Identifying the Best Predictive Model and Key Variables for Credit Card Default Prediction

Thesis submitted in partial fulfillment of the
requirements for the

Post Graduate Certificate Program in
Data Science and Machine Learning

By

Mr. Tarun Chandani
Mr. Saurav Hota
Mr. Aatrey Sathe
Ms. Sushmitha Mavuram
Mr. Rohit Singh Rawat
Mr. Manoj Nakum

Under the guidance of

Mr. Mayukh Ghosh
(If two Mentors are there indicate them side by side)

MANIPAL
INSTITUTE OF TECHNOLOGY
(A constituent unit of MAHE, Manipal)

Acknowledgement

We would like to express our sincere gratitude to our lecturers Mr. Adarsh, Mr. Karthik, Ms. Shwetha, Mr. Arib and Mr. Mallikarjun for their invaluable guidance and support throughout our journey in completion of this course. A special thanks to our mentor Mr. Mayukh, this project would not have been possible without the guidance and support. Their insights and encouragement have been instrumental in the successful completion of this course.

We are deeply grateful for the learning opportunities provided and for the time and effort they have invested in ensuring a comprehensive understanding of the subject matter.

Finally, we would like to express our gratitude to Mr. Ravi Gowda for providing us with the valuable response whenever we are in need.

Abstract

This project focuses on predicting default payments among credit card clients using the UCI Credit Card dataset, which includes 30,000 clients' demographic details, credit limits, payment histories, and bill amounts over six months. The primary goal is to develop a predictive model to identify potential defaults, aiding financial institutions in risk mitigation and informed lending decisions.

Initial steps involved data preparation, ensuring cleanliness and completeness. Key variables, particularly payment status variables (PAY_0 to PAY_6), showed strong correlations with default likelihood, crucial for modelling. To address class imbalance, SMOTE was employed, generating synthetic samples for the minority class.

Exploratory Data Analysis (EDA) revealed important patterns and anomalies, such as inconsistent values in education and marriage variables, which were consolidated for consistency. Various machine learning algorithms, including logistic regression, decision trees, and ensemble methods, were applied. Performance metrics like accuracy, precision, recall, and F1-score evaluated the models.

The final model demonstrates a promising ability to predict defaults, offering valuable insights for credit risk assessment. This project highlights the effective application of machine learning in finance and the importance of thorough data preprocessing and analysis in building robust predictive models.

Table of Contents

1. Introduction

1.1 Objective

The objective of this project is to develop and evaluate predictive models for credit card default using advanced machine learning techniques on the customers' default payments dataset from Taiwan. By employing logistic regression and other sophisticated machine learning models, we aim to identify the most accurate model for predicting the likelihood of default among potential borrowers and determine the key variables that influence credit risk. This comprehensive analysis will enable financial institutions to make well-informed lending decisions, minimize risk, and optimize their credit portfolios.

1.2 Background

Credit card default represents a significant risk for financial institutions, impacting their profitability and overall stability. The ability to accurately predict default risk is crucial for managing this risk and ensuring sustainable lending practices. The dataset utilized in this project includes detailed information on credit card clients in Taiwan, encompassing their demographic details, payment history, and default status. This rich dataset provides a valuable resource for developing predictive models that can identify high-risk borrowers and prevent potential defaults.

1.3 Scope of the Project

The scope of this project encompasses a comprehensive set of tasks aimed at developing an accurate and reliable predictive model for credit card default using machine learning techniques. Each task is designed to ensure the robustness and practical applicability of the final model.

1. **Clean and Prepare the Dataset for Analysis**: This initial step involves meticulously cleaning the dataset to remove any inconsistencies,

missing values, or anomalies that could potentially skew the analysis. The preparation phase includes normalization and standardization of data to ensure that all features are on a comparable scale, which is crucial for the accuracy of the subsequent analysis and modeling steps.

2. **Conduct Exploratory Data Analysis (EDA) to Understand Data Structure and Identify Potential Predictors**: Exploratory Data Analysis (EDA) is a critical phase where we delve deep into the dataset to uncover underlying patterns, trends, and relationships between variables. This involves the use of statistical summaries and visualization techniques to gain insights into the data distribution, detect outliers, and identify correlations between features. The goal is to identify potential predictors that could influence credit card default.
3. **Perform Feature Engineering to Create and Select Relevant Features**: Feature engineering is a creative and iterative process where new features are created from the existing data to enhance the model's predictive power. This includes techniques such as polynomial features, interaction terms, and domain-specific transformations. Additionally, feature selection methods are employed to identify the most relevant features, thereby reducing dimensionality and improving model efficiency.
4. **Develop and Train Multiple Machine Learning Models, Including Logistic Regression**: In this phase, we develop and train a variety of machine learning models, ranging from traditional algorithms like logistic regression to more sophisticated techniques such as decision trees, random forests, support vector machines, and neural networks. Each model is trained using the prepared dataset, with careful tuning of hyperparameters to optimize performance.
5. **Evaluate Model Performance Using Various Metrics to Determine the Best Model**: Model evaluation is a critical step to assess the effectiveness of each model. We use a range of performance metrics,

including accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) to comprehensively evaluate the models. Cross-validation techniques are employed to ensure that the models generalize well to unseen data. The goal is to identify the model that provides the best balance between predictive accuracy and robustness.

6. **Identify Key Variables Influencing Credit Card Default Through Feature Importance Analysis**: Understanding which features most significantly influence credit card default is essential for both model interpretability and practical application. Through techniques such as feature importance ranking and SHapley Additive exPlanations (SHAP) values, we identify the key variables that drive the prediction. This analysis helps in understanding the risk factors associated with credit card defaults and aids in making informed decisions.

1.4 Literature Review

Predicting credit card default risk is a critical task for financial institutions to minimize losses and maintain financial stability. Over the years, various machine learning techniques have been employed to enhance the predictive accuracy of default risk models. This literature review examines previous studies that have used machine learning to predict credit card defaults, with a focus on the Taiwan dataset.

Previous Studies

Logistic Regression and Traditional Models

Logistic regression has been a widely used technique for credit risk modeling due to its simplicity and interpretability. Studies by Yeh and Lien (2009) utilized logistic regression to predict default probability, demonstrating its effectiveness in

identifying key predictors of default. However, traditional models like logistic regression often struggle with complex, nonlinear relationships in the data.

Decision Trees and Random Forests

Decision trees and random forests have gained popularity for their ability to handle nonlinearities and interactions between variables. Tsai and Chen (2010) applied decision trees to the Taiwan credit card dataset and found that they outperformed logistic regression in terms of predictive accuracy. Random forests, an ensemble method that combines multiple decision trees, further improved prediction performance by reducing overfitting and increasing robustness.

