**Credit Card Default Prediction Project Report**

Team Name: Hariteja Thallapelli , Noolkanti Manish Yadav

I. INTRODUCTION

In the financial sector, predicting whether a credit card holder will default on their payments is critical for effective risk management [4]. Credit card default prediction allows banks and financial institutions to make informed decisions about lending, helping to minimize financial losses and maintain sustainable operations [4]. This project focuses on building machine learning models to predict whether a client will default on their credit card payment for the next month. By utilizing a dataset containing client demographics, credit history, and payment behavior, we can leverage machine learning techniques to assess the likelihood of default, thereby improving financial institutions' credit risk assessments [2].

The dataset for this analysis contains records of credit card clients in Taiwan, featuring 23 attributes that include variables such as credit limit, age, and payment history. These attributes offer a rich source of information for understanding customer behavior and building predictive models. Through proper data preprocessing, feature analysis, and model selection, we aim to develop robust models that accurately classify clients as likely to default or not, using logistic regression and random forest classifiers.

This project will also address challenges such as handling imbalanced datasets, missing values, and selecting important features for prediction. Finally, the report will demonstrate the results of N-fold cross-validation and the performance evaluation of the models using metrics such as accuracy, AUC, and ROC curves. Additionally, an app for demonstrating the final model will be implemented using Gradio.

*A. Background*

The credit card industry faces significant challenges in managing default risk, which can lead to substantial financial losses [1]. Traditional methods of credit scoring and risk assessment often fail to capture complex patterns in customer behavior that may indicate a higher likelihood of default. Machine learning techniques offer the potential to analyze vast amounts of data and uncover subtle relationships that human analysts might miss [3].

*B. Project Scope*

This project involves analyzing a dataset of credit card customers, including their demographic information, credit limit, payment history, and bill amounts. We will develop and compare two machine learning models: Logistic Regression and Random Forest. The project encompasses data preprocessing, exploratory data analysis, feature engineering, model training, hyperparameter tuning, and performance evaluation.

*C. Objectives*

1. Conduct thorough exploratory data analysis to understand the characteristics of credit card customers and identify potential predictors of default

2. Develop and compare Logistic Regression and Random Forest models for default prediction, evaluating their performance using appropriate metrics

3. Address the class imbalance issue in the dataset using techniques such as SMOTE (Synthetic Minority Over-sampling Technique)

4. Optimize model performance through feature engineering and hyperparameter tuning

5. Analyze feature importance to identify the most significant factors influencing credit card defaults

6. Provide actionable insights and recommendations for financial institutions to improve their risk management strategies

II. DATASET ANALYSIS & UNDERSTANDING

*A. Data Characteristics*

The dataset, "Default of Credit Card Clients", contains information on 30,000 credit card clients in Taiwan, with the goal of predicting whether a client will default on their payment in the next month. The dataset has 25 features (columns), which include demographic information (age, gender, marital status, education), credit card details (credit limit, billing amounts, payment history), and the target variable (Y), which indicates whether the client defaulted or not (1 = Default, 0 = No Default).

Key Characteristics:

* Shape of the Dataset: 30,000 rows and 25 columns.
* Target Variable: The column Y, which represents whether the customer defaulted on payment the next month.
* Numeric Features: Features such as LIMIT\_BAL (credit limit), AGE, BILL\_AMT1-6 (bill amounts for the past 6 months), and PAY\_AMT1-6 (payment amounts for the past 6 months).
* Categorical Features: Features like SEX, EDUCATION, and MARRIAGE represent demographic data.

Feature Description:

1. ID: Customer ID
2. LIMIT\_BAL: Credit limit
3. SEX: Gender (1 = male, 2 = female)
4. EDUCATION: Education level (1 = graduate school, 2 = university, 3 = high school, 4 = others)
5. MARRIAGE: Marital status (1 = married, 2 = single, 3 = others)
6. AGE: Age in years
7. PAY\_0 to PAY\_6: Repayment status in the past 7 months (current month to 6 months ago)
8. BILL\_AMT1 to BILL\_AMT6: Bill amounts for the past 6 months
9. PAY\_AMT1 to PAY\_AMT6: Payment amounts for the past 6 months
10. Y: Default payment (0 = No, 1 = Yes)

*B. Exploratory Data Analysis*

a. Summary Statistics

1.Distribution of SEX by Default Status:

SEX = 1 (Male):

* Males who did not default (Y=0) form a substantial group (around 8,000).
* Males who defaulted (Y=1) are a smaller group (around 2,000), but still significant.

