Team Project Final Report

Section A, Team 13

Team Members: Samiha Jain (sj472), Alaina Hu (ayh16), Weigao Tan (wt99), Shifei Ruan (sr678), Shicheng Li (sl975), Ricky Han (rh396)

Business Problem & Decision (Business & Data Understanding)

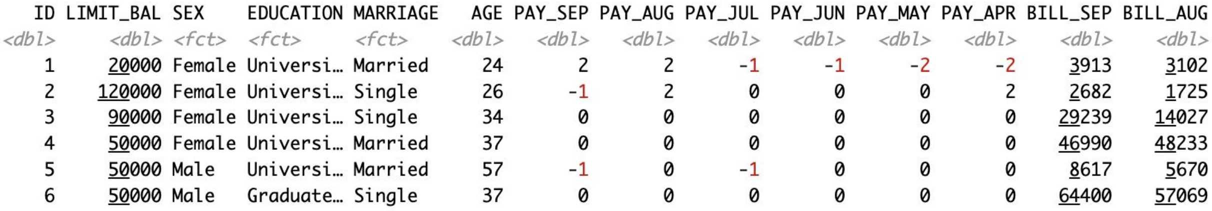
Credit card companies often face the challenge of predicting whether customers will default on their payments. If defaults are not identified in advance, the business may suffer financial losses, higher collection costs, and long-term damage to profitability. At the same time, being overly cautious in granting credit could reduce customer satisfaction and limit growth opportunities. Managing this trade-off between risk control and business development is therefore a critical issue. So our exact business problem is that the credit card company does not know in advance which customers are likely to default on their payments, and as a result, it faces the risk of significant financial losses and operational inefficiencies.

Therefore, we leverage the machine learning project to help the credit card company identify most important factors relative to client reputation and forecast individual default payment probability based on the feature variables. In this way, the risk assessment and business growth can be guaranteed in the investment banking industry.

Data Cleaning (Data Preparation)

The database we apply contains the information about 30,000 credit card clients in Taiwan, which represents their default payment rate based on their demographic characteristics, including credit limit, education level, marital status, previous payment and so on, so the industry belongs to **Consumer Banking**. When conducting data cleaning, we first rename the PAY columns based on the time of the repayment status (e.g. ‘PAY_0’ to ‘PAY_SEP’). Then we filter the other and unknown values in education levels (5,6) and marriage status (3). After filtering, 29163 records left. And we check each variables’ datatype to convert *EDUCATION, MARRIAGE, and default_payment* from numerical variables into factors with readable labels. Finally, we convert the cleaned data into a csv file for further processing and modelling.

After the data cleaning, the first five rows of our database look like this:



```
   ID LIMIT_BAL SEX    EDUCATION MARRIAGE   AGE PAY_SEP PAY_AUG PAY_JUL PAY_JUN PAY_MAY PAY_APR BILL_SEP BILL_AUG
<dbl>     <dbl> <fct>  <fct>     <fct>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>    <dbl>    <dbl>
    1     20000 Female Universi… Married     24       2       2      -1      -1      -2      -2     3913     3102
    2    120000 Female Universi… Single      26      -1       2       0       0       0       2     2682     1725
    3     90000 Female Universi… Single      34       0       0       0       0       0       0    29239    14027
    4     50000 Female Universi… Married     37       0       0       0       0       0       0    46990    48233
    5     50000 Male   Universi… Married     57      -1       0      -1       0       0       0     8617     5670
    6     50000 Male   Graduate… Single      37       0       0       0       0       0       0    64400    57069
```

Issues on Data Cleaning

1. Repayment Status (PAY_0 … PAY_6)

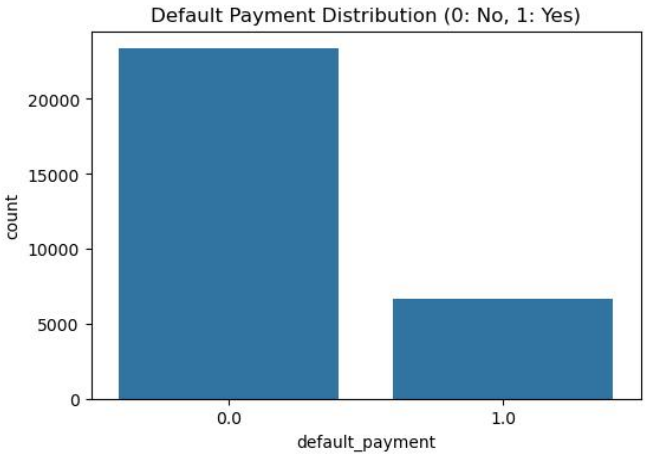
The repayment status variables are designed to take values from -1 to 9, where -1 indicates on-time repayment and higher values indicate the number of months delayed. In the dataset, however, we identified an additional undocumented value of -2. This anomaly appears in 6,561 records, which accounts for roughly 22–23% of the total observations. Because of its relatively large proportion, this irregularity raises concerns: simply removing these records would significantly reduce the dataset size, but keeping them may distort repayment behavior analysis.