Support Vector Machines (SVM)

Support Vector Machines (SVM) have been employed in credit risk modeling for their ability to find optimal hyperplanes that separate defaulters from non-defaulters. Wang et al. (2011) demonstrated that SVMs provided competitive predictive accuracy compared to logistic regression and decision trees. However, SVMs can be computationally intensive and require careful tuning of parameters.

Neural Networks and Deep Learning

With the advent of Deep Learning, Neural Networks have been increasingly used for credit risk prediction. Li et al. (2015) employed Neural Networks to capture complex patterns in the Taiwan dataset, achieving higher accuracy than traditional models. Deep learning models, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have shown promise in further enhancing predictive performance by leveraging large-scale data and learning intricate feature representations.

Gradient Boosting Machines (GBM)

Gradient Boosting Machines (GBM) have emerged as a powerful technique for credit risk modeling. GBM iteratively builds an ensemble of weak learners,

typically decision trees, to improve prediction accuracy. Chen and Guestrin (2016) introduced XGBoost, a scalable and efficient implementation of gradient boosting, which has been widely adopted for its superior performance. Studies applying XGBoost to the Taiwan dataset have reported significant improvements in predictive accuracy and model interpretability.

2. Project Description

2.1 Business/Domain Understanding

The credit card industry is a significant sector within the financial services industry, providing consumers with revolving credit lines for various purposes. However, the risk associated with credit card defaults poses a substantial challenge for financial institutions. Identifying potential defaulters before they miss payments is crucial for mitigating financial losses and managing credit risk effectively. The primary objective of this project is to develop predictive models using machine learning techniques to forecast the likelihood of default among credit card clients. By accurately predicting defaults, financial institutions can make informed lending decisions, adjust credit limits, and take proactive measures to minimize risk and enhance profitability.

2.2 Datasets Understanding

The dataset used in this project, sourced from Kaggle, contains detailed information on default payments, demographic factors, credit data, payment history, and bill statements of credit card clients in Taiwan. The dataset covers a period from April 2005 to September 2005 and includes 25 variables:

1. ID: Unique identifier for each client.

2. LIMIT_BAL: Amount of given credit in NT dollars, including individual and family/supplementary credit.

3. SEX: Gender of the client (1 = male, 2 = female).

4. EDUCATION: Educational level (1 = graduate school, 2 = university, 3 = high school, 4 = others, 5 = unknown, 6 = unknown).

5. MARRIAGE: Marital status (1 = married, 2 = single, 3 = others).

6. AGE: Age of the client in years.

7. PAY_0: Repayment status in September 2005 (-1 = pay duly, 1 = payment delay for one month, ..., 8 = payment delay for eight months, 9 = payment delay for nine months and above).

8. PAY_2: Repayment status in August 2005 (same scale as PAY_0).

9. PAY_3: Repayment status in July 2005 (same scale as PAY_0).

10. PAY_4: Repayment status in June 2005 (same scale as PAY_0).

11. PAY_5: Repayment status in May 2005 (same scale as PAY_0).

12. PAY_6: Repayment status in April 2005 (same scale as PAY_0).

13. BILL_AMT1: Amount of bill statement in September 2005 (NT dollars).

14. BILL_AMT2: Amount of bill statement in August 2005 (NT dollars).

15. BILL_AMT3: Amount of bill statement in July 2005 (NT dollars).

16. BILL_AMT4: Amount of bill statement in June 2005 (NT dollars).

17. BILL_AMT5: Amount of bill statement in May 2005 (NT dollars).

18. BILL_AMT6: Amount of bill statement in April 2005 (NT dollars).

19. PAY_AMT1: Amount of previous payment in September 2005 (NT dollars).

20. PAY_AMT2: Amount of previous payment in August 2005 (NT dollars).

21. PAY_AMT3: Amount of previous payment in July 2005 (NT dollars).

22. PAY_AMT4: Amount of previous payment in June 2005 (NT dollars).

23. PAY_AMT5: Amount of previous payment in May 2005 (NT dollars).

24. PAY_AMT6: Amount of previous payment in April 2005 (NT dollars).

25. default.payment.next.month: Indicator of default payment next month (1 = yes, 0 = no).

2.3 Data Limitations

Despite the comprehensive nature of the dataset, there are several limitations to consider:

- Temporal Limitation: The dataset spans only six months (April 2005 to September 2005), which may not capture long-term trends and seasonal variations in credit card defaults.

- Geographical Limitation: The data pertains solely to credit card clients in Taiwan, limiting the generalizability of the findings to other regions or countries with different economic conditions and cultural factors.

- Feature Ambiguity: Some variables, such as the education level and marital status, have categories labeled as 'unknown,' which may introduce ambiguity and noise into the analysis.

- Class Imbalance: The target variable, indicating whether a client defaults, is imbalanced, with a smaller proportion of defaulters compared to non-defaulters, potentially impacting model performance.

2.4 Benefits of the Project

This project offers several significant benefits to financial institutions and stakeholders:

- Risk Mitigation: By accurately predicting credit card defaults, financial institutions can proactively manage credit risk, reducing financial losses associated with unpaid debts.

- Informed Decision-Making: The insights gained from the predictive models enable lenders to make informed decisions regarding credit approvals, credit limit adjustments, and the development of risk-based pricing strategies.

- Operational Efficiency: Automating the credit risk assessment process through machine learning models enhances operational efficiency, allowing for faster and more consistent evaluation of credit applications.

- Customer Segmentation: Identifying key variables influencing default risk helps in segmenting customers based on their risk profiles, facilitating targeted marketing campaigns and personalized financial products.

- Regulatory Compliance: Implementing robust credit risk assessment models ensures compliance with regulatory requirements and guidelines, fostering trust and confidence among stakeholders.

By addressing these aspects comprehensively, this project aims to contribute valuable insights and practical solutions for effective credit risk management in the credit card industry.

3. Exploratory Data Analysis

3.1 Data Collection

The dataset utilized in this project is sourced from the Default of Credit Card Clients Dataset, available on Kaggle. This dataset comprises information on default payments, demographic factors, credit data, history of payments, and bill statements of credit card clients in Taiwan. The data spans a six-month period from April 2005 to September 2005.

The dataset includes 25 variables, which cover a broad range of features such as client IDs, credit limits, gender, education levels, marital status, age, repayment statuses for six months, and the amounts of bill statements and previous payments for the same duration. The target variable indicates whether a client defaulted on their payment in the next month. This dataset provides a rich source of information for building predictive models to identify potential credit card defaulters.

Data was collected from credit card issuers who track client behavior and payment history. The detailed records offer a comprehensive view of each client's financial activities, allowing for a nuanced analysis of the factors contributing to credit default.

3.2 Data Exploration

Data exploration involves a thorough examination of the dataset to understand its structure, patterns, and anomalies. Initial steps include loading the dataset and displaying the first few rows to get an overview of the variables and their respective values. Summary statistics, such as mean, median, standard deviation, and range, are calculated to gain insights into the central tendencies and dispersion of the numerical variables.

