# Introduction

The dataset provides information about default payments of customers in Taiwan, consisting of 25 variables, and a total of 30,000 rows of data. The variables include details such as customer ID, given credit amount, gender, education level, marital status, age, history of past payments from April to September 2005, bill statement amounts from April to September 2005, and previous payment amounts from April to September 2005. The target variable for machine learning prediction represents the statement of default payment in the next month.

The dataset will serve two purposes. Task 1, will be employed to predict whether a customer will default on the next payment using three classification models. In Task 2, the dataset will be utilized to predict the limit balance for a new customer using three regression models.

# Data Exploration

## Handling Invalid Data

By categorizing meaningless numerical values as "others," I make the representation of features more intuitive and interpretable, preserving the essential characteristics of the data. For example, for the Education feature, the original data had numerical values representing different levels of education, such as 1 for graduate education, 2 for undergraduate education, and 3 for high school education. Numbers 4, 5, 6, and 0, which originally lacked clear meaning, were grouped into a category, namely category 4, representing "others." This simplification streamlines the representation of the Education feature while retaining the essential information. The associated code is shown in Figure 1.

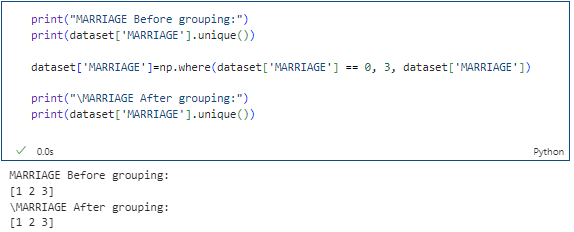
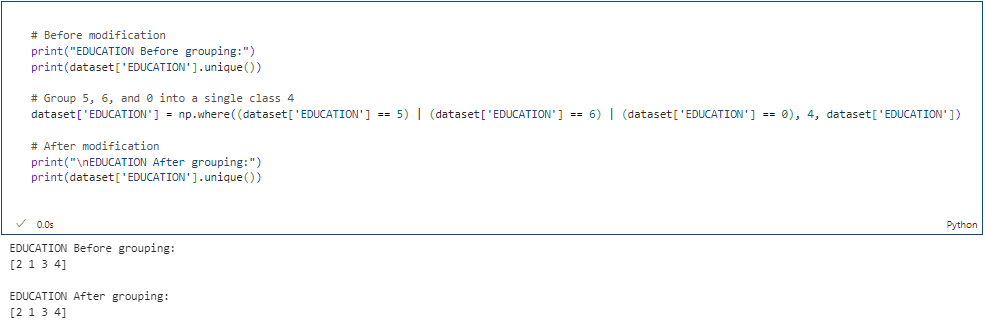


Figure 1. Code of handling invalid data

## 2.2 Aggregate statistics on the data

The average credit card limit stands at 167,484.3 NT dollars, exhibiting a standard deviation of 129,747.3 NT dollars, encompassing a range from 10,000 to 1 million NT dollars. Education-wise, most individuals hold either a master's or undergraduate degree. Marital status within the dataset is predominantly composed of married or single individuals. The mean age is 35.5 years, accompanied by a standard deviation of 9.2 years. The associated code is shown in Figure 2.

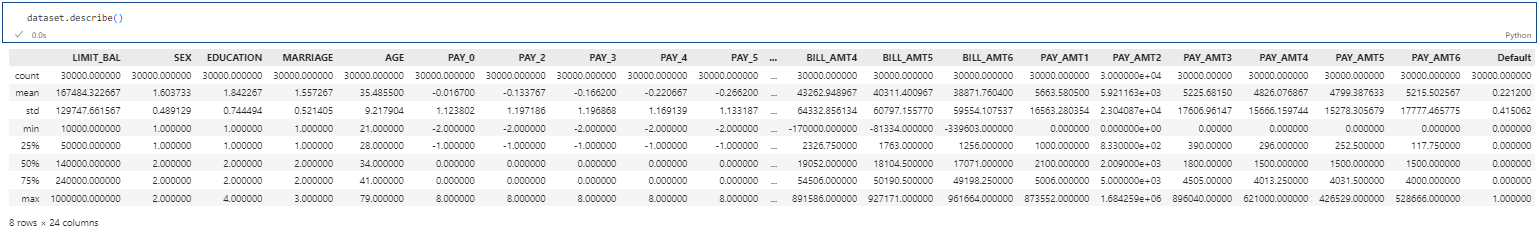


Figure2. Code of data description

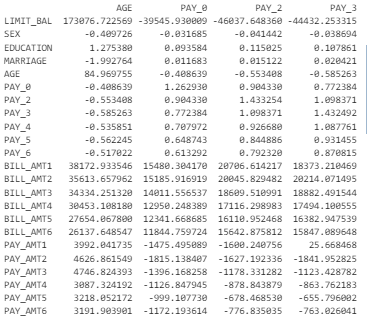
The information on aggregate statistics is shown in Table 1.

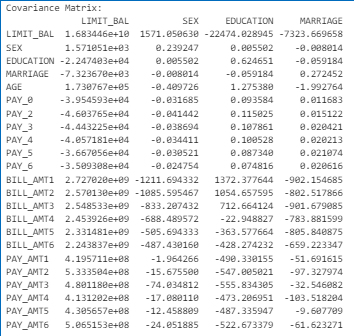
Table 1 Statistics of the data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Count** | **Mean** | **Standard Deviation** | **Minimum** | **Maximum** |
| LIMIT\_BAL | 30000 | 167484.3 | 129747.7 | 10000 | 1000000 |
| SEX | 30000 | 1.6 | 0.5 | 1 | 2 |
| EDUCATION | 30000 | 1.8 | 0.5 | 1 | 4 |
| MARRIAGE | 30000 | 1.5 | 0.5 | 1 | 3 |
| AGE | 30000 | 35.5 | 9.2 | 21 | 79 |

## 2.3 Overall Data Relationships

### 2.3.1 Covariance

Covariance measures the linear relationship between two variables [1]. High covariance between features may suggest redundancy. The covariance matrix indicates generally high values for LIMIT\_BAL. However, this doesn't necessarily imply a meaningful relationship. Analyzing the original dataset reveals LIMIT\_BAL's high covariance may be due to large values rather than a strong correlation with other variables. This indicates a scale difference, not a linear relationship. The associated code is shown in Figure 3.