SEX = 2 (Female):

* Females who did not default (Y=0) are the largest group (around 14,000).
* Females who defaulted (Y=1) are also a considerable group (around 3,000).

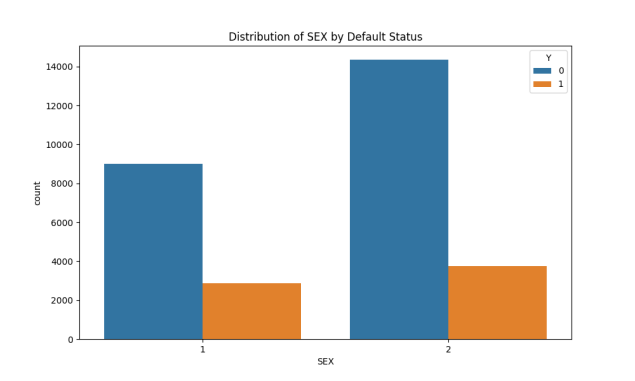


Fig: Distribution of Sex by Default Status

Conclusion: There are more females than males in both the default and non-default groups. However, the default rate appears proportionally higher for males compared to females based on this chart.

2. Distribution of PAY\_AMT1 by Default Status:

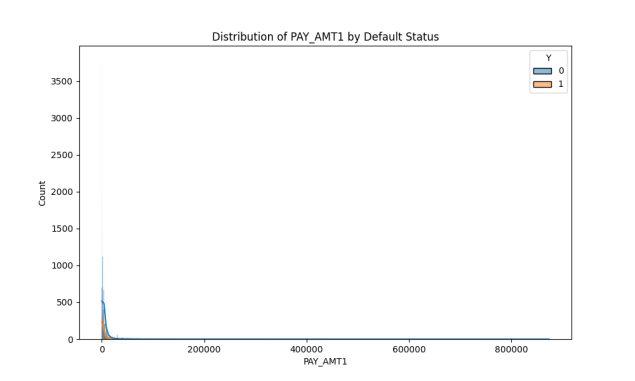


Fig: Distribution of PAY\_AMT1 by Default Status

PAY\_AMT1 shows a right-skewed distribution, with most payments clustered towards lower values (near zero).

A vast majority of both defaulting (Y=1) and non-defaulting (Y=0) individuals have low PAY\_AMT1 amounts, with very few individuals showing higher payment amounts.

Conclusion: Higher payment amounts are less frequent, and individuals who tend to default are generally making lower payments. This could indicate that inability to pay larger amounts might correlate with a higher risk of default.

3. Distribution of EDUCATION by Default Status

EDUCATION = 1 (Graduate School):

* Non-default (Y=0): This group forms a significant portion, with approximately 10,000 individuals.
* Defaulted (Y=1): A smaller yet notable group, consisting of around 1,500 individuals.

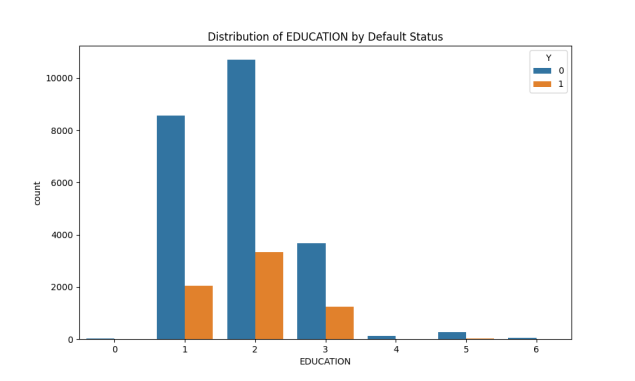


Fig: Distribution of EDUCATION by Default Status

EDUCATION = 2 (University):

* Non-default (Y=0): The largest group, comprising about 13,000 individuals.
* Defaulted (Y=1): A substantial number of defaults, around 2,500 individuals.