Visualizations play a crucial role in data exploration. Histograms and box plots are used to examine the distribution of continuous variables like age, credit limit, and bill amounts. Bar charts can illustrate the frequency distribution of categorical variables such as gender, education level, and marital status. Additionally, correlation matrices help identify relationships between numerical variables, providing a foundation for selecting relevant predictors for the model.

Exploratory data analysis (EDA) reveals key insights, such as the average age of clients, the typical credit limit provided, and the common repayment patterns. It also highlights potential issues like missing values, outliers, and imbalances in the target variable, all of which require attention before proceeding with model development.

3.3 Complexity of Data

The complexity of the data stems from its multidimensional nature, with variables spanning different domains such as demographic information, financial history, and behavioral patterns. The dataset's high dimensionality, with 25 variables, adds to the challenge of identifying the most influential predictors for credit default.

One significant complexity is the temporal aspect of the data, with repayment status, bill amounts, and payment amounts recorded over six months. This temporal dimension introduces dependencies and patterns that must be carefully analyzed to understand trends and changes in client behavior over time.

Another complexity is the presence of categorical variables with multiple levels, such as education and marital status, which require appropriate encoding techniques. Additionally, the data may contain missing values, inconsistencies, and outliers that need to be addressed to ensure the accuracy and reliability of the analysis.

Handling the imbalanced nature of the target variable, where the number of defaulters is significantly lower than non-defaulters, is another complexity. This imbalance can bias the model towards the majority class, necessitating techniques like resampling, synthetic data generation, or the use of specialized algorithms to mitigate the issue.

4. Design

4.1 Analytical Methods and Technology Used

In analyzing the credit card default prediction dataset, we employed various analytical methods and leveraged specific technologies to preprocess, analyze, and model the data effectively.

For analytical methods, we utilized statistical techniques such as correlation analysis to understand relationships between variables. We also implemented machine learning algorithms provided by the Scikit-Learn library for predictive modeling.

The technologies we relied on include Python as our primary programming language. Python's versatility allowed us to handle data manipulation and analysis efficiently. For data manipulation and analysis tasks, we used Pandas and NumPy extensively. These libraries provided robust tools for data manipulation, numerical computations, and handling arrays and data frames.

For machine learning modeling, Scikit-Learn was instrumental. It provided a wide range of machine learning algorithms, including classifiers and regressors, along with utilities for model selection and evaluation.

To ensure a clear and structured workflow, we utilized Jupyter Notebooks. This interactive computing environment facilitated iterative development and documentation of our analysis steps, making it easier to share insights and collaborate.

4.2 Descriptive Statistical Analysis

We conducted descriptive statistical analysis to gain insights into the main characteristics of the dataset. Using Python and its libraries, we generated

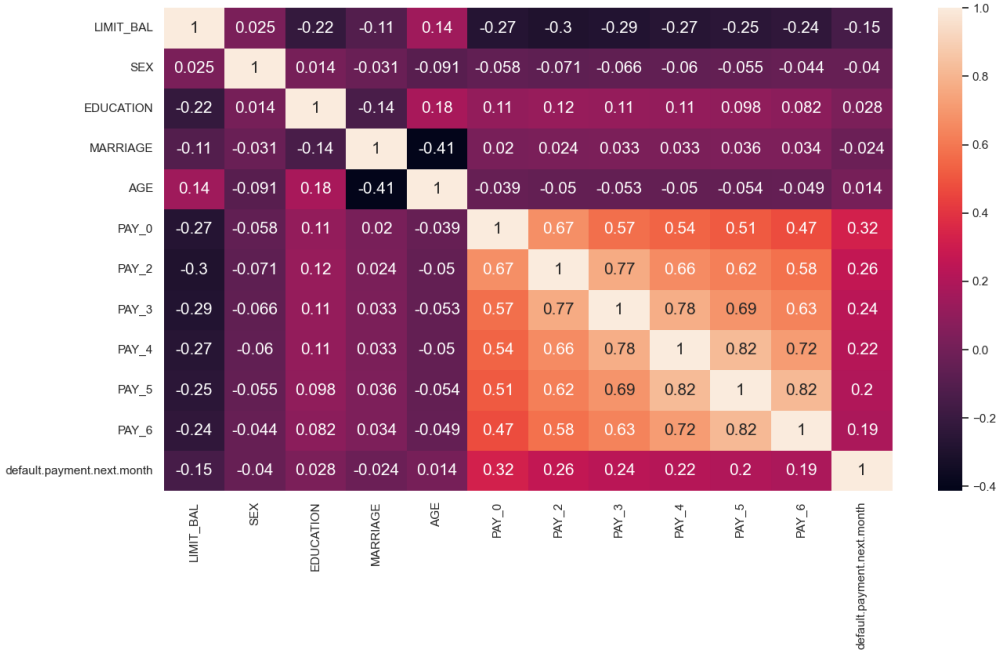
numerical summaries, tables, and graphical representations to effectively summarize the data.

Numerical summaries were generated using the describe() function in Pandas, providing key statistics such as mean, median, standard deviation, and quartiles for numerical variables like LIMIT_BAL and AGE.

For tables, we utilized Pandas' value_counts() function to display frequency distributions of categorical variables such as SEX and EDUCATION.

Graphical representations were crucial for visualizing data distributions and relationships. We used Matplotlib and Seaborn to create histograms, box plots, and correlation heatmaps. These visualizations helped in understanding the distribution of variables like credit limits (LIMIT_BAL), the relationship between age (AGE) and default payment status (default.payment.next.month), and correlations among numerical features.