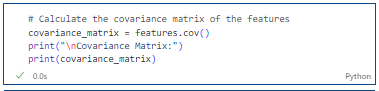


Figure 3. Code of data covariance and Example of the printed result

### 2.3.2 Correlation Heatmap

Correlation coefficients are preferred as they measure both linearity and variable scale. The range is -1 to 1, where values near -1 or 1 indicate strong linearity, while values near 0 suggest weak linearity. From the last row of the heatmap, the correlation between features and the target can be observed. Excluding the correlation of 1 with itself, the strongest correlation is observed between Pay0-Pay6, indicating a strong relationship between the repayment records of the past six months and whether the user will default in the next month. Moreover, education and age may be two variables worth paying attention to. The associated code and results are shown in Figure 4.

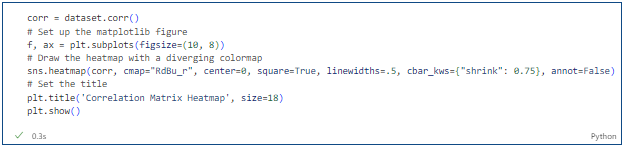
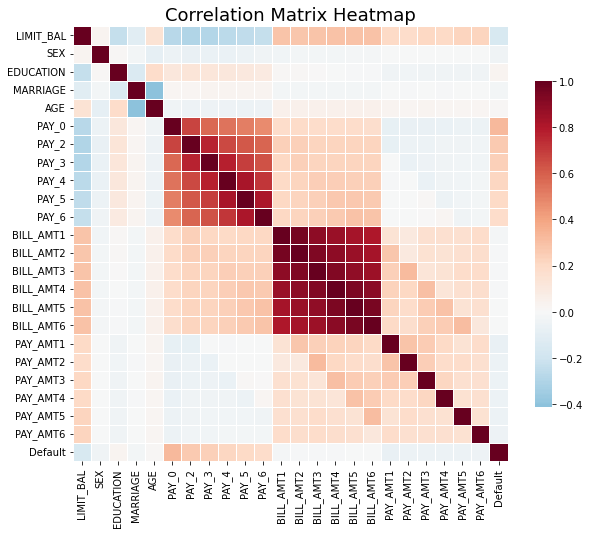
 

Figure 4. Code and result of the heatmap

Outliers, impacting certain models like linear regression, have minimal effect on decision tree and random forest models. Relationship plots show that "outliers" make up a negligible proportion, almost inconsequential to influencing data results, and the data distribution aligns with a normal distribution.

Multicollinearity in regression occurs when some predictors are highly correlated, leading to unstable regression coefficients. The absence of collinearity among features, observed in relationship plots, is a positive finding. The associated results are shown in Figure 5.

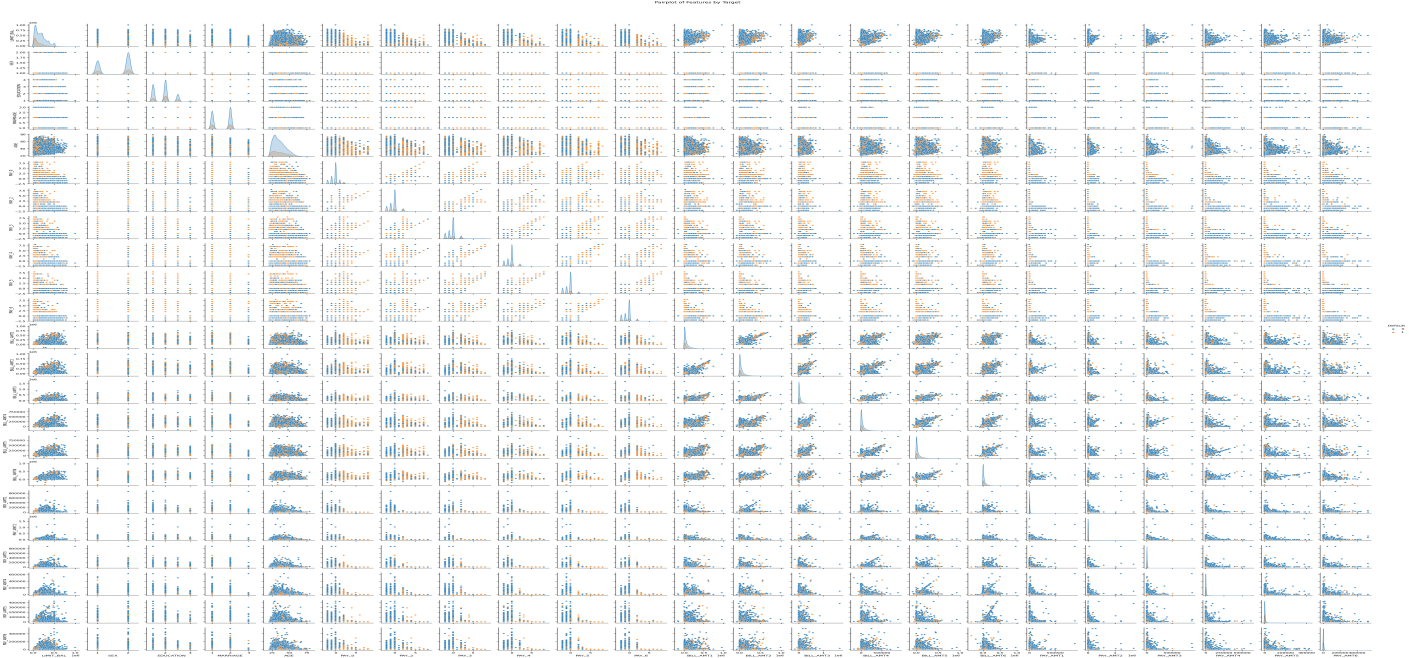
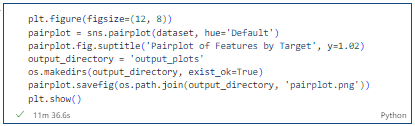


Figure 5. Associated Results

### 2.3.3 Histogram

The histogram reveals a much larger quantity of default payment data compared to non-default payment data. Both default and non-default payment data distributions appear to approximately follow a normal distribution. The associated code and results are shown in Figure 6.

