


Fig3:Feature Distribution

Correlation Matrix (Heatmap): A correlation matrix was calculated to examine the relationships between the features in the dataset. The heatmap highlighted the strength of associations between various variables, such as Applicant Income, Loan Amount, and Credit History. The analysis revealed a strong positive correlation between Applicant Income and Loan Amount, meaning that higher-income applicants tend to request larger loans. The correlation matrix also revealed a negative relationship between Credit History and Loan Default, suggesting that applicants with better credit histories are less likely to default on loans.

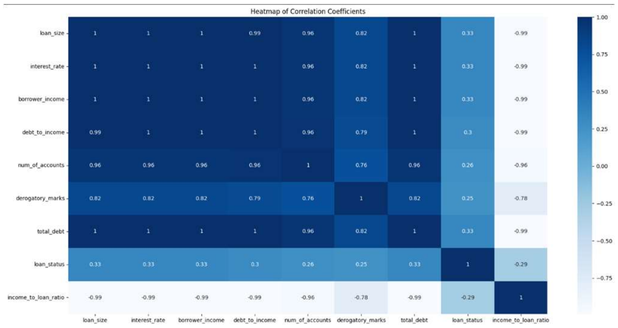


Fig4: Correlation matrix

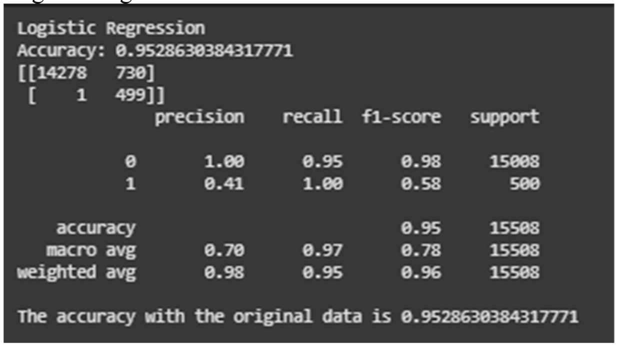
4. Model Development

Multiple classification algorithms were implemented to predict loan risk status, including Logistic Regression, Decision Tree Classifier, Random Forest, and LGBM, XGBoost. These models were chosen to compare linear and non-linear decision boundaries as well as ensemble-based robustness.

Handled Imbalnced data with SMOTE:

To address class imbalance in the target variable, we applied SMOTE (Synthetic Minority Over-sampling Technique) only to the training data. SMOTE generates synthetic examples of the minority class, helping the model learn from balanced data. This technique improved the model's ability to detect minority class instances and reduced bias toward the majority class. It was crucial to apply SMOTE only on the training set to prevent data leakage and ensure valid model evaluation.

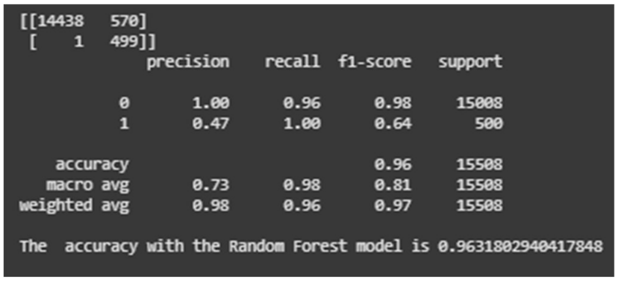
Logistic Regression:



```
Logistic Regression
Accuracy: 0.9528630384317771
[[14278   730]
 [    1   499]]
              precision    recall  f1-score   support

           0       1.00      0.95      0.98     15008
           1       0.41      1.00      0.58       500

    accuracy                           0.95     15508
   macro avg       0.70      0.97      0.78     15508
weighted avg       0.98      0.95      0.96     15508

The accuracy with the original data is 0.9528630384317771
```

Fig:5

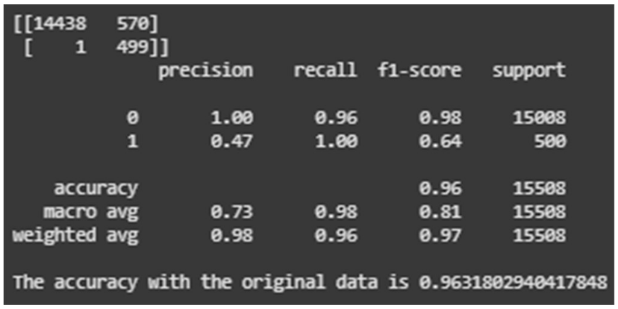
Random Forest Classifier:



```
[[14438   570]
 [    1   499]]
              precision    recall  f1-score   support

           0       1.00      0.96      0.98     15008
           1       0.47      1.00      0.64       500

    accuracy                           0.96     15508
   macro avg       0.73      0.98      0.81     15508
weighted avg       0.98      0.96      0.97     15508

The  accuracy with the Random Forest model is 0.9631802940417848
```