2. Bill Statement Amounts (BILL_AMT1 … BILL_AMT6)

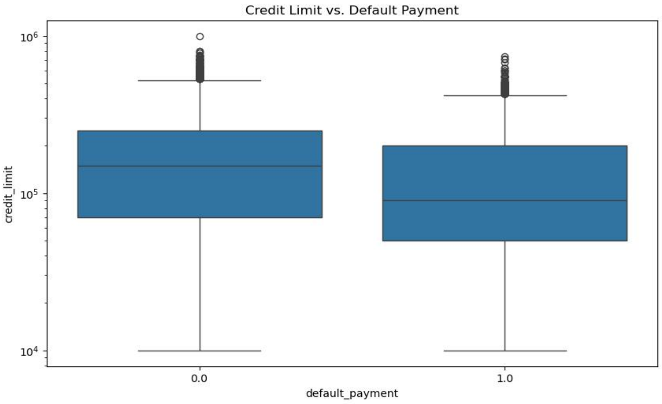
The bill statement variables represent the billed amount for each of the past six months. In principle, these values should be non-negative. However, we identified 1,930 records (approximately 6.4% of the dataset) that contain negative values in at least one of the bill amount variables. These negative values may reflect over-payments or surplus depositsby customers, but they remain inconsistent with the expected definition of billing amounts. This anomaly requires further consideration in the data cleaning process, as it could potentially distort the distribution of financial features.

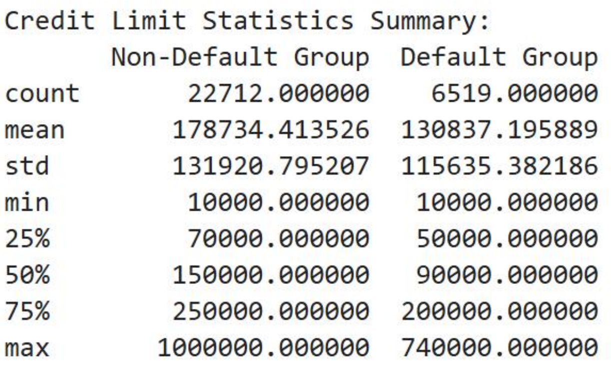
Data Analyzing (Data Preparation)

During the data analyzing process, we first focus on the target variable: default_payment (0: No 1: Yes), the core feature directly reflecting whether the client's payment is overdue or not. The result shows that the average default rate among our 30,000 clients is 22.30%. In other words, the ration between punctual payment and default payment is approximately 4:1.



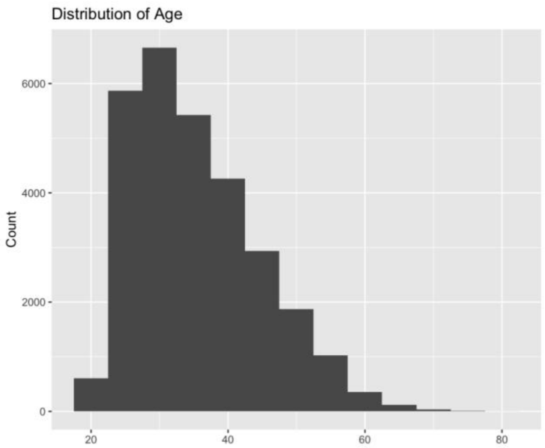
Next, we explore the relationship between credit limit and default payment. We calculate the detailed statistics for default and non-default group. The statistics show that the credit limit for non-default group is higher than the default group on all levels. Here's the data distribution and boxplot for credit limit between non-default and default group：