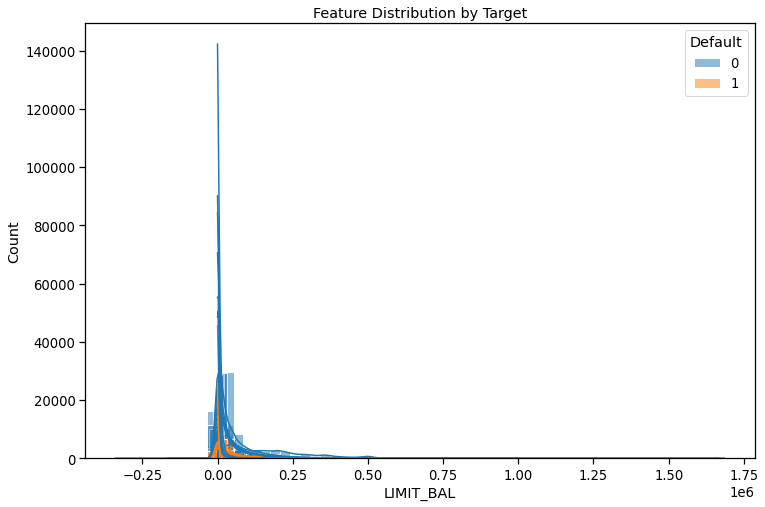
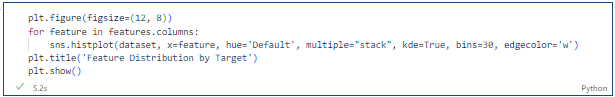


Figure 6. Code of histogram

The graph highlights dataset imbalance, with non-default instances outnumbering default cases by threefold. This dataset's inherent structure acknowledges the rarity of credit card defaults. Resampling risks losing vital minority class details. Artificial samples lack real-world diversity, impacting model authenticity. In imbalanced scenarios, precise metric choices like precision, recall, F1, AUC, and ROC are critical for a nuanced evaluation beyond accuracy. The associated code and results are shown in Figure 7.

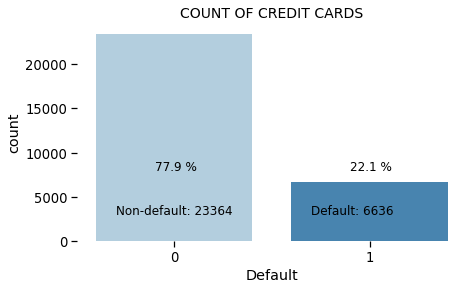
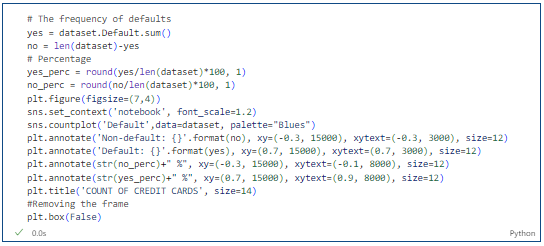


Figure7. Code and result of counts about labels

# 3. Feature Engineering

## 3.1 One-Hot Encoding

Within the dataset, one-hot encoding has been employed on Gender (1 = male; 2 = female), Education (1 = graduate school; 2 = university; 3 = high school; 4 = others), and Marital status (1 = married; 2 = single; 3 = others) features into binary vectors, making them more suitable for machine learning models. Take Gender as an example, where 1 denotes male and 2 denotes female, encoding will result in two columns: SEX\_1 and SEX\_2. Additionally, this process has resulted in the generation of new features. The associated code and results are shown in Figure 8.

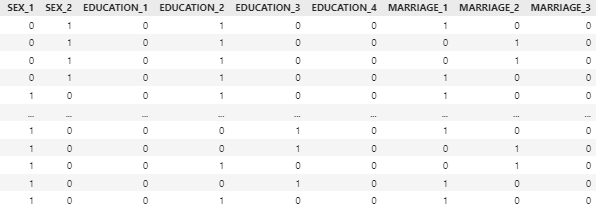
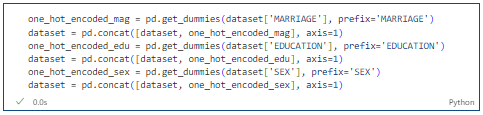


Figure8. Code and result of one-hot encoding

## 3.2 SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) is a resampling methodology of paramount importance in situations such as credit card default prediction, where instances of default are conspicuously less frequent compared to non-default instances [2]. This approach serves to augment the presence of the minority class, thereby cultivating a dataset characterized by improved balance. Within the scope of this project, SMOTE is implemented in the training set to address class imbalance. The corresponding code implementation and outcomes are elucidated in Figure 9.

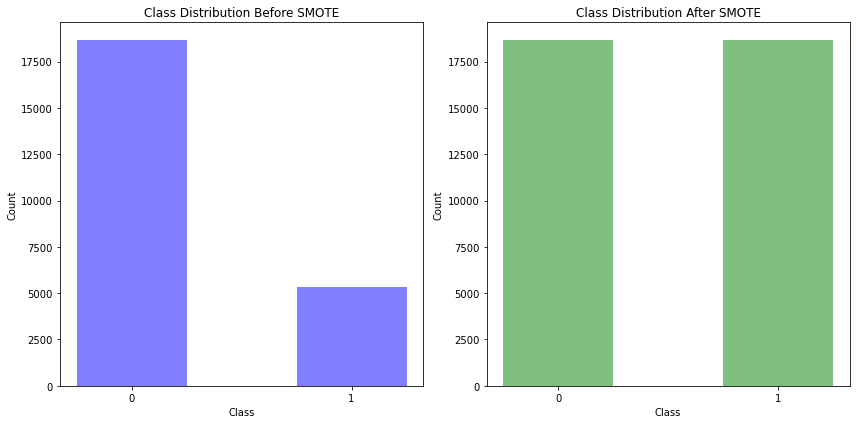


Figure9. Code and result of dataset distribution before and after SMOTE

## 3.3 Data Scaling

In machine learning, certain algorithms exhibit sensitivity to feature scaling, necessitating uniform scaling to prevent disproportionate feature impacts. Regression models, support vector machines (SVM), and neural networks are typically sensitive to feature scale, making scaling beneficial for enhanced model convergence and overall performance. In contrast, tree models like decision trees and random forests are unaffected by feature scale during the splitting process. These models construct decision rules via recursive data splitting, where thresholds depend on feature value comparisons. The associated code is shown in Figure 10.

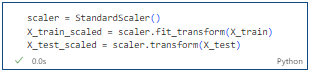


Figure10. Code of scaling

# 4. Model Training

The dataset is split into a test set (25%) and a training set using a random seed of 33. Classification tasks include four models which are Logistic Regression, SVM, a Random Forest model, and Gradient boosting. regression tasks include four models which are Linear Regression, Random Forest, an ANN model, and gradient boosting.