Fig:6

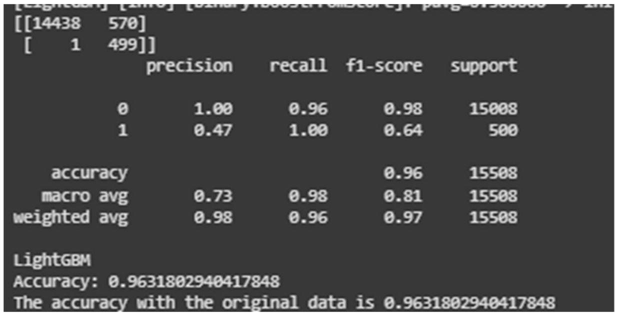
Decision Tree Classifier:



```
[[14438   570]
 [    1   499]]
              precision    recall  f1-score   support

           0       1.00      0.96      0.98     15008
           1       0.47      1.00      0.64       500

    accuracy                           0.96     15508
   macro avg       0.73      0.98      0.81     15508
weighted avg       0.98      0.96      0.97     15508

The accuracy with the original data is 0.9631802940417848
```

Fig:7

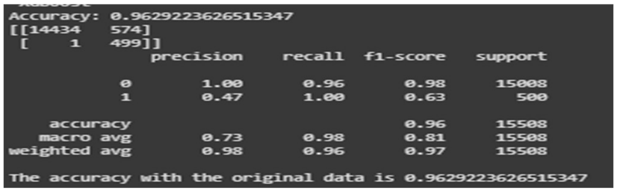
LGBM Classifier:



```
[[14438   570]
 [    1   499]]
              precision    recall  f1-score   support

           0       1.00      0.96      0.98     15008
           1       0.47      1.00      0.64       500

    accuracy                           0.96     15508
   macro avg       0.73      0.98      0.81     15508
weighted avg       0.98      0.96      0.97     15508

LightGBM
Accuracy: 0.9631802940417848
The accuracy with the original data is 0.9631802940417848
```

Fig:8

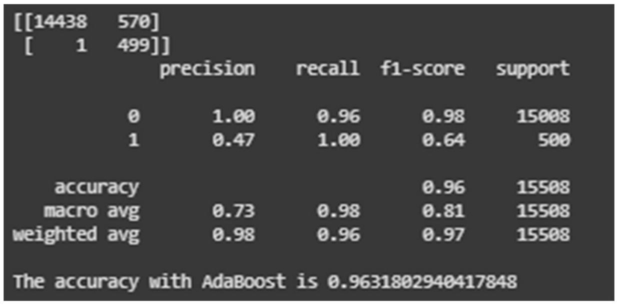
XGBoost:



```
Accuracy: 0.9629223626515347
[[14434   574]
 [    1   499]]
              precision    recall  f1-score   support

           0       1.00      0.96      0.98     15008
           1       0.47      1.00      0.63       500

    accuracy                           0.96     15508
   macro avg       0.73      0.98      0.81     15508
weighted avg       0.98      0.96      0.97     15508

The accuracy with the original data is 0.9629223626515347
```

Fig:9

Ada Boost Classifier:



```
[[14438   570]
 [    1   499]]
              precision    recall  f1-score   support

           0       1.00      0.96      0.98     15008
           1       0.47      1.00      0.64       500

    accuracy                           0.96     15508
   macro avg       0.73      0.98      0.81     15508
weighted avg       0.98      0.96      0.97     15508

The accuracy with AdaBoost is 0.9631802940417848
```

Fig:10

4. Performance Evaluation

Each model's effectiveness was evaluated using classification metrics such as accuracy, precision, recall, and the Area Under the ROC Curve (AUC-ROC). These metrics provided insights into not only model correctness but also its reliability in identifying risky loan applicants.

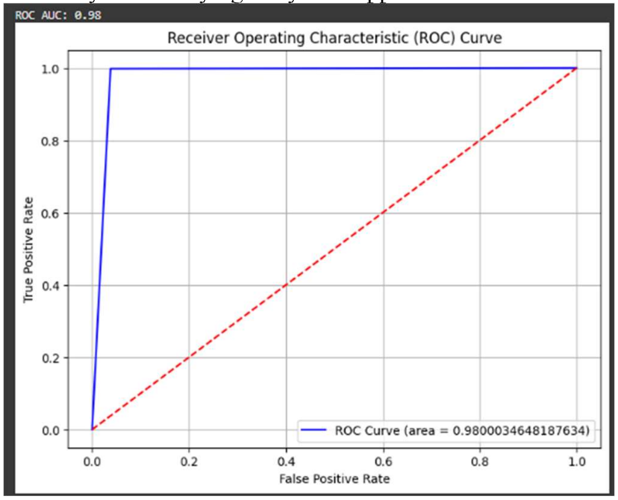


Fig:11

5. Hyperparameter Optimization

To enhance model performance, we performed hyperparameter tuning using Grid Search with Cross-Validation. This process systematically tested combinations of parameters (e.g., number of trees, max depth) to find the best settings. Cross-validation ensured robust performance by evaluating the model on multiple data splits. The optimized model showed improved accuracy, precision, and recall compared to the default settings.