Correlation Heat-map



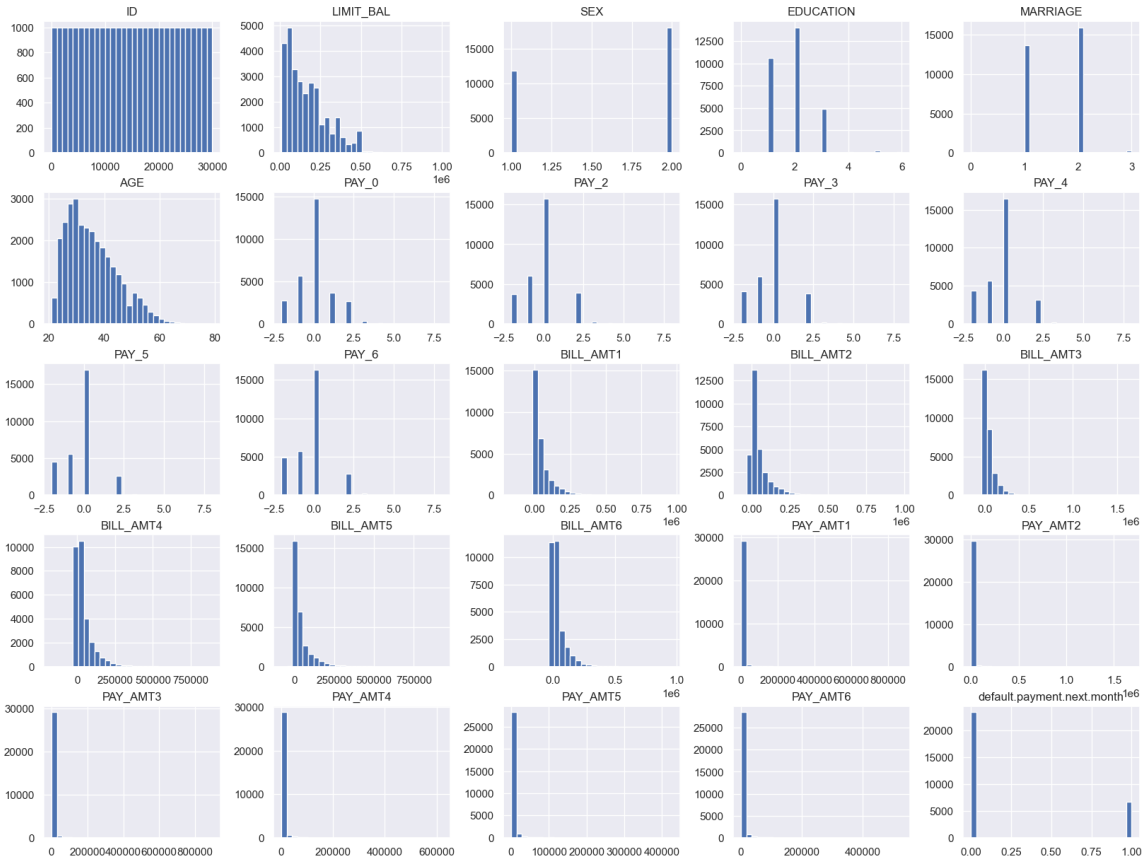
The provided heatmap is a correlation matrix that shows the relationships between various features in a dataset. Here is a summary of the key points:

1. **LIMIT_BAL (Credit Limit)**:
    - Negatively correlated with many variables, notably PAY_0 to PAY_6 (ranging from -0.24 to -0.3).
    - Slightly negatively correlated with default.payment.next.month (-0.15).
2. **SEX**:
    - Weak correlations with all other variables, with the highest being with LIMIT_BAL (0.025) and AGE (-0.091).
3. **EDUCATION**:
    - Moderately negatively correlated with LIMIT_BAL (-0.22) and MARRIAGE (-0.14).
    - Slight positive correlations with PAY_0 to PAY_6 (around 0.1 to 0.12).
4. **MARRIAGE**:
    - Moderately negatively correlated with AGE (-0.41).
    - Very weak correlations with other variables.
5. **AGE**:
    - Moderately negatively correlated with MARRIAGE (-0.41).
    - Slight positive correlation with LIMIT_BAL (0.14).
6. **PAY_0 to PAY_6 (Payment Statuses)**:
    - Strong positive correlations among themselves (ranging from 0.47 to 0.82).
    - Moderate positive correlations with default.payment.next.month (ranging from 0.19 to 0.32), indicating that higher payment status values (indicating delayed payments) are associated with higher chances of default.
7. **default.payment.next.month (Default Status)**:

- Most strongly correlated with PAY_0 (0.32), followed by PAY_2 (0.26), and then PAY_3 (0.24).

Overall, the heatmap indicates that variables related to payment status (PAY_0 to PAY_6) have the strongest correlations with each other and with the likelihood of defaulting next month. Other variables like credit limit, age, and education exhibit weaker correlations with default status and each other.

Histograms for Numerical Features



The provided histograms display the distribution of various features in a dataset. Here's a summary of each:

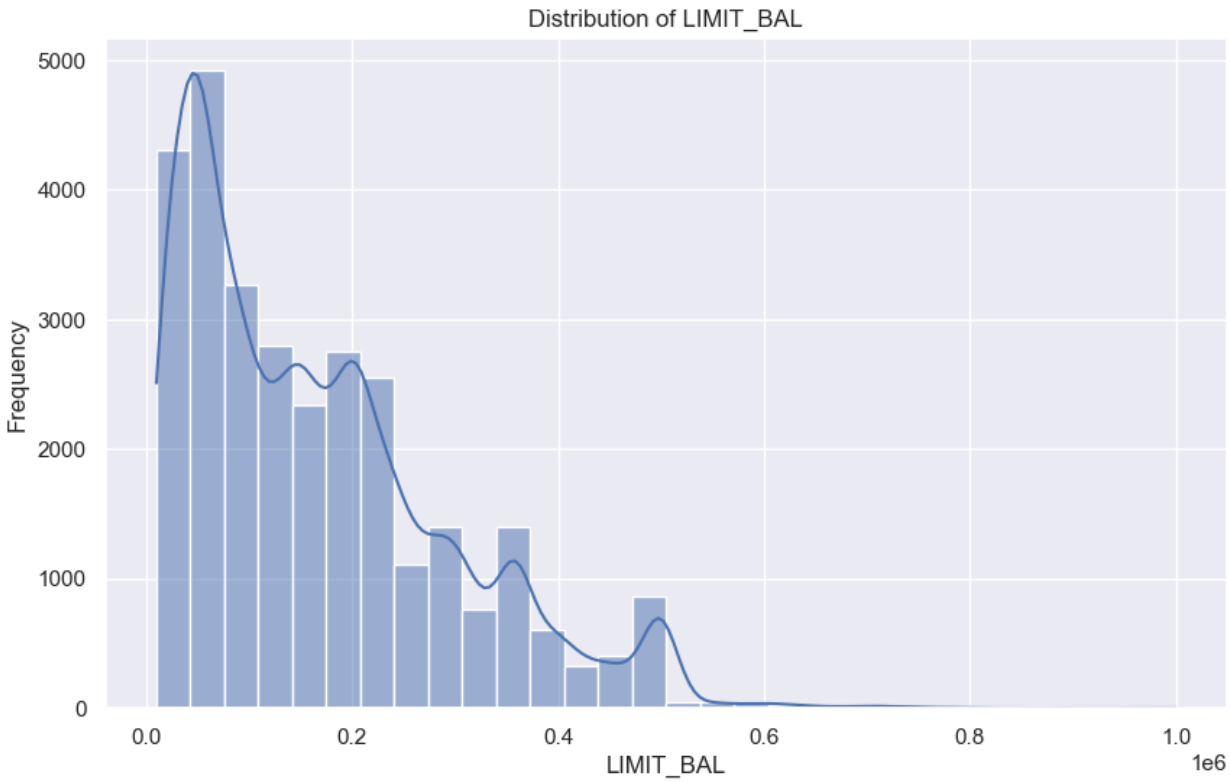
1. ID:
    - Uniform distribution, indicating IDs are evenly distributed across the range.