## 4.1 Classification model

**Logistic Regression** uses the sigmoid function to map linear predictions between 0 and 1, making it suitable for probability estimation. **SVM** finds the decision boundary that maximizes the margin, particularly effective in handling high-dimensional data. **Random Forest** mitigates overfitting by ensembling multiple decision trees, especially beneficial in discrete target variable classification tasks, such as task 1, which involves binary classification. **Gradient Boosting** iteratively trains models, focusing on previous model errors, progressively enhancing predictive performance, especially adept at handling nonlinear relationships and outliers.

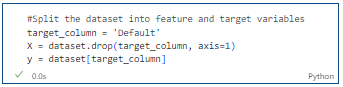


Figure11. Training and Testing dataset preparation

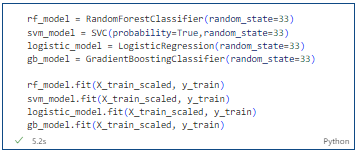
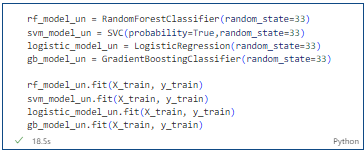


Figure12. Constructing Machine Learning Models and Model Training: Unscaled (Left) vs. Scaled (Right)

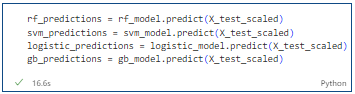
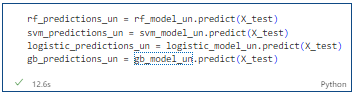


Figure13. Predicting the Test Set: Unscaled (Left) vs. Scaled (Right)

## 4.2 Regression model

**Linear Regression** models the relationship between dependent and independent variables with a linear equation, suitable for predicting continuous outcomes. **Random Forest** predicts continuous output values in a regression model. **Artificial Neural Network (ANN)** models, inspired by human brain structure, use interconnected nodes to learn complex relationships, making them powerful for regression tasks with nonlinear patterns. **Gradient Boosting**, the output is a prediction of continuous values.

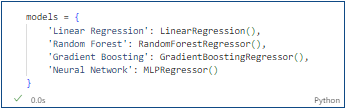


Figure14. Create a model dictionary applying sci-kit-learn

## 4.3 Cross-validation

Cross-validation is a versatile evaluation method that comprehensively enhances model reliability and performance, especially in the dataset with a small dataset or class imbalance. It mitigates overfitting risks by iteratively validating different data subsets, ensuring a robust assessment of generalization performance. By utilizing 10-fold cross-validation, a more robust performance evaluation can be achieved, reducing the randomness introduced by a single dataset split.

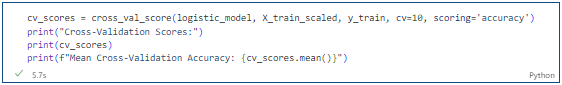


Figure15. Sample code of 10-fold cross-validation

## 4.4. Model Optimization

Finding the optimal combination of model performance by trying different sets of hyperparameters. Utilizing **Grid Search** (GridSearchCV) and **Bayesian Optimization** for hyperparameter tuning in the models.

Taking random forest as an example, in this project, parameters such as n\_estimators and max\_depth were fine-tuned for the random forest (RF) model, employing grid search (GridSearchCV) and Bayesian Optimization. The associated hyperparameter optimization code is shown in Figure 16.

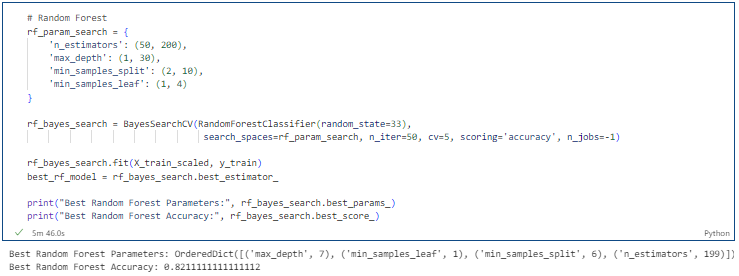


Figure 16. Example code about hyperparameter optimization: Grid Search (Left) vs. Bayesian Optimization (Right)

## 4.5 Model evaluation

### 4.5.1 Classification task evaluation index

In classification tasks, common evaluation metrics include **Accuracy**, **Precision**, **Recall**, **F1-score**, and **AUC**. **Accuracy** measures the proportion of correctly predicted samples out of the total samples. **Precision** evaluates the proportion of true positive predictions among the samples predicted as positive. **Recall** measures the proportion of true positive predictions among the actual positive samples. **F1-score** is the harmonic mean of Precision and Recall, providing a balanced assessment of model accuracy and completeness. **AUC** stands for Area Under the ROC Curve. The ROC curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) across different thresholds for a model's predicted probabilities.

### 4.5.2 Regression task evaluation index

In regression tasks, common evaluation metrics include , , and **Explained Variance**. **RMSE** measures the average magnitude of the residuals (the differences between predicted and actual values) and is sensitive to larger errors. **MAE** measures the average absolute difference between the predicted and actual values. It is less sensitive to outliers compared to RMSE. represents the proportion of variance in the dependent variable (target) explained by the independent variables (features). A higher indicates a better fit, with 1 being a perfect fit. **Explained Variance** quantifies the proportion of variance in the target variable that the model can explain. A higher value indicates better explanatory power.

# 5. Results

Using the models mentioned in Sections 4.1 and 4.2, along with techniques such as 10-fold cross-validation and Grid Search, Bayesian Optimization for model training, optimal classification and regression results were obtained for each model. The relevant outcomes are presented in this section.

## 5.1 Results of classification

This project uses four classification models for training and prediction on the test set data. The relevant code is shown in the figure 17.

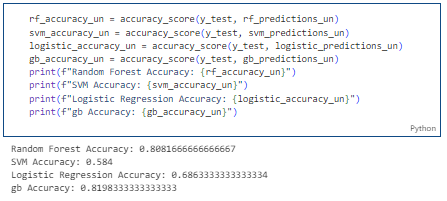


Figure 17. Code and result of classification model predictions (unscaled)

The performance of predictions obtained using the unscaled dataset is shown in Table 2.

Table 2 The performance of predictions using the unscaled dataset

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Random Forest | 0.8153 |
| SVM | 0.7823 |
| Logistic Regression | 0.7821 |
| Gradient Boosting Tree | 0.8198 |

Although the accuracy of the three models is decent when using the unscaled dataset, in reality, considering multiple performance metrics is essential due to the dataset's imbalance, not just accuracy. The F1 score is a metric that comprehensively considers both precision and recall, making it particularly suitable for situations with imbalanced classes.