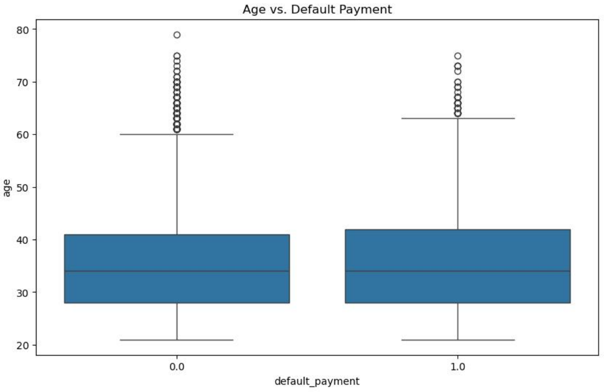


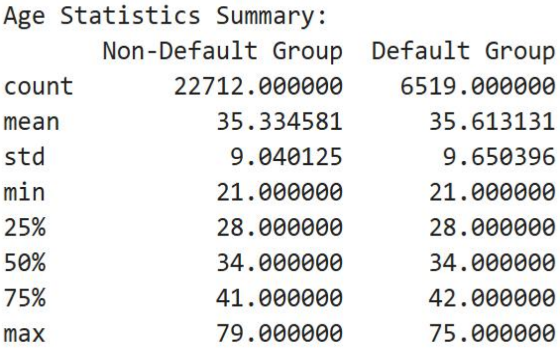


```
Credit Limit Statistics Summary:
       Non-Default Group  Default Group
count       22712.000000    6519.000000
mean       178734.413526  130837.195889
std        131920.795207  115635.382186
min         10000.000000   10000.000000
25%         70000.000000   50000.000000
50%        150000.000000   90000.000000
75%        250000.000000  200000.000000
max       1000000.000000  740000.000000
```

Then, we look at the relationship between age and default payment. The age distribution is right-skewed, with most customers clustered around their late-20s to mid-40s (peak in the early-30s) and a long thin tail into older ages. Very few observations fall below 20 or above 70. However the distribution of clients' age between the non-default and default group is approximately the same centering around 35 years old with standard deviation of 9. The only difference is that the maximum value of age for non-default group is 79, which is 4 years older than the default one.

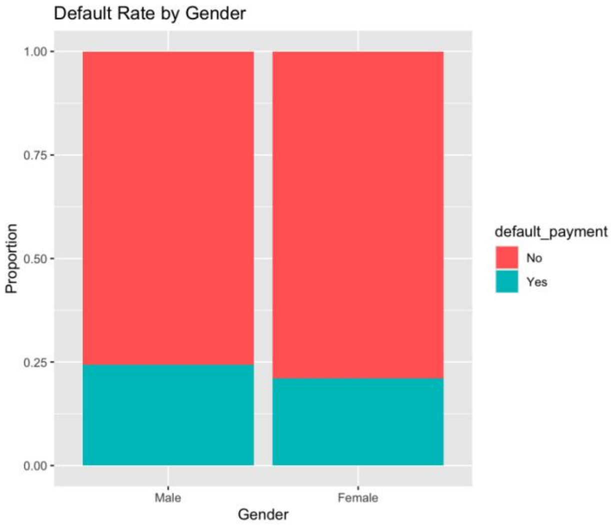




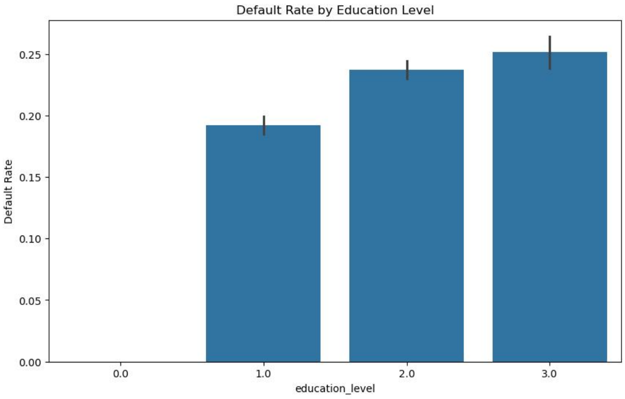


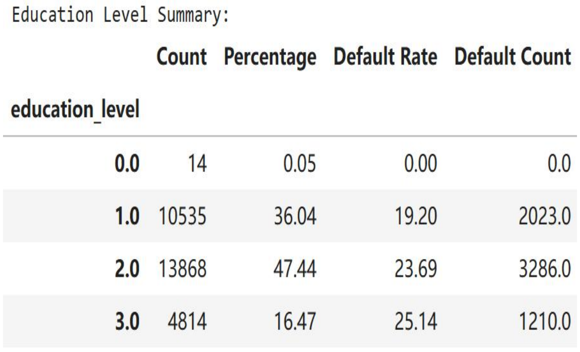
```
Age Statistics Summary:
       Non-Default Group  Default Group
count       22712.000000    6519.000000
mean           35.334581      35.613131
std             9.040125       9.650396
min            21.000000      21.000000
25%            28.000000      28.000000
50%            34.000000      34.000000
75%            41.000000      42.000000
max            79.000000      75.000000
```