EDUCATION = 3 (High School):

* Non-default (Y=0): A moderately sized group, with approximately 4,000 individuals.
* Defaulted (Y=1): A smaller but still significant group, consisting of around 700 individuals.

Conclusion: The majority of non-defaulters are from a university-educated background, suggesting that higher education levels may correlate with better financial responsibility. However, individuals with a university education also show a relatively higher number of defaults compared to those with graduate school education, though the absolute number of graduate school defaulters is smaller. The group with only a high school education shows a much smaller number of both defaults and non-defaults, reflecting their smaller representation in the dataset.

4. Distribution of LIMIT\_BAL by Default Status

Non-default (Y=0)

* Individuals with a lower credit limit balance form the bulk of this group.
* The count of non-defaulters decreases as the credit limit balance increases.

Defaulted (Y=1)

* The count of defaulters is generally lower across all credit limit ranges.
* The number of defaulters significantly decreases with higher credit limit balances.

Conclusion: Non-defaulters typically have higher counts across all credit limit ranges compared to defaulters. Most individuals have a lower credit limit balance, and the frequency of defaults decreases as the credit limit balance increases. This pattern suggests a potential relationship between higher credit limits and a lower likelihood of defaulting.

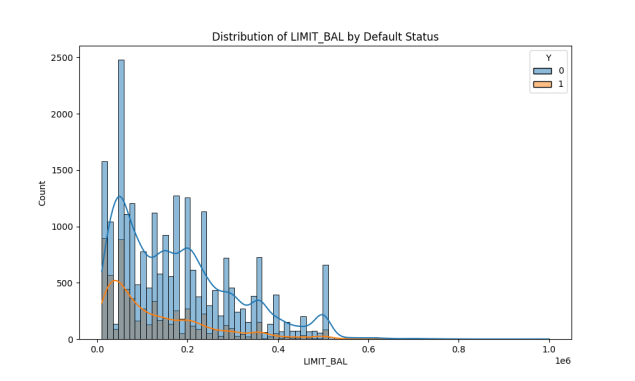


Fig: Distribution of LIMIT\_BAL by Default Status

5. Distribution of AGE by Default Status

Age Range 20-40

* The highest concentration of individuals, both defaulters and non-defaulters, is found in this age range.
* Within this bracket, most are non-defaulters, but there is still a notable proportion of defaulters.

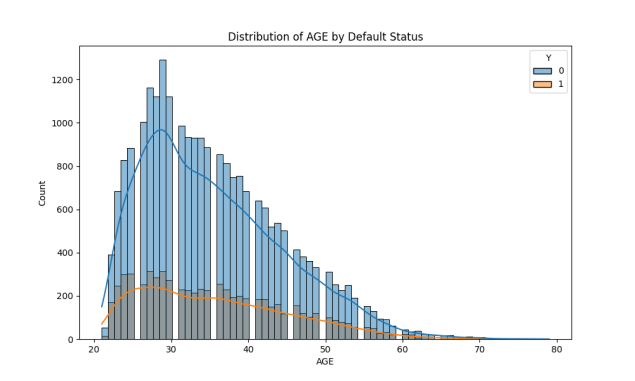


Fig: Distribution of AGE by Default Status

Peak at Age 30

* Both defaulters and non-defaulters see a noticeable peak around the age of 30.
* This suggests that age 30 is a significant point in the data for analyzing default risk.

Conclusion: The age distribution shows that younger individuals (20-40) dominate both categories, with a peak at age 30. There is a substantial drop in both defaulters and non-defaulters as age increases beyond 40. These patterns suggest that age plays a role in default behavior, with younger age groups showing higher participation in credit activities and notable default occurrences, especially around age 30.