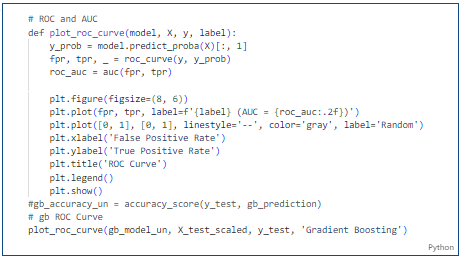
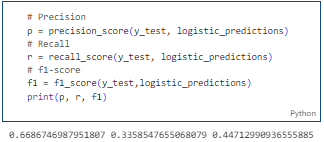


Figure 17. Sample code and result of calculating classification indexes

Finally, this work obtained the performance results of scaled and unscaled data on the four classifiers, as shown in Table 3.

Table 3. The performance of predictions of classification

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Unscaled\_data | | Scaled\_data | |  |
| F1score | AUC | F1score | AUC | Accuracy |
| SVM | 0.42 | 0.63 | 0.44 | **0.72** | **0.8153** |
| Logistic Regression | 0.38 | 0.58 | 0.35 | **0.72** | **0.7823** |
| Random Forest | 0.49 | 0.49 | 0.46 | **0.77** | **0.7821** |
| Gradient Boosting Tree | 0.50 | 0.72 | 0.50 | **0.78** | **0.8198** |

After scaling, the performance of the three models has improved to varying degrees, with Random Forest showing the smallest change, supporting the notion that tree models do not require scaling. Both SVM and Logistic Regression have shown significant improvements in predicting class 1 (default). The Random Forest model has achieved the best classification effect, and its F1 value can be as high as 46%.

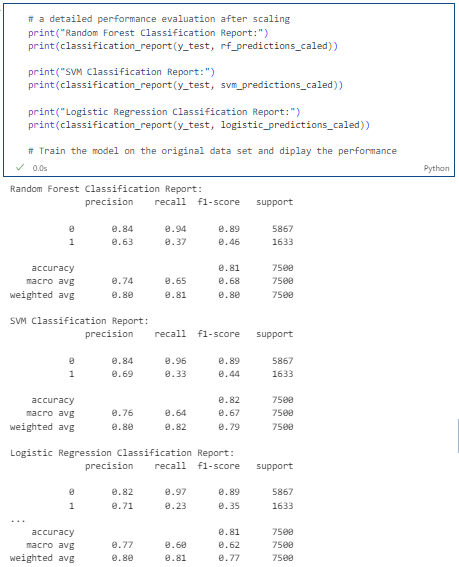
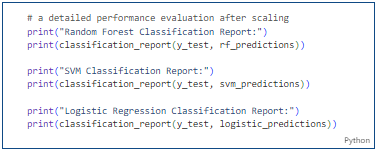
The overall F1 score is relatively low, but across various models, the F1 score for the class representing normal credit card repayments (Value = 0) consistently exceeds 88%. However, for the class representing credit card defaults (Value = 1), the F1 score is only around 50%, significantly diminishing the overall F1 score result.

Figure 18. Code and Sample result of F1-score for default and not default classes

The classification metrics corresponding to different categories are presented in Table 4.

Table 4 F1-score for default and not default classes

|  |  |  |
| --- | --- | --- |
|  | Class | F1score |
| SVM | 0 | 0.89 |
| 1 | 0.46 |
| Logistic Regression | 0 | 0.89 |
| 1 | 0.44 |
| Random Forest | 0 | 0.89 |
| 1 | 0.35 |

The primary reason for this phenomenon is the imbalance in the categorical data, even though the SMOTE method was employed for data resampling, it did not sufficiently address the issue. Another contributing factor is the absence of hyperparameter tuning during the model training process, resulting in suboptimal training outcomes. To ameliorate this, methods such as Bayesian optimization were utilized for hyperparameter optimization. Taking the random forest model as an example, the optimal hyperparameters obtained following the parameter adjustment are presented in Table 5. The optimized accuracy after this adjustment is 0.82.

Table 5 The optimal hyperparameters

|  |  |
| --- | --- |
| **Parameter** | **Values** |
| max\_depth | 7 |
| min\_samples\_leaf | 1 |
| min\_samples\_split | 6 |
| n\_estimators | 199 |

Additionally, model performance is assessed through ROC curves. A high AUC-ROC value typically indicates that the model can effectively distinguish between positive and negative samples. Gradient Boosting has the best performance. ROC curves of each model are shown in Figure 11.

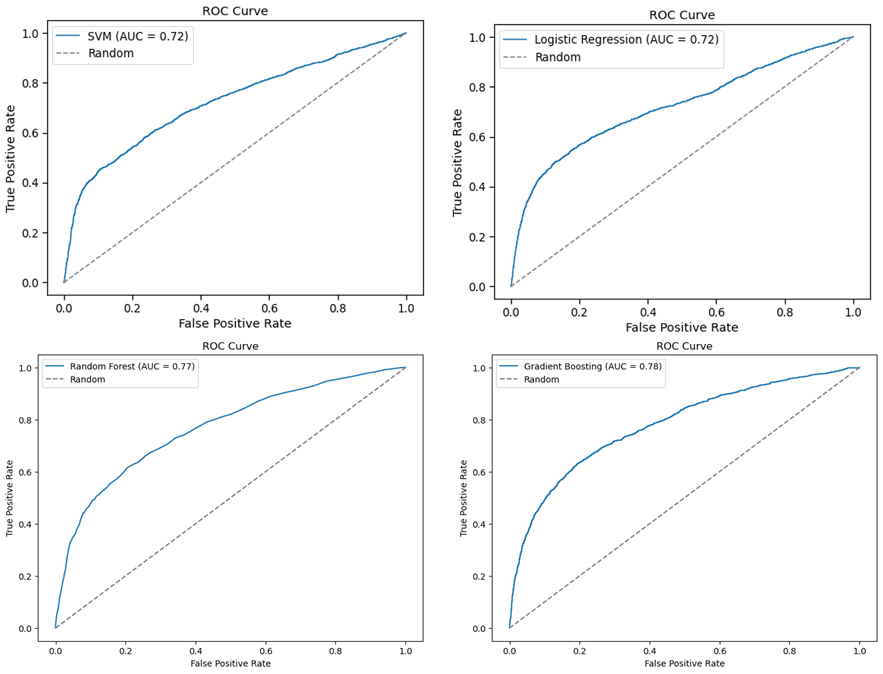


Figure 11. Results of ROC curves

From the AUC results, it can be observed that the performance of the best model reaches up to 78%. The classification outcomes of the model are considered acceptable based on this metric.

