Credit Card Default Prediction

1. Introduction and Domain Knowledge:

The history of credit cards began in the 20th century with the introduction of charge cards in the 1950s, which allowed consumers to make purchases on credit. As credit card usage increased, the need for effective risk management strategies to identify potential defaulters became increasingly important. Historical events, such as the 2008 financial crisis, underscored the necessity for robust predictive models, as rising default rates among credit card holders contributed to systemic risks and financial instability. Consequently, accurately predicting credit defaults is essential for reducing risk and fostering greater stability within the financial system[1].

Credit card default prediction is a crucial task for financial institutions, directly impacting their ability to manage lending risk, optimize credit limits, and mitigate potential losses. When a significant number of customers fail to repay their credit card debts, it leads to an increased likelihood of bad debts, higher interest rates, and stricter credit policies. This situation not only affects individual financial institutions but also places a strain on economic conditions, creating a ripple effect that can influence job markets and consumer spending.

In this project, we aim to develop a predictive model to determine whether a credit card holder will default on their payment in the following month. The analysis will utilize the UCI Default of Credit Card Clients Dataset, which contains records of 30,000 Taiwanese credit card holders. This dataset includes various demographic features, such as gender, age, education level, and marital status, along with historical financial and payment data. The primary target variable in this analysis is whether the customer defaulted on their payment the next month, indicated by a value of 1 for default and 0 for no default.

Through this project, we hope to provide valuable insights that can enhance risk management strategies for financial institutions, ultimately fostering a more stable economic environment.

2. Dataset Analysis & Understanding

In this project, we analyzed the "Default of Credit Card Clients" dataset to develop a predictive model for credit card defaults. The first step we had loaded the dataset and examined its structure, followed by descriptive statistics to understand the distribution of features and identification of missing values

A. Data Characteristics

The dataset utilized in this analysis consists of 30,000 rows and 25 columns, which include one target variable and 24 input features. The input features can be categorized into two main groups[2]:

1. *Demographic Features:*
   - SEX: Gender of the cardholder (1 = male, 2 = female).
   - EDUCATION: Education level of the cardholder (1 = graduate school, 2 = university, 3 = high school, 4 = others).
   - MARRIAGE: Marital status of the cardholder (1 = married, 2 = single, 3 = others).
   - AGE: Age of the credit card holder.

2. *Financial Features:*
   - LIMIT_BAL: Credit limit assigned to the cardholder.
   - BILL_AMT1-6: Amount of bill statements over the last six months.
   - PAY_AMT1-6: Payment amounts made by the cardholder in the last six months.
   - PAY_0 to PAY_6: History of delayed payments over the last six months, where PAY_0 refers to repayment status in September, PAY_2 in August, and so on.

B. Target Variable: The target variable in this analysis indicates whether a credit card holder defaulted on their payment in the following month. This binary classification (1 = default, 0 = no default) is crucial for assessing the model's predictive performance regarding credit risk.

C. **Data Quality Assessment:** We assessed the dataset for missing values and duplicates, ensuring data integrity. The target variable, "default payment next month," was visualized to evaluate class balance, which is crucial for model training.

D. **Feature Selection and Cleaning:** ID Filed were removed to streamline the dataset. A correlation analysis was conducted to identify relationships between features, guiding the selection of relevant predictors.

E. **Data Visualization:** Visualization for made including the distribution of credit limits and payment histories, to gain insights into their distributions and potential outliers.

F. **Preprocessing Steps:** With the help of one-hot encoding the Categorical variables were transformed into a numerical format suitable for machine learning algorithms.

G. **Train-Test Split and Model Preparation**: The dataset was split into training and testing sets, enabling model validation. We implemented N-fold cross-validation to evaluate model performance reliably.

H. **Addressing Class Imbalance:** To handle class imbalance, we applied the Synthetic Minority Over-sampling Technique (SMOTE), which enhanced model training by balancing the distribution of classes.