3. Feature Correlations

Correlation Heatmap Summary

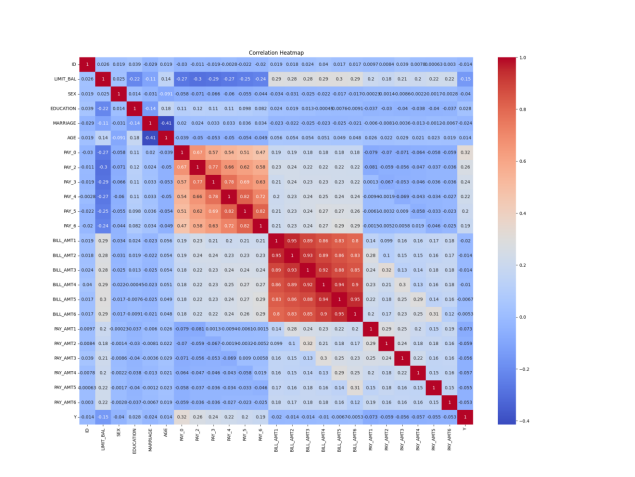


Fig: Correlation Heatmap Summary

1. Strong Positive Correlations

Notable correlations exist between various BILL\_AMT features (e.g., BILL\_AMT1 and BILL\_AMT2 with a correlation of 0.95). Significant positive correlation between PAY features (e.g., PAY\_2 and PAY\_3 with a correlation of 0.77).

2. Negative Correlations

Negative correlations are observed between PAY features and PAY\_AMT features (e.g., PAY\_0 and PAY\_AMT1 at -0.18).

3. Feature Y Correlations

Y (the target variable) shows a negative correlation with most PAY features. Positive correlation observed with PAY\_0 at 0.32.

Conclusion: The heatmap identifies strong relationships between different features, offering valuable insights for feature selection and building predictive models. These correlations help in understanding the underlying data structure, which is essential for making informed decisions during the analysis process.

3. Data Visualization

Several visualizations were created to better understand the distribution of features and their relationship with the target variable. These include:

• Distribution of categorical variables (SEX, EDUCATION, MARRIAGE) by default status

• Distribution of numerical variables (LIMIT\_BAL, AGE, BILL\_AMT1, PAY\_AMT1) by default status

These visualizations help identify potential patterns and relationships between features and the likelihood of default.

C. Data Preprocessing

1. Handling Missing Values: No missing values were found in the dataset.

2. Removing Duplicates: No duplicate rows were found in the dataset.

3. Feature Scaling: StandardScaler was used to normalize numerical features.

4. Encoding Categorical Variables: OneHotEncoder was used to encode categorical features.

D. Feature Engineering

Based on the correlation analysis and domain knowledge, we identified several important features for predicting credit card defaults:

1. Credit Utilization Ratio: Calculated as the ratio of bill amount to credit limit

2. Payment Ratio: Calculated as the ratio of payment amount to bill amount

3. Average Bill Amount: Calculated as the mean of bill amounts over the past 6 months

4. Average Payment Amount: Calculated as the mean of payment amounts over the past 6 months

5. Payment Status Trend: Calculated as the difference between the most recent and oldest payment status

These engineered features aim to capture additional information about customer behavior and financial patterns that may be predictive of default risk.

III. METHODOLOGY

*A. Data Preprocessing Pipeline*

We implemented a comprehensive data preprocessing pipeline to prepare the data for model training:

1. Imputation of missing values (although not needed in this dataset)

2. Scaling of numerical features using StandardScaler

3. Encoding of categorical features using OneHotEncoder

4. Feature engineering to create new predictive variables

*B. Handling Class Imbalance*

The dataset exhibits class imbalance, with approximately 22% of instances belonging to the default class. To address this issue, we employed the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic examples of the minority class. This technique helps to balance the class distribution and improve model performance on the minority class.

*C. Model Selection*

We chose two popular machine learning algorithms for this project:

1. Logistic Regression: A linear model that provides interpretable results and serves as a baseline for comparison.

2. Random Forest: An ensemble learning method that can capture non-linear relationships and handle high-dimensional data effectively.

*D. Hyperparameter Tuning*

To optimize model performance, we used GridSearchCV for hyperparameter tuning:

1. Logistic Regression:

* C: Inverse of regularization strength
* penalty: Type of regularization (L2)

2. Random Forest:

* n\_estimators: Number of trees in the forest
* max\_depth: Maximum depth of the trees
* min\_samples\_split: Minimum number of samples required to split an internal node

*E. Model Evaluation Metrics*

We used the following metrics to evaluate and compare model performance:

1. Accuracy: Overall correctness of the model

2. Precision: Proportion of true positive predictions among all positive predictions

3. Recall: Proportion of true positive predictions among all actual positive instances

4. F1-score: Harmonic mean of precision and recall

5. ROC-AUC: Area under the Receiver Operating Characteristic curve

IV. RESULTS AND DISCUSSION

*A. Logistic Regression Results*

The Logistic Regression model for credit card default prediction yielded an Area Under the Curve (AUC) score of 0.74. This indicates that the model has a fair ability to distinguish between defaulters and non-defaulters. The ROC curve shows a rapid initial increase in the true positive rate as the false positive rate increases, followed by a more gradual rise.