## 5.2 Results of regression

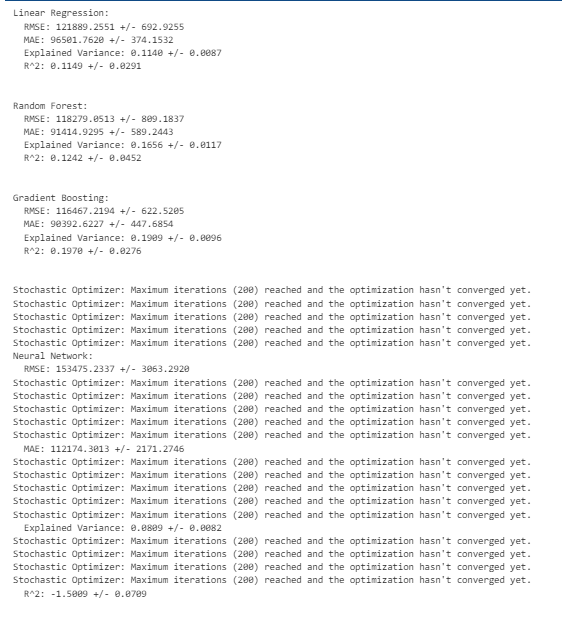
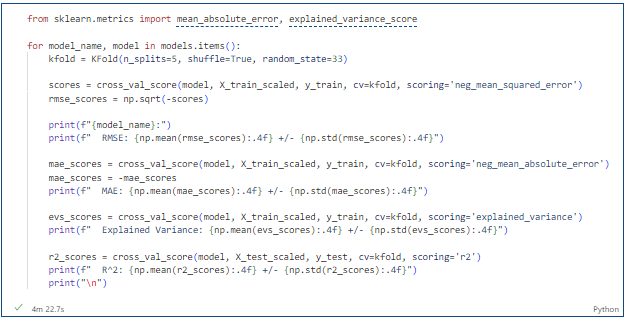


Figure 17. Code and sample result of calculating regression indexes

In Task 2, four models were trained and evaluated using four regression metrics. Additionally, grid search was employed in this project to obtain relatively optimal model parameters and their corresponding training results. Table 5 displays the outcomes of the four regression models.

Table 5 The results of the regression

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **MAE** | | **RMSE** |  | **Explained Variance** |
| Linear Regression | | 96501.7 | 121889.3 | 0.1149 | 0.1140 |
| Random Forest | | 91414.9 | 116467.2 | 0.1970 | 0.1656 |
| Gradient Boosting | | **90392.6** | **116467.2** | **0.1970** | **0.1909** |
| Neural Network | | 112174.3 | 153475.2 | -1.5009 | 0.0809 |

From the regression results, it is evident that the best-performing model is the Gradient Boosting Tree, followed by the Random Forest. Both ensemble learning models exhibit similar RMSE and metrics, but the Gradient Boosting (GB) model demonstrates a slightly lower MAE, indicating a slightly superior performance compared to the Random Forest.

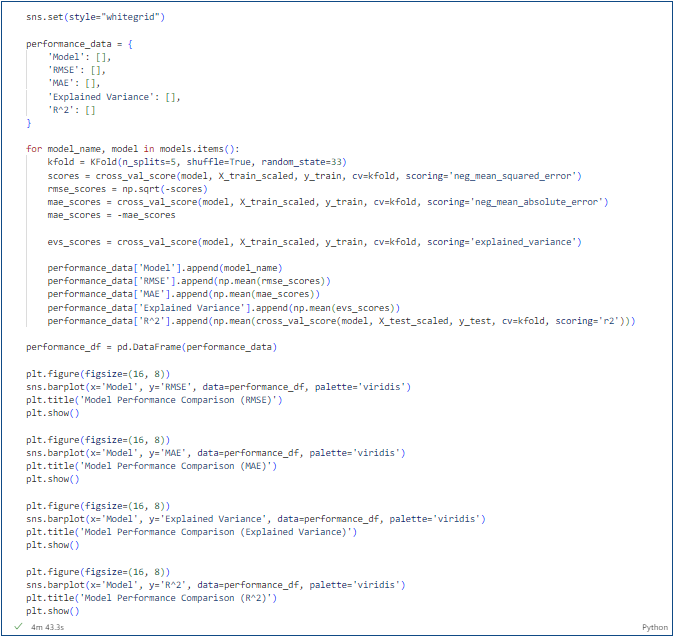


Figure 18. Code of visualizing the performance indexes

The histogram of the four regression models is shown in the Figure 19.

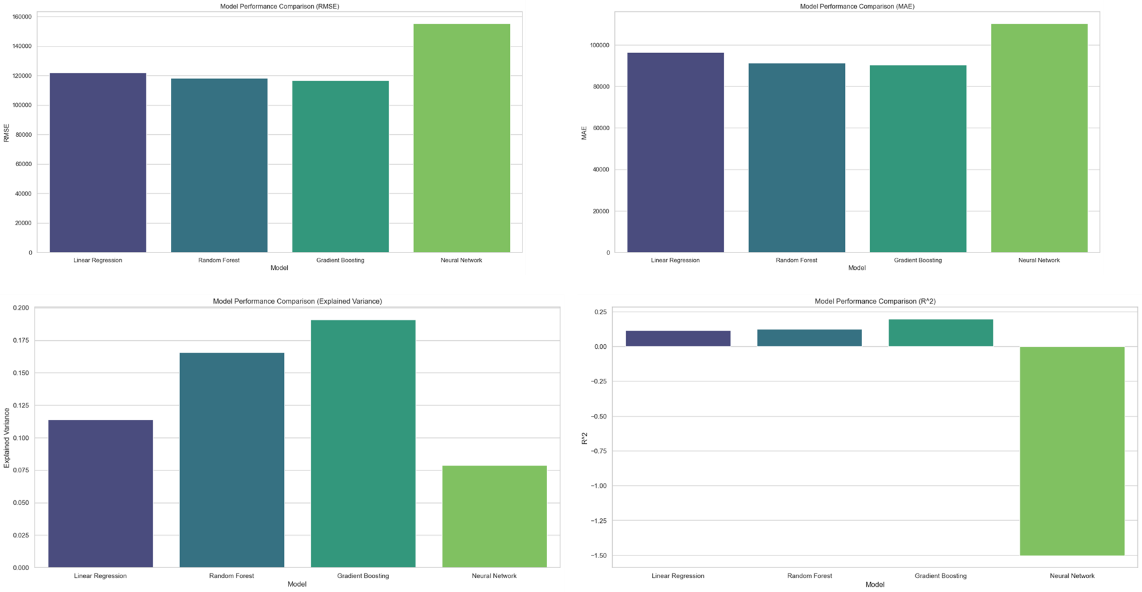


Figure 12. Histogram of the results

From the regression results, the ensemble learning models exhibit a hierarchy in performance, with Gradient Boosting (GB) outperforming Random Forest (RF), and Neural Network (NN) displaying the lowest performance. Given that both GB and RF are ensemble learning models, their overall performance tends to surpass that of non-ensemble learning models. The training process of neural networks necessitates substantial parameter tuning, and their performance on structured data might not be as competitive as that of ensemble learning models. Hence, the observed results align with certain realistic expectations.