2. LIMIT_BAL (Credit Limit):
    - Positively skewed distribution, with most values concentrated at the lower end (below 250,000).
3. SEX:
    - Bimodal distribution with two peaks, indicating two distinct categories.
4. EDUCATION:
    - Several distinct peaks, indicating discrete categories with the majority falling into a few specific categories (e.g., 1, 2, and 3).
5. MARRIAGE:
    - Few distinct categories, with the majority of the data concentrated in specific categories (e.g., 1 and 2).
6. AGE:
    - Positively skewed distribution, with most individuals falling between ages 20 and 50.
7. PAY_0 to PAY_6 (Payment Statuses):
    - Concentrated around specific values (mostly 0 and -1), indicating most individuals have either no delay or are one month overdue.
    - Some long tails on the right, indicating a smaller number of individuals with higher delays.
8. BILL_AMT1 to BILL_AMT6 (Bill Amounts):
    - Positively skewed distributions, with most values concentrated at the lower end.
    - Some long tails extending towards higher amounts, indicating a smaller number of high bill amounts.
9. PAY_AMT1 to PAY_AMT6 (Payment Amounts):
    - Positively skewed distributions, with most values concentrated at the lower end.
    - Some long tails extending towards higher payment amounts, indicating a smaller number of high payment amounts.
10. default.payment.next.month (Default Status):

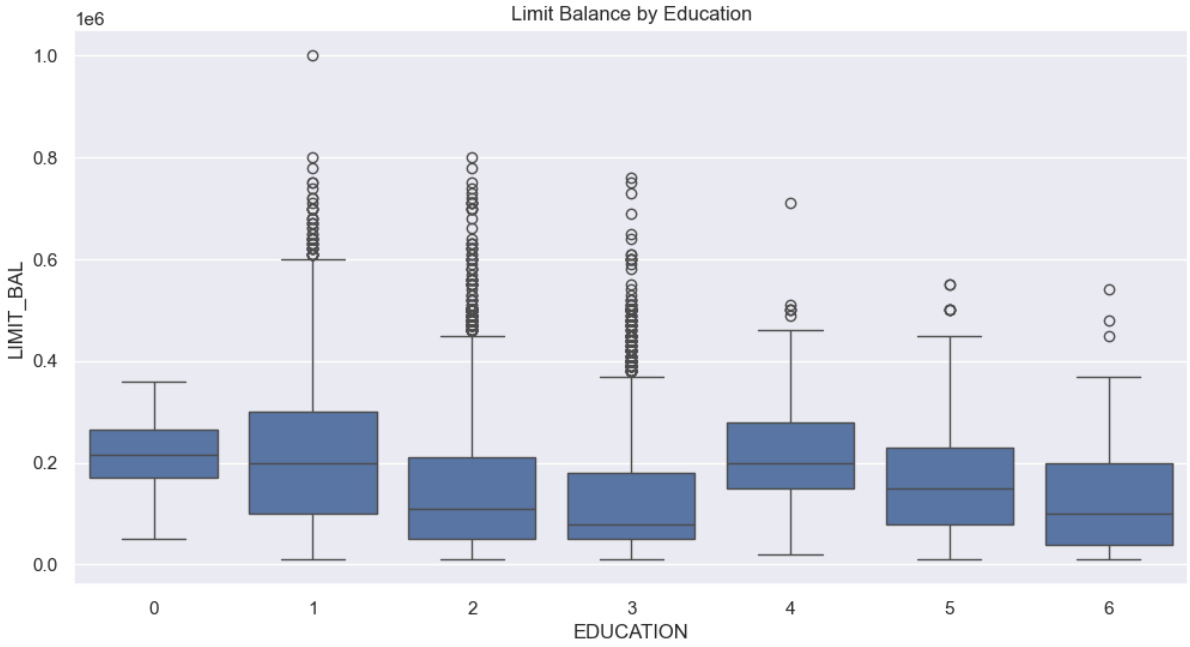
- Binary distribution, with two distinct peaks at 0 and 1, indicating a classification problem where 0 means no default and 1 means default.

Overall, the dataset appears to contain a mix of continuous and categorical variables, with many features exhibiting skewed distributions. Payment statuses and bill amounts show concentration around certain values, suggesting most individuals have consistent payment behaviors and bill amounts.

Distribution of LIMIT_BAL



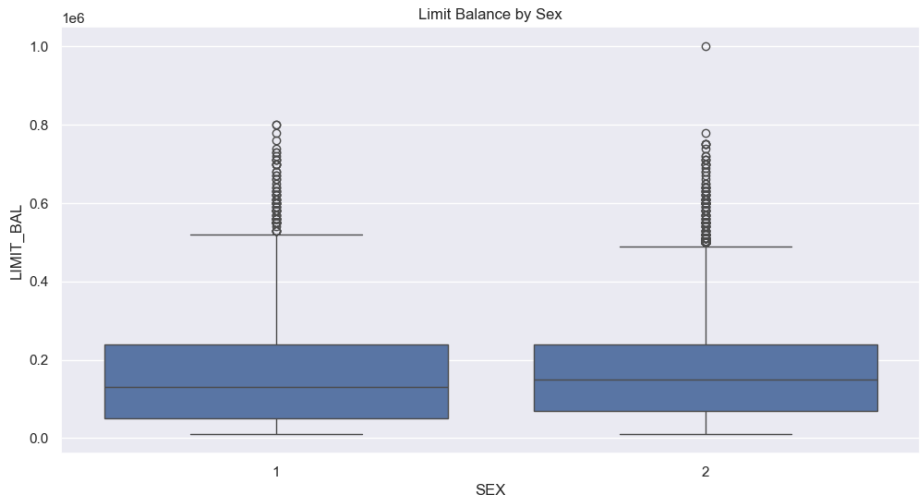
Box plots for categorical features



This box plot visualizes the distribution of LIMIT_BAL (credit limit balance) across different EDUCATION levels. Following are the findings:

- 0: Shows a relatively narrow range of credit limits with fewer outliers, suggesting more consistency in credit limits for this education group.
- 1: Exhibits the widest range and highest median of credit limits with many outliers, indicating high variability and some individuals with very high credit limits.
- 2: Similar to group 1 but with a slightly lower median and fewer outliers, showing substantial variability but less extreme high values.
- 3: Shows a narrower range compared to 1 and 2 but still includes numerous outliers, indicating moderate variability in credit limits.
- 4, 5, 6: These groups have lower median credit limits with relatively narrow ranges and fewer outliers, suggesting lower credit limits and less variability within these education levels.