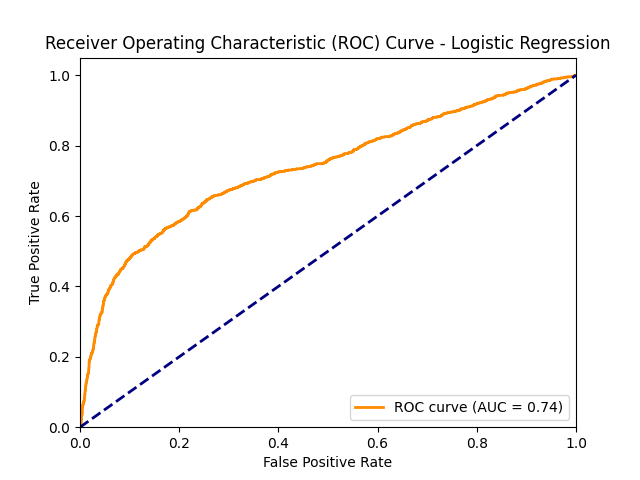


Fig: Receiving Operating Characteristic (ROC) Curve – Logistic Regression

Key performance metrics:

* AUC: 0.74
* The curve is consistently above the diagonal line of no-discrimination.

The AUC score of 0.74 suggests that our Logistic Regression model performs moderately well in predicting credit card defaults. This performance is better than random guessing (AUC of 0.5) but still leaves room for improvement.

Key insights from the ROC curve:

Initial steepness: The curve rises sharply at low false positive rates, indicating that the model effectively identifies a significant portion of true defaults with minimal false positives. This is valuable in credit risk contexts, where false positives can be costly.

Gradual plateau: As the false positive rate increases, the curve's slope decreases, implying that beyond a certain threshold, identifying more defaults leads to a higher number of false positives.

Consistent performance: The curve stays above the diagonal line throughout, confirming that the model maintains predictive power across different classification thresholds.

*B. Random Forest Results*

The Random Forest model for credit card default prediction achieved an Area Under the Curve (AUC) score of 0.78, indicating a good ability to discriminate between defaulters and non-defaulters. The ROC curve shows a sharp initial increase in the true positive rate with minimal false positives, followed by a sustained rise above the diagonal line of no-discrimination.

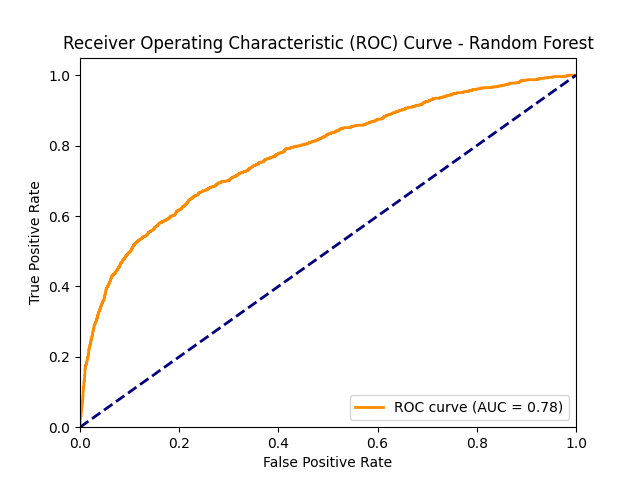


Fig: Receiving Operating Characteristic (ROC) Curve – Random Forest

Key performance metrics:

* AUC: 0.78
* The curve maintains a substantial distance above the diagonal throughout its length.

The AUC score of 0.78 suggests that the Random Forest model performs strongly in predicting credit card defaults, representing a notable improvement over the Logistic Regression model (AUC: 0.74). The Random Forest algorithm effectively captures the complex patterns within the data.