## 5.3 Interpretability of the results

The fundamental principle of SHAP values is to assign a contribution value to each feature, indicating the degree of impact of that feature on the model's output [4]. SHAP values take into account every possible combination of features and calculate the average contribution of each feature to the model's output. This averaging property ensures fairness, preventing undue weighting of a specific combination.

### 5.4.1 SHAP values of classification

The code for interpretable classification using SHAP is depicted in Figure 13.

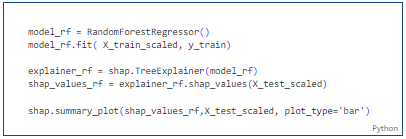


Figure 13. The code of SHAP

The feature importance ranking is shown in Figure 14.

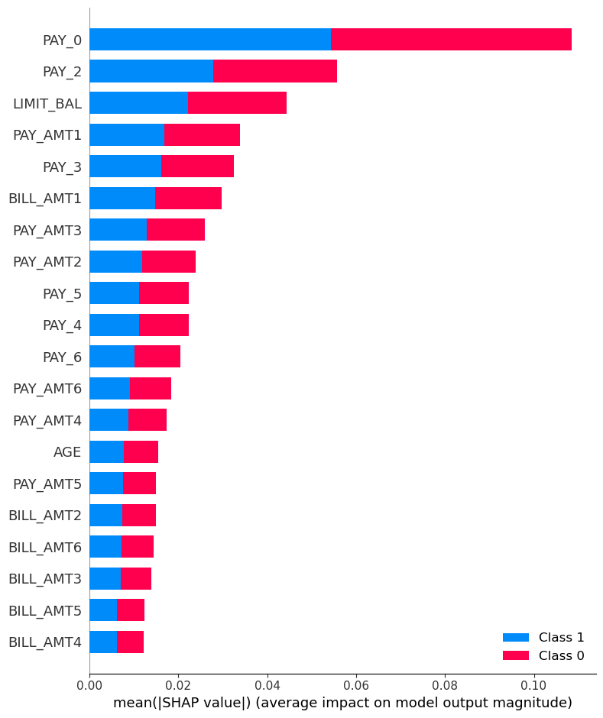


Figure 14. The feature importance of SHAP

Figure 14 shows the average SHAP value for each feature. The higher the average SHAP value, the more stable the influence of the feature on the prediction result. PAY\_0 features are the most important for classification results, followed by PAY-2, LIMIT\_BAL, etc. Their average SHAP values are high, indicating that they are important factors in the model prediction results. The observed SHAP values in the classification models align with real-world financial transaction scenarios. In Chapter 2, the project examined the heatmap of feature correlations from the perspective of data correlation, revealing a strong correlation between classification outcomes and features such as PAY\_0, PAY\_2, and LIMIT\_BAL. The SHAP values derived from classification models, particularly Random Forest, effectively highlight feature importance. These results are consistent with the characteristics observed in real financial scenarios, providing valuable insights and serving as a meaningful reference.

The SHAP value of a single sample is shown in Figure 15.

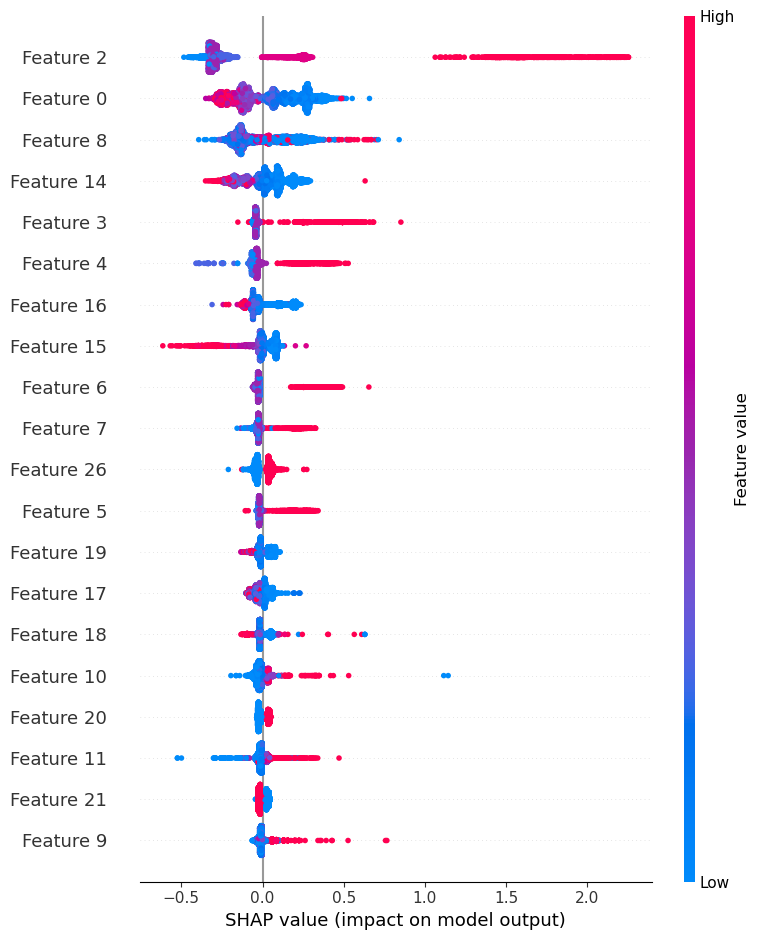


Figure 15. The feature importance of SHAP value about an instance

Figure 15 shows that Feature2, that is, the classification result is the most sensitive to feature-2, that is, pay\_0. Next is Limit\_Bal. Similar to the overall feature importance depicted by SHAP values, for individual samples in classification models, features such as PAY\_0 and LIMIT\_BAL continue to contribute significantly to the model's predictions. Therefore, whether from a global or local perspective, based on machine learning models and data-driven methods, the work can assert that features like PAY\_0 and LIMIT\_BAL are crucial in the task of predicting whether a credit card will be borrowed.