We also plot a graph to explore the gender distribution based on the response variable 'default_payment'. Both genders are predominantly non-defaulters, but males exhibit a slightly higher default proportion than females; the gap is modest.



Meanwhile, we segment the clients based on different education levels. The default rate and corresponding percentage for each education level are shown below, among which we can see that the default rate is highest for the high-school population which reaches 25.14%. And the default rate for university and graduate school group is 23.69% and 19.20% respectively. Since the sample size for value 4,5 & 6 is all too small, we later delete those subjects.

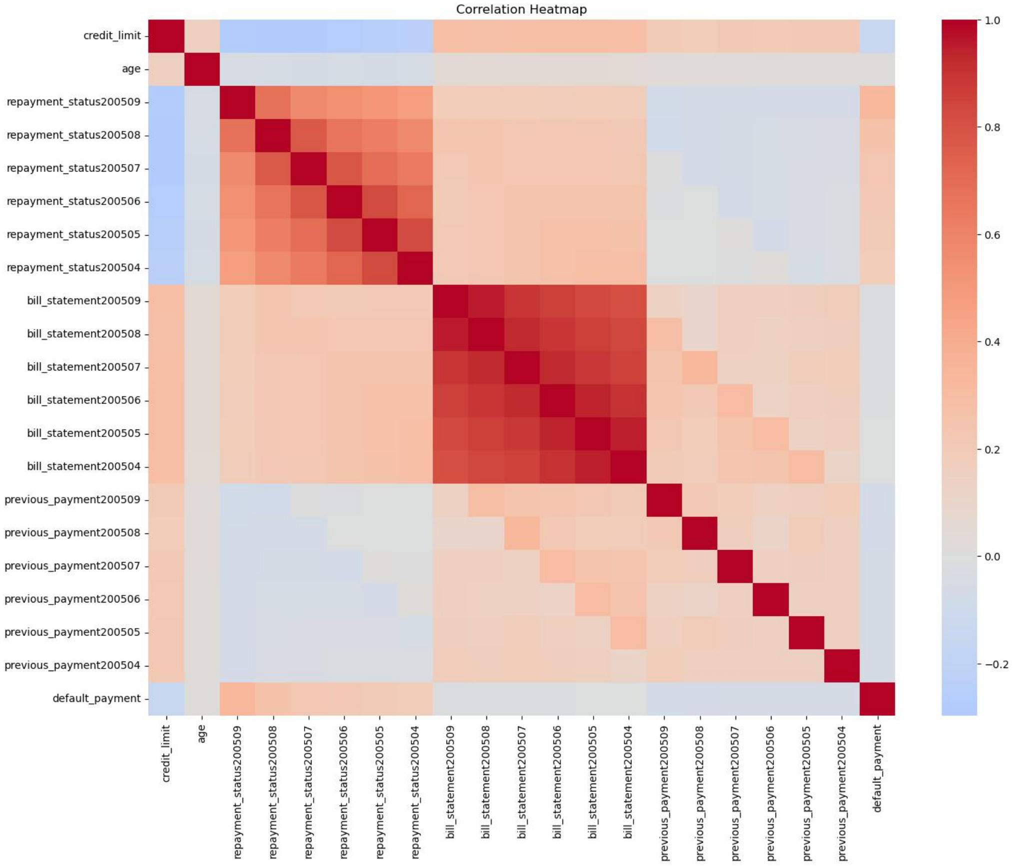




Education Level Summary:

education_level	Count	Percentage	Default Rate	Default Count
0.0	14	0.05	0.00	0.0
1.0	10535	36.04	19.20	2023.0
2.0	13868	47.44	23.69	3286.0
3.0	4814	16.47	25.14	1210.0

Finally, we create a correlation heatmap between all feature variables and the target variable (default_payment). The default payment is only positively correlated with repayment status from April 2005 to September 2005 and client's age. Conversely, the correlation between default payment and bill statement, previous payment, and credit limit is all negative.