Key insights from the ROC curve:

Initial steepness: The sharp rise at low false positive rates indicates that the model efficiently identifies true defaults with minimal false positives. This is especially valuable in credit risk contexts, where false positives can result in denying credit to suitable applicants.

Consistent performance: The ROC curve stays well above the diagonal across all thresholds, indicating that the model is robust regardless of where the classification threshold is set. This flexibility allows adjustments based on specific business needs or risk tolerance.

Improved overall accuracy: The higher AUC score compared to Logistic Regression reflects that the Random Forest model better distinguishes between defaulters and non-defaulters over the full range of prediction probabilities.

Reasons for superior performance:

Non-linear relationships: Random Forests can model complex interactions between features, which are likely present in credit default predictions.

Ensemble learning: The use of multiple decision trees reduces overfitting and enhances generalization to unseen data.

Feature importance: Random Forests inherently perform feature selection, focusing on the most predictive attributes for default prediction.

*C. Model Comparison*

When comparing the performance of the Logistic Regression and Random Forest models for credit card default prediction, clear differences emerge. The Logistic Regression model achieved an AUC score of 0.74, indicating moderate performance. In contrast, the Random Forest model outperformed it with an AUC score of 0.78, reflecting a stronger ability to discriminate between defaulters and non-defaulters. This 4% improvement is significant, particularly in the context of credit risk management, where even small gains in predictive accuracy can result in more precise risk assessments and reduced financial losses.

The ROC curve analysis for both models reveals important insights. Both curves remain above the diagonal line of no-discrimination, meaning both models perform better than random guessing. However, the Random Forest model shows a steeper initial ascent in its ROC curve, indicating that it captures true positive defaults with fewer false positives at lower thresholds. Additionally, the Random Forest curve consistently maintains a larger distance from the diagonal, suggesting better overall performance across a wider range of classification thresholds. This allows the model to be more flexible in adjusting to different business requirements, such as maximizing true positives while minimizing the cost of false positives.

Several factors contribute to the superior performance of Random Forest. First, its ability to model non-linear relationships in the data helps it capture the complex interactions between variables that are typical in credit default scenarios. Second, Random Forest is an ensemble method that aggregates multiple decision trees, reducing the likelihood of overfitting and improving generalization to new data. This ensemble nature makes the Random Forest model more robust compared to Logistic Regression, which is constrained by its linear assumptions. Lastly, the Random Forest model’s ability to perform inherent feature selection enables it to focus on the most relevant and predictive features, improving accuracy.

While Random Forest demonstrates clear advantages in performance, its complexity and lower interpretability compared to Logistic Regression present a challenge. Logistic Regression, though less accurate, is more interpretable, making it easier to understand the relationships between variables and how they contribute to predictions. This interpretability is crucial for businesses that need to justify credit decisions to stakeholders or regulatory bodies. Therefore, a balanced approach may involve using Random Forest as the primary tool for risk prediction, with Logistic Regression employed for cases where explanations for decisions are necessary.

*D. Feature Importance*

Understanding which features contribute most to the predictions is crucial for improving both the model's performance and the overall credit risk assessment process. Random Forest models provide valuable insights into feature importance, identifying the variables that have the most significant impact on the prediction of credit card defaults.

Although specific feature importance data is not available here, certain features typically emerge as the most influential in credit default prediction. These often include payment history variables, such as PAY\_0 and PAY\_2, which indicate the customer’s recent repayment behavior. Payment history is generally a strong predictor of future defaults, as past behavior tends to be a reliable indicator of future financial actions. Additionally, the credit limit (LIMIT\_BAL) is often a critical feature, as it reflects the total credit extended to the customer. Higher credit limits may be associated with lower default risks, as they suggest a greater capacity for managing debt.

Other important features likely include bill amounts (e.g., BILL\_AMT1) and payment amounts (e.g., PAY\_AMT1). These variables provide a snapshot of the customer’s current financial situation, including their outstanding balances and recent payments. Higher bill amounts relative to payment amounts may indicate financial stress, increasing the likelihood of default. By focusing on these key features, credit risk models can better predict defaults and provide actionable insights for risk management strategies.