### 5.4.2 SHAP values of regression

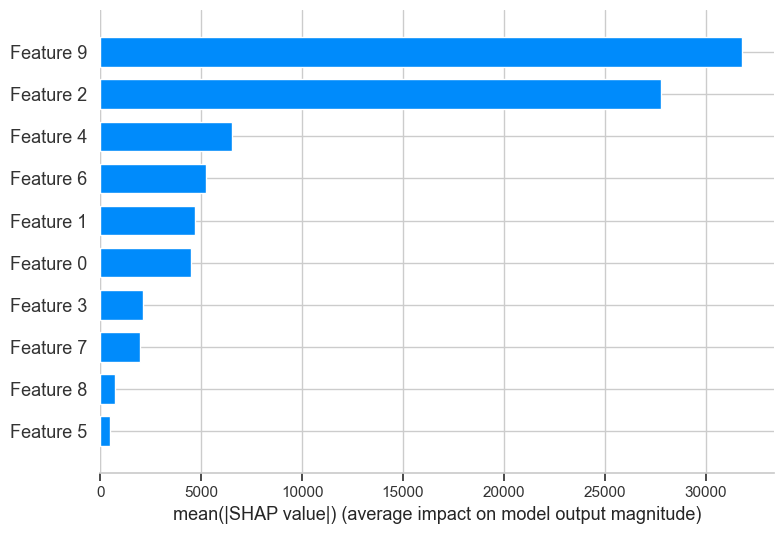


Figure 16. The feature importance of SHAP values

As illustrated in Figure 16, it can be seen that Feature 9, as “Age” is the most important feature, followed by Feature 2\_Master degree. From the overall SHAP values in the regression task, features such as Feature9, Feature2, and Feature4 exhibit significant contributions to the model's predictions. This aligns with real-world credit card limit-setting criteria. In the heatmap presented in Chapter 2, it was observed that Feature9, Feature2, and Feature4 have strong correlations with LIMIT\_BAL, further validating the data's correlation patterns. This consistency between the SHAP values and the observed data correlations reinforces the importance of these features in the context of credit limit determination.

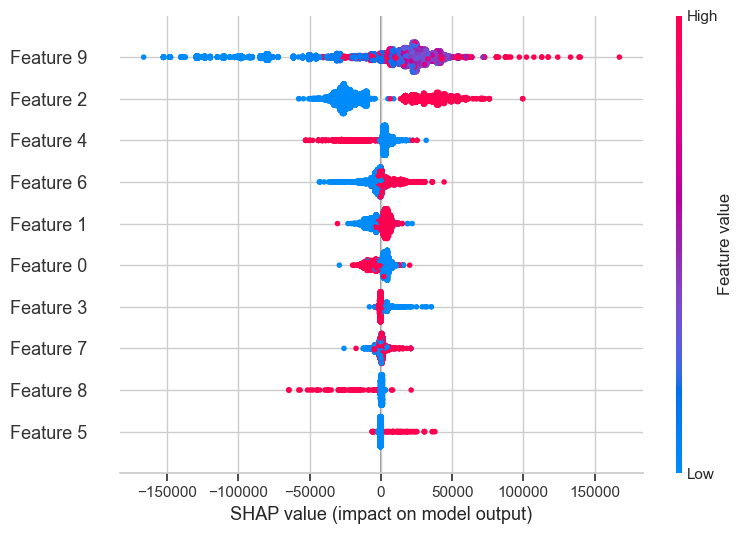


Figure 17. The feature importance of SHAP values about an instance

From the impact of the SHAP value of a single sample on the result, Feature 9\_Age, Feature 2\_Master degree, and Feature 4\_high school degree are the three most important features. Feature 4 is inversely proportional to the predicted result path. The examination of SHAP values for individual samples in the regression task indicates that features such as Feature 9, Feature 4, and Feature 2 have substantial impacts on the model predictions. This consistency with the overall feature importance in the regression model suggests that Feature 9, Feature 4, and Feature 2 indeed have significant influences on the regression outcomes, aligning with real-world scenarios.

# 6. Conclusion

The study is centered around predicting credit card payment default probability and limit balances using a comprehensive dataset that includes detailed credit card information. It encompasses both classification and regression tasks, covering default payments, demographic factors, credit data, and the payment and billing history of credit card customers in Taiwan from April 2005 to September 2005.

The significance of this research lies in its ability to construct predictive models for credit card defaults and credit limits [5]. The employed methodology has been proven feasible and valid, substantiated by performance indicators such as F1 scores. The robustness of the approach is further ensured through meticulous parameter adjustment and the comparative analysis of proportional results. Various machine learning models are distinctly applied for classification and regression tasks, incorporating diverse data preprocessing techniques such as one-hot encoding, and SMOTE, and utilizing methodologies like 10-fold cross-validation and hyperparameter optimization. The culmination of these processes yields highly effective classification and regression models, the outcomes of which find practical application in real-world scenarios.

In addition to model construction, the study delves into the interpretability of machine learning models [6], specifically leveraging SHAP values for model interpretation. Visualizing the impact of various characteristics highlights the significance of age, education, and PAY\_0, aligning with practical scenarios where credit card loans or limits are determined.

While models have exhibited reasonable effectiveness in the classification task, there exists potential for improvement. Consequently, future endeavors may explore the incorporation of deep learning methods to elevate the performance of models.

# References

[1] Subasi A, Cankurt S. Prediction of default payment of credit card clients using Data Mining Techniques[C]//2019 International Engineering Conference (IEC). IEEE, 2019: 115-120.

[2] Sakaguchi H, Stewart N, Gathergood J, et al. Default effects of credit card minimum payments[J]. Journal of Marketing Research, 2022, 59(4): 775-796.

[3] Kelly B, Xiu D. Financial machine learning[J]. Foundations and Trends® in Finance, 2023, 13(3-4): 205-363.

[4] Man X, Chan E P. The best way to select features? comparing mda, lime, and shap[J]. The Journal of Financial Data Science, 2021, 3(1): 127-139.

[5] Ganaie M A, Hu M, Malik A K, et al. Ensemble deep learning: A review[J]. Engineering Applications of Artificial Intelligence, 2022, 115: 105151.

[6] Lin K, Gao Y. Model interpretability of financial fraud detection by group SHAP[J]. Expert Systems with Applications, 2022, 210: 118354.