3. Feature Analysis & Selection

Feature selection is crucial for identifying the most relevant features that contribute to predicting the target variable, "default payment next month."[4][5]

- ***Demographic Features:*** Although features like SEX, AGE, EDUCATION, and MARRIAGE are included in the dataset, initial correlation analysis indicates that they have a weak correlation with the target variable. Nonetheless, these features are retained in the model as they may add value when combined with other relevant features.
- ***Financial Features:*** The Payment History features (PAY_0, PAY_2, etc.) exhibit a strong correlation with the likelihood of default. Customers who have been late on payments in previous months are more likely to default. Additionally,

the BILL_AMT and PAY_AMT series, which represent billing and payment behavior, provide insights into a customer's financial discipline. However, while BILL_AMT shows only a moderate correlation with the target variable, PAY_AMT appears to have a stronger predictive power.

- ***Conclusion****:* In summary, the Payment History features (PAY_0 to PAY_6) and Payment Amount features (PAY_AMT1 to PAY_AMT6) are likely to be the most critical predictors of credit default.

C. Data Cleaning/Preprocessing

Several preprocessing steps were implemented to ensure the dataset is clean and ready for modeling:

- ***Handling Missing Values:*** No missing values were identified in the dataset, eliminating the need for imputation.
- ***Removing Duplicates:*** A small number of duplicate records were detected and removed to maintain data quality.
- ***Dropping Irrelevant Features:*** The ID column, serving as a unique identifier, was dropped due to its lack of predictive power.
- ***Encoding Categorical Variables*****:** Categorical variables such as SEX, EDUCATION, and MARRIAGE were encoded using one-hot encoding to transform them into a numerical format suitable for model training.
- ***Scaling Numeric Features*****:** Features like LIMIT_BAL, BILL_AMT1-6, and PAY_AMT1-6 were scaled using StandardScaler to normalize their distributions. This scaling is particularly important for models like logistic regression that are sensitive to feature scaling.

4. Data Visualization - Independent Features

Visualizing the data is crucial for understanding feature distributions and their relationships with the target variable.

- ***Target Variable Distribution:*** The target variable, indicating default payment next month, is highly imbalanced. The majority of customers (~22,000) did not default (label 0), while only ~5,000 customers did default (label 1). This imbalance necessitates careful consideration in model training to ensure effective learning from the minority class.

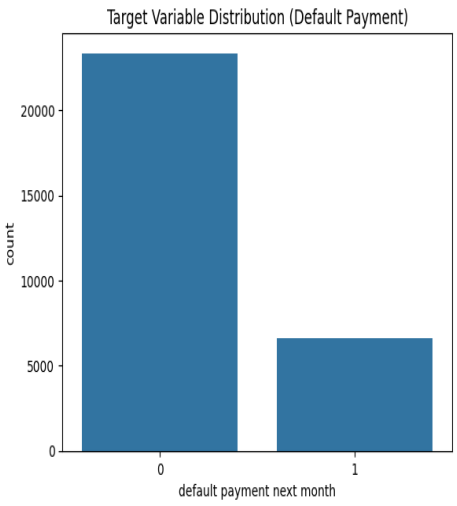

Fig -1:Target Variable Distribution:

- ***Credit Limit Distribution:*** Most customers possess a credit limit below $200,000, with a long tail showing that a small number of customers have significantly higher credit limits. The mean limit balance, represented by a red line in the distribution plot, is slightly above $200,000, highlighting the concentration of customers around lower credit limits.