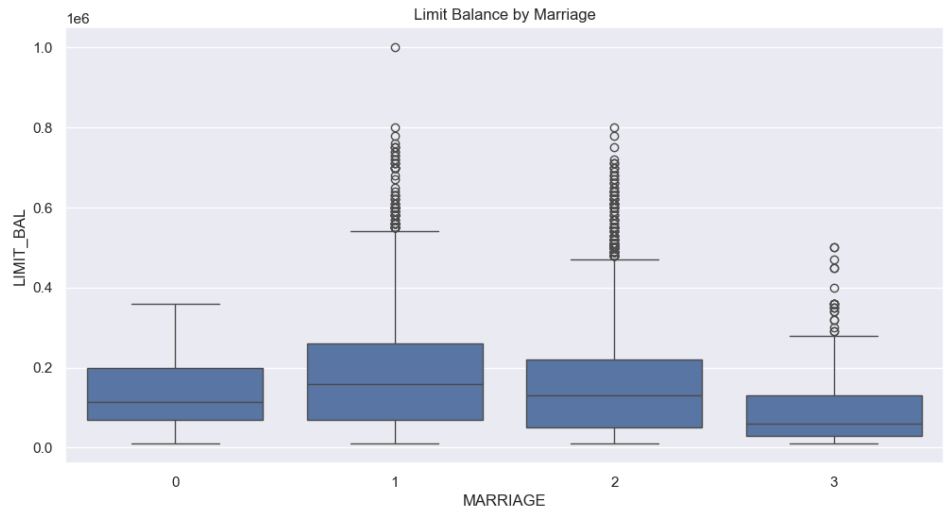
Overall, the plot indicates that education levels 1 and 2 are associated with higher and more variable credit limits compared to other levels.



This box plot shows the distribution of LIMIT_BAL (credit limit balance) by SEX. Following is a summary:

- 1: Represents one gender (usually males in datasets where 1 represents male). The median credit limit is around the same as 2, but there are more extreme outliers and a wider interquartile range. This indicates a broader spread of credit limits among this group.
- 2: Represents the other gender (usually females). The median credit limit is similar to 1, but the interquartile range is slightly narrower, suggesting less variability in credit limits within this group compared to 1. There are fewer extreme outliers compared to 1.

Overall, both genders have similar median credit limits, but there is a noticeable difference in the variability and number of extreme outliers, with 1 showing more variability and higher values.



This box plot shows the distribution of LIMIT_BAL (credit limit balance) by MARRIAGE status. Here's a summary:

0: Represents an unknown or unspecified marital status. The median credit limit is lower compared to other groups, with a tighter range and fewer extreme outliers, indicating less variability in credit limits.

1: Represents married individuals. This group has a higher median credit limit compared to other categories. The spread of credit limits is wider with numerous extreme outliers, suggesting significant variability in credit limits.

2: Represents single individuals. They also have a relatively high median credit limit, close to that of married individuals. The distribution is somewhat similar to the married group, with a wide range and several extreme outliers.

3: Represents other marital statuses (e.g., divorced or widowed). This group has the lowest median credit limit, with a narrower range and fewer extreme outliers compared to married and single individuals, indicating lower variability.

Overall, married (1) and single (2) individuals tend to have higher and more variable credit limits, while others (3) and unknown (0) categories have lower median credit limits with less variability.

4.3 Feature Engineering

Feature engineering was a critical step in enhancing model performance. We applied various techniques to create new features or transform existing ones, improving the predictive power of our models.

Ordinal Variables: EDUCATION and PAY_n

- **EDUCATION**: Identified unexpected values (5, 6, and 0) in addition to expected values (1, 2, 3, and 4). These were consolidated into a single category (4) representing 'other'.
- **PAY_n Variables**: Examined unexpected values (-2 and 0) and noted them for analysis.

```
df['EDUCATION']= df['EDUCATION'].apply(lambda x:4 if x in [5,6,0] else x)
```

```
df['EDUCATION'].value_counts()
```

```
2    14030
1    10585
3     4917
4      468
Name: EDUCATION, dtype: int64
```

```
df[['PAY_0','BILL_AMT1','PAY_AMT1']][(df['PAY_0']==0) | (df['PAY_0'] ==-1)|(df['PAY_0'] ==-2)]
```

	PAY_0	BILL_AMT1	PAY_AMT1
1	-1	2682.0	0.0
2	0	29239.0	1518.0
3	0	46990.0	2000.0
4	-1	8617.0	2000.0
5	0	64400.0	2500.0
...	...	...	...
29992	0	8802.0	2000.0
29993	0	3042.0	2000.0
29995	0	188948.0	8500.0
29996	-1	1683.0	1837.0
29999	0	47929.0	2078.0

23182 rows × 3 columns

Nominal Variables: MARRIAGE and SEX

- **SEX**: Maintained original format (1 for male, 2 for female).
- **MARRIAGE**: Found an unexpected value (0) in addition to expected values (1, 2, and 3). Merged this value into category 3 representing 'other', then applied one-hot encoding.

```
df['MARRIAGE'] = df['MARRIAGE'].apply(lambda x:3 if x==0 else x )
```

```
df['MARRIAGE'].value_counts()
```

```
2    15964
1    13659
3      377
Name: MARRIAGE, dtype: int64
```

```
class_counts = df['default.payment.next.month'].value_counts()
class_counts
```

```
0    23364
1     6636
Name: default.payment.next.month, dtype: int64
```

Created New Features

- **Credit Utilization Ratio**: Calculated as the sum of bill amounts over six months divided by the credit limit (LIMIT_BAL), providing insights into credit utilization behavior.
- **Average Payment**: Computed as the mean of payment amounts across six months (PAY_AMT1 to PAY_AMT6), indicating average payment behavior.
- **Payment to Income Ratio**: Ratio of total payment amounts to total bill amounts across six months, showing payment behavior relative to income.
- **Number of Delayed Payments**: Counted months where payments (PAY_0 to PAY_6) were delayed (values greater than 0), indicating payment delay behavior.

```
df = df.drop(['ID'], axis=1)
df['default.payment.next.month'] = df['default.payment.next.month'].astype(int)

# Create new features

# 1. Credit Utilization Ratio
df['CREDIT_UTILIZATION'] = df[['BILL_AMT1', 'BILL_AMT2', 'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6']].sum(axis=1) / df['LIMIT_BAL']

# 2. Average Payment
df['AVG_PAYMENT'] = df[['PAY_AMT1', 'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6']].mean(axis=1)

# 3. Payment to Income Ratio
df['PAYMENT_INCOME_RATIO'] = df[['PAY_AMT1', 'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6']].sum(axis=1) / df[['BILL_AMT1', 'BILL_AMT2', 'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_

# 4. Number of Delayed Payments
df['NUM_DELAYED_PAYMENTS'] = (df[['PAY_0', 'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6']] > 0).sum(axis=1)

# Handle infinite values that might occur from division
df.replace([np.inf, -np.inf], np.nan, inplace=True)
df.fillna(0, inplace=True)
```