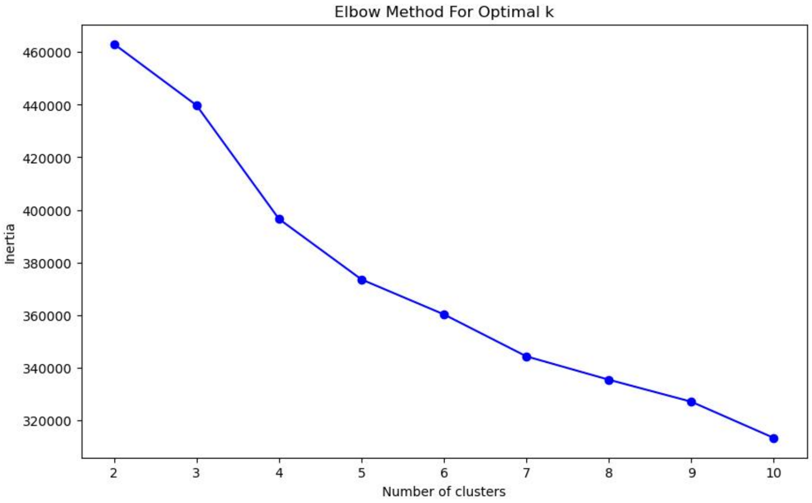
Data Modeling

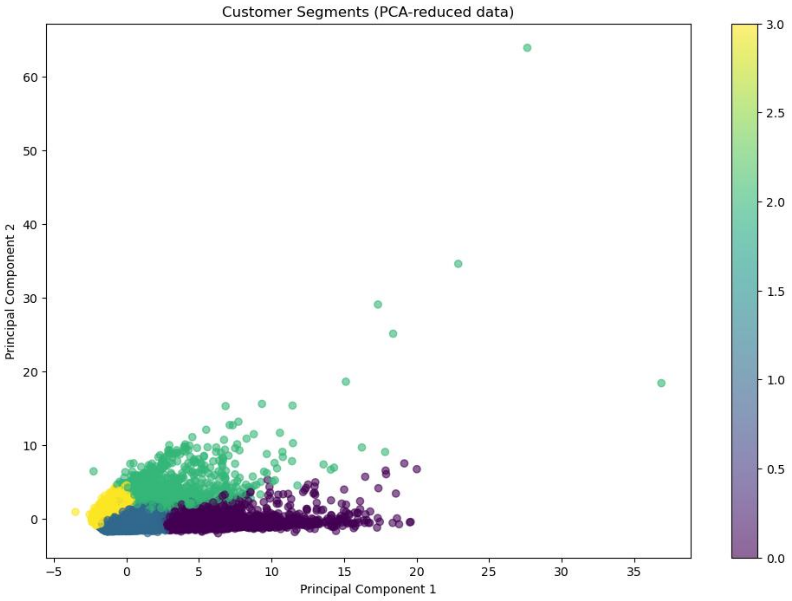
1. K-Means Clustering with PCA Visualization

We are now in the data modeling procedure. We first extract the target variable from the characteristic variables to split the original dataset into training and test group based on the ratio 8:2. Next, we build up processing pipeline transformer for numeric and categorical features independently and create two identical preprocessors - one for modeling and the other one for clustering, in order to apply all the preprocessing steps to the entire feature dataset. The final standardized, encoded version of data is suitable for performing clustering.

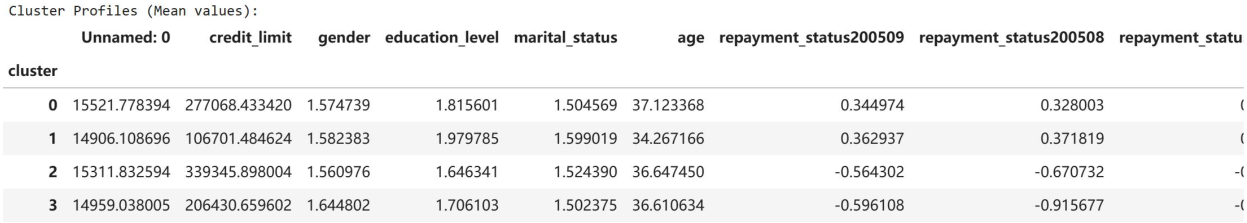
We then use Principle Component Analysis to reduce the high-dimension data to 2 dimensions. This allows us to visualize the complex data in a simple 2D plot while preserving the most important patterns. After that, we test different numbers (from 2 to 10) to find the optimal one for clustering. The line plot below represents how inertia changes with different cluster numbers: The “elbow” point (where the line bends) indicates the optimal number of clusters. Thus, we leverage K-means clustering with the chosen optimal number of clusters (4), assigning each client to one of the 4 clusters based on their feature similarity.