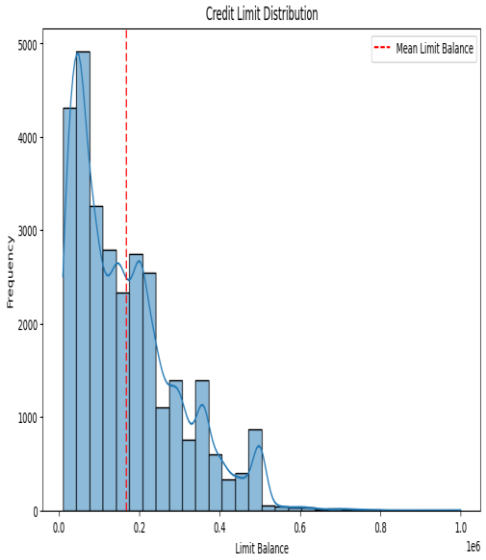

Fig2 - Credit Limit Distribution

- ***Payment History Visualization:***

  A significant portion of customers maintains a repayment status (PAY_0) of 0, indicating on-time payments. Conversely, a smaller group has delayed payments, with very few customers exhibiting severely delayed repayment statuses. This insight into payment behavior is critical for predicting default risk.

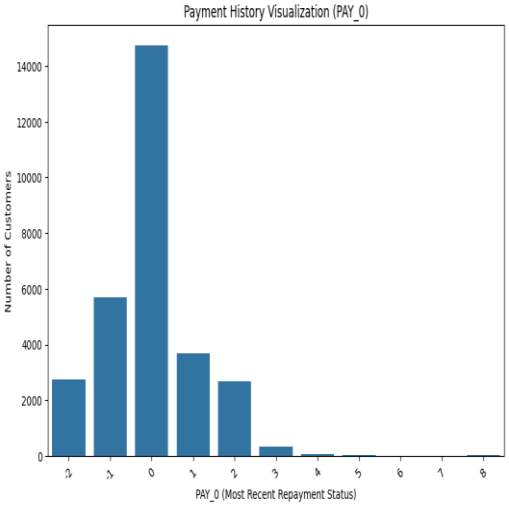

Fig3- Payment History Visualization

- ***Correlation Matrix:*** A correlation heatmap was generated to visualize relationships between features. It revealed that the repayment statuses (PAY_0 to PAY_6) exhibit a positive correlation with the target variable, suggesting that timely payments are associated with a lower likelihood of default. In contrast, the billing amounts (BILL_AMT1-6) demonstrated lower correlations with default behavior, indicating they may not be as influential in predicting default risk.

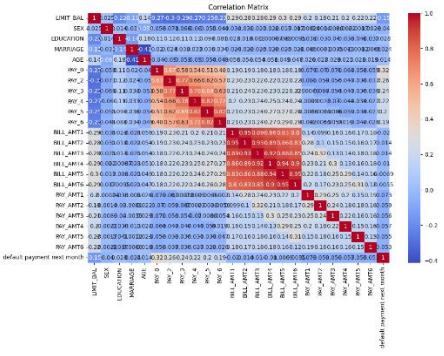

Fig 4: Correlation

- ***Model Performance Metrics***

The performance of the models was evaluated using precision, recall, and f1-score metrics, as summarized in the tables below:

	precision	recall	f1-score	support
0	0.78	1.00	0.88	4673
1	0.29	0.00	0.00	1320
accuracy			0.78	5993
macro avg	0.53	0.50	0.44	5993
weighted avg	0.67	0.78	0.68	5993

	precision	recall	f1-score	support
0	0.84	0.95	0.89	4673
1	0.65	0.36	0.46	1320
accuracy			0.82	5993
macro avg	0.74	0.65	0.67	5993
weighted avg	0.80	0.82	0.79	5993

Table 1: Model Performance Metrics

5. Model Building & Evaluation Plans

Building on the insights gleaned from the exploratory data analysis, the next phase involves developing predictive models. Our initial strategy comprises the following steps:

- ***Balancing the Dataset:*** Given the imbalanced nature of the target variable, we will employ SMOTE (Synthetic Minority Over-sampling Technique) to over-sample the minority class (defaulters). This approach aims to provide the model with a balanced dataset, facilitating improved learning from both classes.
- ***Model Selection:*** We plan to initiate the modeling process with straightforward models such as Logistic Regression and Random Forest Classifier. Both models will be assessed using 10-fold cross-validation to ensure robust generalization. Logistic Regression will be utilized for its interpretability, while Random Forests are preferred for their capacity to capture intricate relationships among features, as evidenced by our earlier experiments where it achieved a maximum accuracy score of 74.1% and an F1 score of 75.0%.
- ***Evaluation Metrics:*** Due to the dataset's imbalance, relying solely on accuracy metrics will be inadequate. Therefore, we will employ additional metrics such as Precision, Recall, F1-score, and ROC-AUC to comprehensively evaluate model performance, particularly focusing on the minority class (defaulters).
- ***Deployment:*** Following training and evaluation, the final stage involves constructing a Gradio interface to facilitate user interaction with the model, allowing predictions on whether a new customer is likely to default.

6. Conclusion:

1. Lessons Learned:

Throughout the process of developing the credit card default prediction model, I had learned important insights into both the data mining techniques and the importance of scientific writing. One of the key lessons was the importance of understanding the domain context when working on real-world datasets. The historical perspective on credit card usage and default rates, especially in case of events like the 2008 financial crisis, highlighted the critical nature of accurate predictive modelling in financial settings. Working with the "Default of Credit Card Clients" dataset presented challenges that went beyond what we typically encounter in clean and homogeneous datasets. The need for thorough data cleaning and preprocessing was important ,

particularly in handling categorical features and ensuring data integrity. One major experience was learning about the implementation of one-hot encoding on model dimensionality, which led to a clear understanding of the "dummy variable trap" and its impact on performance metrics. The exploration of class imbalance was another area of growth. Researching techniques like SMOTE for oversampling the minority class introduced me to effective strategies for enhancing model training. Additionally, I recognized that feature selection and engineering are critical steps in building a robust model. The correlation analysis conducted during this phase [7]

provided valuable insights, particularly in identifying which features held the most predictive power. Finally, the iterative nature of model evaluation and parameter tuning challenged me to adopt a more systematic approach. Implementing cross-validation and experimenting with various feature had made my understanding clean on how different aspects of data preparation and model selection can influence outcomes.

on precision, recall, and F1-score as key performance metrics. Overall, the insights gained thus far provide a clear direction for the development of a robust predictive model, which will enhance credit risk assessment and decision-making processes in the financial sector.

2. *Future Considerations:*

Explore sophisticated models like ensemble methods or deep learning to enhance predictive accuracy. Investigate advanced imputation techniques for handling missing data. Continue exploring methods to address class imbalance, such as cost-sensitive learning[6].

3. *Conclusion:*

By working with data on the UCI Default of Credit Card Clients Dataset, we found that payment history features significantly influence default likelihood, while demographic factors show minimal correlation. The imbalanced nature of the dataset necessitates specific strategies like SMOTE for effective model training. We aim to implement Logistic Regression and Random Forest Classifier, focusing

References:

1. Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33(4), 847-856. https://doi.org/10.1016/j.eswa.2006.07.007
2. Xu, Z., Chu, Q., Song, X., Hu, P., & Peng, L. (2023). Explainable prediction of loan default based on machine learning models. *Data Science and Management*, 6(3), 123-133. https://doi.org/10.1016/j.dsm.2023.04.003
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
4. Fuster, A., Goldsmith Pinkham, P., Ramadorai, T., et al. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(1), 5-47. https://doi.org/10.1111/jofi.12915

5. Giudici, P., Hadji-Misheva, B., & Spelta, A. (2020). Network-based credit risk models. *Quality Engineering*, 32(2), 199-211. https://doi.org/10.1080/08982112.2019.1655159
6. Le Cun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, 521, 436-444. https://doi.org/10.1038/nature14539
7. Lyn, T., Edelman, D., & Crook, J. (2002). *Credit Scoring and its Applications*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898718317