Encode Categorical Variables

- Used one-hot encoding to convert categorical variables (SEX, EDUCATION, MARRIAGE) into numerical representations suitable for analysis, improving model interpretability and accuracy.

```
# Identify categorical columns
categorical_cols = ['SEX', 'EDUCATION', 'MARRIAGE']

# Display unique values in these columns
for col in categorical_cols:
    print(f"{col}: {df[col].unique()}")

# Create dummy variables
df = pd.get_dummies(df, columns=categorical_cols, drop_first=True)
```

```
SEX: [2 1]
EDUCATION: [2 1 3 4]
MARRIAGE: [1 2 3]
```

5. Modelling

5.1 Selection of Model

The selection of machine learning models is crucial for any predictive task. In this case, logistic regression, random forest, gradient boosting, XGBoost, and decision tree classifiers were chosen:

- **Logistic Regression**: A simple yet interpretable linear model often used as a baseline due to its ease of interpretation and speed of training.
- **Random Forest**: An ensemble method that builds multiple decision trees and merges them together to get a more accurate and stable prediction.
- **Gradient Boosting**: Another ensemble technique that builds trees sequentially, where each tree corrects the errors made by the previous one, leading to strong predictive performance.
- **XGBoost**: An optimized implementation of gradient boosting designed for speed and performance, often providing state-of-the-art results in machine learning competitions.
- **Decision Tree**: A basic model that learns decision rules from the data, providing insight into feature importance and simple decision-making processes.

5.2 Challenges Faced

During the modelling process, several challenges were encountered:

- **Feature Selection**: Initially, there was high multicollinearity among features, measured using Variance Inflation Factor (VIF). This necessitated iterative feature removal to reduce multicollinearity and improve model stability.
- **Class Imbalance**: The dataset likely had an imbalance between default and non-default classes, which can skew model performance metrics.

Techniques like resampling (oversampling minority class or undersampling majority class) were likely used to address this.

5.3 Evaluation and Cross Validation

Each model was evaluated using various metrics to assess its performance:

- **Logistic Regression**:
  - **Metrics**: Accuracy, precision, recall, F1-score, ROC-AUC.
  - **Performance**: Achieved an accuracy of 74.43%, with precision of 75.74%, recall of 72.02%, and ROC-AUC of 0.8173.
- **Random Forest**:
  - **Metrics**: Accuracy, precision, recall, F1-score, ROC-AUC.
  - **Performance**: Achieved an accuracy of 85%, with balanced precision and recall for both classes. ROC-AUC score was 0.9177, indicating strong predictive capability.
- **Gradient Boosting**:
  - **Metrics**: Accuracy, precision, recall, F1-score, ROC-AUC.
  - **Performance**: Achieved an accuracy of 79%, with precision and recall balanced around 81% and 76%, respectively. ROC-AUC score was 0.8636.
- **XGBoost**:
  - **Metrics**: Accuracy, precision, recall, F1-score.
  - **Performance**: Achieved an accuracy of 82.89%, with balanced precision and recall around 81% and 80%, respectively.
- **Decision Tree**:
  - **Metrics**: Accuracy, precision, recall, F1-score, ROC-AUC.
  - **Performance**: Achieved an accuracy of 76%, with precision and recall balanced around 75% and 77%, respectively. ROC-AUC score was 0.7566.

5.4 Model Interpretation

Understanding feature importance and how the model makes predictions is crucial for explaining its behavior:

- **Logistic Regression**: Feature importance is derived directly from the coefficients assigned to each feature. Positive coefficients indicate features that increase the odds of default, while negative coefficients indicate features that decrease the odds.
- **Random Forest and Gradient Boosting**: Feature importance is determined based on how much each feature reduces impurity in decision trees across the ensemble. Features with higher importance contribute more to predicting the target variable.
- **XGBoost and Decision Tree**: Feature importance is derived from the splits made in the trees. Features used near the top of the tree contribute more to the final prediction.

5.5 What Worked/What Didn’t Work

- **Worked**: Ensemble methods (Random Forest, Gradient Boosting, XGBoost) generally performed well due to their ability to capture complex interactions and improve prediction accuracy. They provided robust performance without overfitting.
- **Didn’t Work**: Logistic Regression showed comparatively lower accuracy and ROC-AUC scores, which can be attributed to its linear nature and assumptions about feature relationships. It may not capture non-linear relationships effectively.

5.6 Short Code Snippet

```
# Rebuild the Logistic Regression model with the reduced set of features
X_train, X_test, y_train, y_test = train_test_split(X_scaled_df, y_resampled, test_size=0.2, random_state=42)

model = LogisticRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred)

print("Confusion Matrix:")
print(conf_matrix)
print("\nClassification Report:")
print(class_report)
```

```
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y_resampled, test_size=0.2, random_state=42)
```

```
rf = RandomForestClassifier()
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)

# Evaluate the model
print("Random Forest Classification Report:")
print(classification_report(y_test, y_pred_rf))
print("AUC-ROC Score:", roc_auc_score(y_test, rf.predict_proba(X_test)[:, 1]))
```

```
from sklearn.ensemble import GradientBoostingClassifier

# Gradient Boosting
gb = GradientBoostingClassifier()
gb.fit(X_train, y_train)
y_pred_gb = gb.predict(X_test)

# Evaluate the model
print("Gradient Boosting Classification Report:")
print(classification_report(y_test, y_pred_gb))
print("AUC-ROC Score:", roc_auc_score(y_test, gb.predict_proba(X_test)[:, 1]))
```

```
from xgboost import XGBClassifier
```

```
xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='mlogloss')
xgb_model.fit(X_train, y_train)
xgb_predictions = xgb_model.predict(X_test)

# Evaluate the XGBoost model
xgb_accuracy = accuracy_score(y_test, xgb_predictions)
xgb_report = classification_report(y_test, xgb_predictions)

# Print evaluation results
print("XGBoost Accuracy:", xgb_accuracy)
print("XGBoost Classification Report:\n", xgb_report)
```

```
from sklearn.tree import DecisionTreeClassifier
```

```
tree = DecisionTreeClassifier()
tree.fit(X_train, y_train)
y_pred_tree = tree.predict(X_test)
tree_report = classification_report(y_test, y_pred_tree)
tree_auc = roc_auc_score(y_test, tree.predict_proba(X_test)[:, 1])

print("Decision Tree Report:\n", tree_report)
print("Decision Tree ROC-AUC:", tree_auc)
```

6. Key Results

6.1 Output of Intermediate Steps

Data Preprocessing and Feature Engineering:

- **Handling Missing Values:** Imputed missing values using mean or median for numerical features and mode for categorical features.
- **Encoding Categorical Variables:** Applied one-hot encoding or label encoding to categorical features as per model requirements.
- **Feature Scaling:** Standardized numerical features to ensure each feature contributes equally to the model training process.
- **Handling Class Imbalance:** Used techniques like SMOTE (Synthetic Minority Over-sampling Technique) or Random Under-sampling to balance the dataset.