The scatter plot of the PCA-reduced data colors points according to their cluster assignment, showing how the different customer segments are distributed in 2D space.



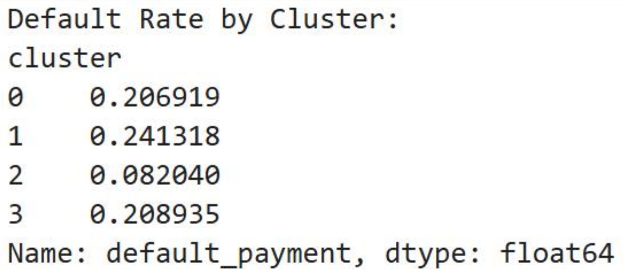


From the cluster profile and default rate bar chart, we know that the default rate is highest in Cluster 1 (24.13%), the group with the highest repayment status and lowest credit limit and previous payment on average. Meanwhile, the education level metrics of Cluster 1 approximately equals 2 (nearest to 3 (high-school diploma) among the four clusters). Oppositely, the cluster with lowest default rate - Cluster 2 (8.20%), possesses the significantly largest credit limit, bill statement and previous payment, as well as the repayment status all between (-1,0), which means paying on time. The clustering result above all perfectly align with our discoveries made in the data processing process.

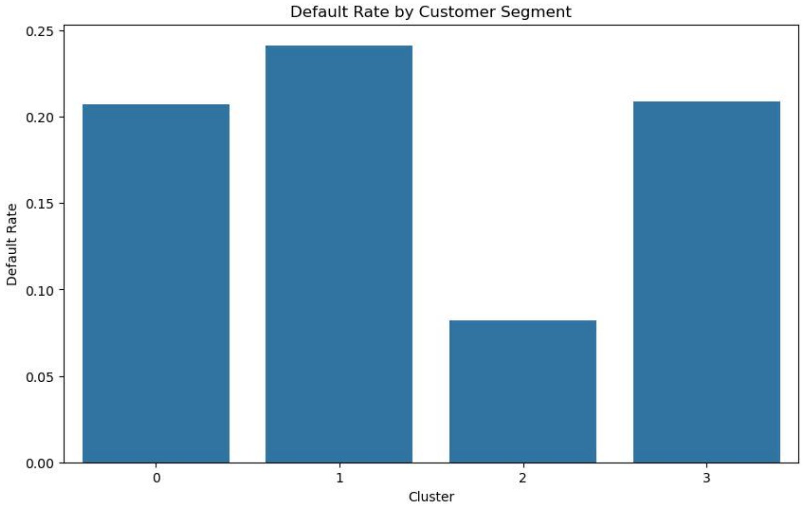


Cluster Profiles (Mean values):

	Unnamed: 0	credit_limit	gender	education_level	marital_status	age	repayment_status200509	repayment_status200508	repayment_statu
cluster									
0	15521.778394	277068.433420	1.574739	1.815601	1.504569	37.123368	0.344974	0.328003	(
1	14906.108696	106701.484624	1.582383	1.979785	1.599019	34.267166	0.362937	0.371819	(
2	15311.832594	339345.898004	1.560976	1.646341	1.524390	36.647450	-0.564302	-0.670732	-(
3	14959.038005	206430.659602	1.644802	1.706103	1.502375	36.610634	-0.596108	-0.915677	-(



```
Default Rate by Cluster:
cluster
0    0.206919
1    0.241318
2    0.082040
3    0.208935
Name: default_payment, dtype: float64
```