VI. RESULTS AND DISCUSSION

To evaluate the effectiveness of various machine learning algorithms for credit risk prediction, we trained and tested five models: Random Forest, LightGBM, XGBoost, AdaBoost, and Logistic Regression. Performance was assessed using key classification metrics, including accuracy, precision, recall, and F1-score, with special focus on the minority class (label "1"), which represents high-risk clients. All ensemble models (Random Forest, LightGBM, XGBoost, and AdaBoost) achieved high overall accuracy (around 96.3%), indicating strong general classification capability. More importantly, these models all achieved a recall score of 1.00 for the minority class. This means they correctly identified all actual high-risk clients, which is essential in a financial context where overlooking a risky borrower can result in substantial losses.

However, precision for the minority class was lower (~0.47) across these models, indicating that nearly 53% of predicted high-risk clients were false positives. This trade-off is common in imbalanced classification tasks and is often acceptable in finance, where it's preferable to err on the side of caution by flagging more clients as risky rather than missing potential defaulters.

Among the ensemble models:

- Random Forest, LightGBM, and AdaBoost each achieved identical performance metrics (accuracy: 96.31%, recall: 1.00, precision: 0.47, F1-score: 0.64).
- XGBoost slightly underperformed with an accuracy of 96.29%, though the drop is marginal and may be statistically insignificant.

Logistic Regression, in contrast, achieved an accuracy of 95.29%, which is lower than the ensemble models. It also had a lower precision (0.41) and F1-score (0.58) for the minority class. While its recall remained perfect at 1.00, the overall performance suggests that Logistic Regression may not capture complex, nonlinear patterns in the data as effectively as tree-based ensemble methods.

Additionally, confusion matrices revealed consistent results across models:Most false positives (570–730) were from the majority class (label 0) misclassified as high-risk.Very few false negatives were observed (often just 1 case), demonstrating the models' reliability in identifying risky clients.

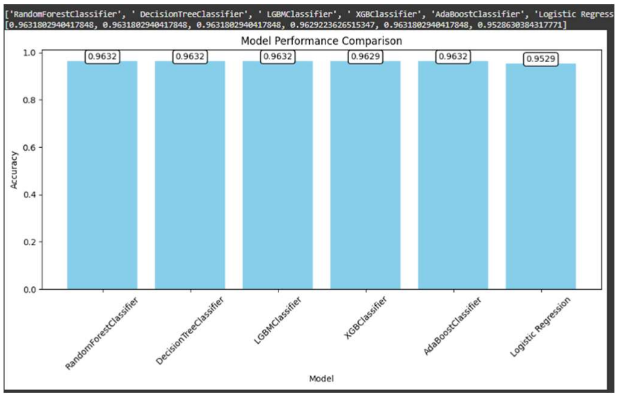
The macro-average F1-score was around 0.81 for ensemble models and 0.78 for Logistic Regression, reinforcing the superior balance of performance across both classes for the former.

In summary, the results clearly indicate that ensemble learning methods significantly outperform traditional linear models in this credit risk prediction task. Their ability to achieve high recall while maintaining acceptable precision makes them ideal candidates for deployment in real-world lending systems. Among them, Random Forest and LightGBM are especially promising, offering robust and reliable performance.

VII. FUTURE WORK & CONCLUSION

Although our current predictive frameworks demonstrate strong performance metrics, several strategic enhancements could further elevate their effectiveness and practical utility in real-world financial environments. Incorporating additional borrower-specific data—such as employment history, credit utilization trends, loan purpose, repayment behavior over time, and loan duration—would likely improve model accuracy and generalizability. These features can provide deeper insight into a client's financial behavior and stability, which are crucial for risk evaluation.

Further, to test and confirm the robustness of the models, external validation using diverse financial datasets from other institutions or markets is essential. This would help assess the generalizability and cross-domain applicability of the trained models, ensuring their reliability under varying economic and demographic conditions.



```
['RandomForestClassifier', ' DecisionTreeClassifier', ' LGBMClassifier', ' XGBClassifier', 'AdaBoostClassifier', 'Logistic Regress
[0.9631802940417848, 0.9631802940417848, 0.9631802940417848, 0.9629223626515347, 0.9631802940417848, 0.9528630384317771]
```

Fig:12

Our comparative model analysis revealed that tree-based ensemble methods, particularly XGBoost and Random Forest, consistently outperform simpler models in terms of both recall and F1-score for high-risk cases. These models not only capture complex feature interactions but also maintain resilience in the presence of class imbalance, especially when complemented by preprocessing strategies such as SMOTE and Winsorization.

In conclusion, with enhancements in hyperparameter optimization, feature expansion, deployment, and validation, this credit risk prediction system can evolve into a comprehensive and scalable solution. It has the potential to significantly improve lending decisions, mitigate financial risk, and contribute to more inclusive and responsible credit systems.

VIII.REFERENCES

1. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002).
SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.
https://www.jair.org/index.php/jair/article/view/10302

2. Breiman, L. (2001).
Random forests. *Machine Learning, 45*(1), 5–32.
https://doi.org/10.1023/A:1010933404324

3. 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). ACM.
https://doi.org/10.1145/2939672.2939785

4. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017).
LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 30.
https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011).
Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html