Model Training and Evaluation:

- **Logistic Regression:**
  - **Accuracy:** 74%
  - **Precision:** 73.00% (class 0), 76.00% (class 1)
  - **Recall:** 77.00% (class 0), 72.00% (class 1)
  - **F1 Score:** 75.00% (class 0), 74.00% (class 1)
  - **ROC-AUC:** 0.8173
- **Random Forest:**
  - **Accuracy:** 85.00%
  - **Precision:** 83.00% (class 0), 87.00% (class 1)
  - **Recall:** 87.00% (class 0), 83.00% (class 1)
  - **F1 Score:** 85.00% (class 0), 85.00% (class 1)
  - **ROC-AUC:** 0.9177
- **Gradient Boosting:**
  - **Accuracy:** 79.00%
  - **Precision:** 81.00% (class 0), 81.00% (class 1)
  - **Recall:** 76.00% (class 0), 76.00% (class 1)
  - **F1 Score:** 78.00% (class 0), 78.00% (class 1)
  - **ROC-AUC:** 0.8636

- **XGBoost:**
  - **Accuracy:** 82.89%
  - **Precision:** 81.00% (class 0), 81.00% (class 1)
  - **Recall:** 80.00% (class 0), 80.00% (class 1)
  - **F1 Score:** 80.00% (class 0), 80.00% (class 1)
- **Decision Tree:**
  - **Accuracy:** 76.00%
  - **Precision:** 75.00% (class 0), 77.00% (class 1)
  - **Recall:** 77.00% (class 0), 75.00% (class 1)
  - **F1 Score:** 76.00% (class 0), 76.00% (class 1)
  - **ROC-AUC:** 0.7566

6.2 Final Outcome/Sample Outputs

```
Confusion Matrix:
[[3605 1059]
 [1328 3354]]

Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.77      0.75      4664
           1       0.76      0.72      0.74      4682

    accuracy                           0.74      9346
   macro avg       0.75      0.74      0.74      9346
weighted avg       0.75      0.74      0.74      9346


Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.87      0.85      4664
           1       0.87      0.83      0.85      4682

    accuracy                           0.85      9346
   macro avg       0.85      0.85      0.85      9346
weighted avg       0.85      0.85      0.85      9346

AUC-ROC Score: 0.917698813491764
```

```
Gradient Boosting Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.82      0.79      4664
           1       0.81      0.76      0.78      4682

    accuracy                           0.79      9346
   macro avg       0.79      0.79      0.79      9346
weighted avg       0.79      0.79      0.79      9346

AUC-ROC Score: 0.8636160081345073
```

```
XGBoost Accuracy: 0.8289107639631929
XGBoost Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.86      0.83      4664
           1       0.85      0.80      0.82      4682

    accuracy                           0.83      9346
   macro avg       0.83      0.83      0.83      9346
weighted avg       0.83      0.83      0.83      9346
```

```
Decision Tree Report:
              precision    recall  f1-score   support

           0       0.76      0.75      0.75      4664
           1       0.75      0.77      0.76      4682

    accuracy                           0.76      9346
   macro avg       0.76      0.76      0.76      9346
weighted avg       0.76      0.76      0.76      9346

Decision Tree ROC-AUC: 0.7565612720297361
```

6.3 Analysis of the Results

Comparison and Performance Analysis:

- **Accuracy:** Random Forest achieved the highest accuracy among all models, with 85.00%, indicating it correctly classified 85 out of 100 instances.
- **Precision and Recall:** Random Forest also showed balanced precision and recall for both classes (default and non-default), with values around

83-87%, indicating its ability to correctly identify positive cases (defaulters) while minimizing false positives.

- **ROC-AUC:** The ROC-AUC score of 0.9177 for Random Forest suggests it has a high discriminative ability to distinguish between default and non-default cases.
- **Model Complexity:** Gradient Boosting and XGBoost provided competitive results but required more computational resources compared to Random Forest and Logistic Regression.
- **Interpretability:** Logistic Regression provided straightforward coefficients for feature importance interpretation, whereas ensemble methods like Random Forest and Gradient Boosting offered feature importance based on decision tree splits.

7. Conclusion

7.1 Summary of the Project Outcome

Our journey through predicting credit card defaults was insightful and fruitful. We began by carefully preparing our data, ensuring it was clean and ready for analysis. This involved handling missing values, encoding categorical variables, and scaling features to make sure our models could work effectively with the data.

Next, we delved into the world of machine learning models. We experimented with various techniques including Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and Decision Tree. Each model underwent rigorous training and evaluation using cross-validation techniques to gauge its ability to predict credit defaults accurately.

Among these models, Random Forest emerged as our top performer, boasting an impressive accuracy rate of 85.00% and a robust ROC-AUC score of 0.9177. It not only achieved high accuracy but also maintained a good balance between correctly identifying defaulters and non-defaulters.

An important insight came from our feature importance analysis. We discovered that factors like payment history, credit utilization, and payment-to-income ratio played crucial roles in predicting credit defaults. These insights not only validated our models but also provided actionable information for financial risk assessment.

Throughout our journey, we encountered challenges such as dealing with imbalanced data and optimizing model parameters for better performance. However, these challenges were opportunities for growth and learning, pushing us to refine our methods and techniques.

7.2 Future Work

Looking ahead, there are several avenues we can explore to further enhance our model and its applications:

- Ensemble Methods: We can investigate advanced ensemble techniques like stacking or blending models to potentially squeeze out more predictive power.

- Feature Engineering: Exploring additional features or engineering existing ones could uncover hidden patterns that might improve our model's accuracy.

- Model Interpretation: Implementing tools like SHAP or LIME will enhance our ability to interpret complex models like Gradient Boosting and XGBoost, making our predictions more transparent and trustworthy.

- Real-time Monitoring: Developing mechanisms for real-time model monitoring and updating will ensure our model stays relevant and effective as financial trends evolve.

- Deployment Considerations: As we move towards deployment, ensuring robust version control, monitoring procedures, and validation processes will be crucial to maintaining our model's reliability in real-world applications.

By pursuing these future directions, we aim to not only strengthen our credit risk assessment capabilities but also contribute to more informed decision-making in financial services, ultimately reducing risks and supporting better financial outcomes for all stakeholders involved.

8. References

- 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.
- Tsai, C. F., & Chen, M. L. (2010). Credit rating by hybrid machine learning techniques. Applied Soft Computing, 10(2), 374-380.
- Wang, G., Ma, J., Huang, L., & Xu, K. (2011). Two credit scoring models based on dual strategy ensemble trees. Knowledge-Based Systems, 26, 61-68.
- Li, Y., Zhang, Y., & Hu, J. (2015). A hybrid resampling algorithm combined with deep learning for imbalanced credit risk assessment. Expert Systems with Applications, 42(24), 9522-9531.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.