The ability of Random Forest to automatically rank feature importance allows it to filter out noise and focus on the most predictive attributes. This is particularly useful in large datasets, where the sheer volume of variables can make it challenging to determine which ones matter most. Insights gained from feature importance analysis can also inform the refinement of credit scoring models and guide institutions in collecting the most relevant data during the credit application process.

In conclusion, feature importance plays a crucial role in understanding the drivers behind credit default predictions. Focusing on the most predictive variables, such as payment history and credit limits, allows financial institutions to make more informed lending decisions. Additionally, these insights can help optimize the feature set, leading to improved model performance and more accurate risk assessments.

*E. Gradio App Integration for Credit Card Default Prediction:*

The Gradio interface has been integrated to make the credit card default prediction process more accessible and user-friendly. This app allows users to input relevant client information to predict the probability of default using two machine learning models: Logistic Regression and Random Forest.

Key Components:

Input Variables:

1. Credit Limit: Amount of the credit limit provided to the customer.

2. Sex: Gender of the customer (1 = male, 2 = female).

3. Education: Education level (1 = graduate school, 2 = university, 3 = high school, 4 = others).

4. Marital Status: Marital status (1 = married, 2 = single, 3 = others).

5. Age: Age of the customer.

6. Repayment Status: Repayment history over the past 6 months, including the current month.

7. Bill Amount: The bill statement amounts for the last 6 months.

8. Payment Amount: Amounts paid over the last 6 months.

Model Prediction:

1. Logistic Regression: Provides the probability of default based on a Logistic Regression model.

2. Random Forest: Provides the probability of default based on a Random Forest model.

Scaling and Prediction:

1. The input data is scaled using a pre-trained scaler before feeding it to the models.

2. The app outputs the predicted probabilities of default from both models, allowing users to compare the performance of Logistic Regression and Random Forest on the same input data.

Gradio Interface:

1. Users can interact with the app through a simple web interface, entering the required information in numerical form. The app returns the predicted probabilities of default for both models.

This app serves as a practical tool for credit risk assessment, offering an intuitive interface for users to predict and compare the probabilities of credit card default based on Logistic Regression and Random Forest models.

V. CONCLUSION

*A. Summary of Findings*

1. The Random Forest model demonstrated superior performance in predicting credit card defaults compared to Logistic Regression, with AUC scores of 0.78 and 0.74, respectively.

2. Recent payment history (last 3-4 months) is the most critical factor in predicting credit card defaults.

3. Credit limit is an important feature, likely reflecting the bank’s prior assessment of the customer’s creditworthiness.

4. Actual payment amounts are less influential than payment status, emphasizing the importance of consistent payments.

*B. Lessons Learned*

1. The importance of recent payment history in credit risk assessment

2. The value of ensemble methods like Random Forest in capturing complex relationships in financial data

3. The need for careful feature engineering and selection to improve model performance

4. The effectiveness of ROC-AUC as a metric for comparing model performance in imbalanced classification problems

*C. Recommendations for Financial Institutions*

1. Focus on monitoring recent payment behavior (last 3-4 months) as a key indicator of default risk.

2. Develop early intervention strategies for customers showing signs of payment difficulties in recent months.

3. Regularly reassess credit limits based on customers’ payment behavior and other relevant factors.

4. Implement a tiered risk assessment system that weighs recent payment history more heavily than long-term payment amounts.

5. Consider using ensemble methods like Random Forest for more accurate default prediction.

*D. Future Work*

1. Explore other advanced algorithms such as Gradient Boosting (e.g., XGBoost, LightGBM) for potential performance improvements.

2. Investigate more sophisticated feature engineering techniques, including time-series analysis of payment patterns.

3. Develop an ensemble model combining multiple algorithms to leverage their individual strengths.

4. Conduct a more in-depth analysis of feature interactions and their impact on default prediction.

5. Explore the potential for developing a real-time risk assessment system based on the trained models.

6. Investigate the economic impact of model predictions by conducting a cost-benefit analysis of different classification thresholds.

VI. REFERENCES

1. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.
2. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136.
3. 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.
4. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
5. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.
6. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
7. McKinney, W. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51-56).
8. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(3), 90-95.
9. Waskom, M. L. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.
10. Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research, 18(1), 559-563.