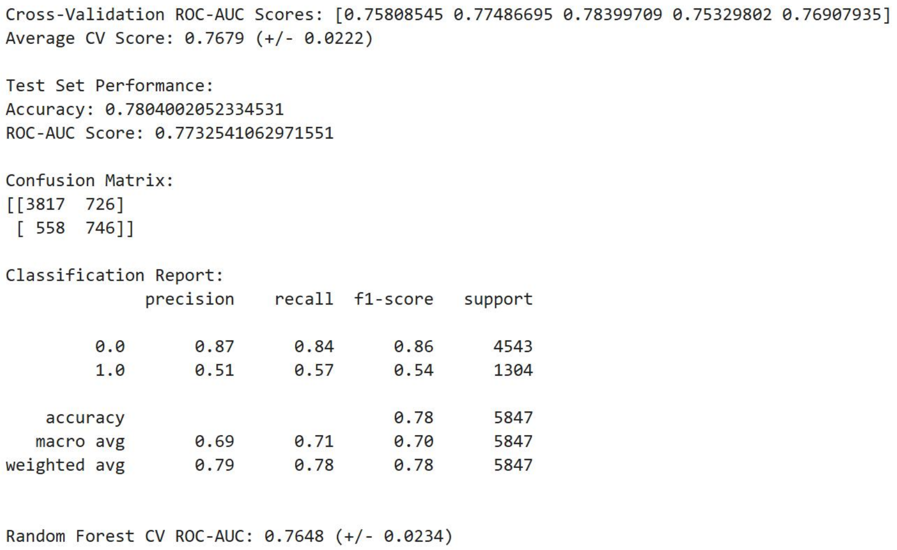


2. K-Fold Cross Validation

We apply 5-Fold Cross Validation to evaluate the function of model pipeline. The average Cross Validation ROC-AUC Score ± 2*Standard Error is between [0.7457,0.7901]. Our prediction model shows a 76.8% discrimination accuracy in internal validation, and remains stable on different subsets of data with fluctuations within 2.2%, indicating the reliability of the model's prediction results.

When evaluating the final model based on the test dataset, the confusion matrix displays the distribution of actual vs. predicted classification: 3817 True Negatives(actual non-default payment correctly predicted as non-default), 726 False Positives(actual non-default payment inaccurately predicted as default), 558 False Negatives(actual default payment inaccurately predicted as non-default) and 746 True Positives(actual default payment correctly predicted as default). Through the classification report, we can see that the predict accuracy for Classification 0 (non-default payment group) is much greater than the default payment group with

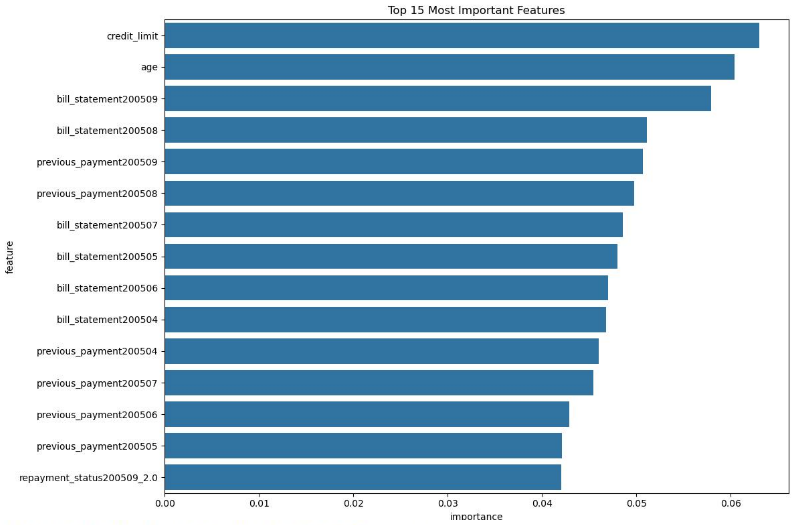
87% precision rate and 84% recall rate, however, the moderate Classification 1 performance still has room for improvement.

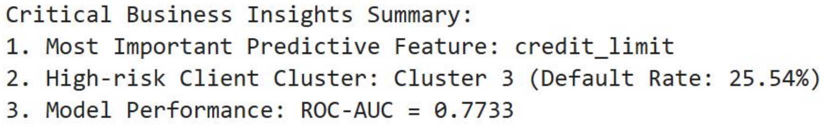


```
Cross-Validation ROC-AUC Scores: [0.75808545 0.77486695 0.78399709 0.75329802 0.76907935]
Average CV Score: 0.7679 (+/- 0.0222)

Test Set Performance:
Accuracy: 0.7804002052334531
ROC-AUC Score: 0.7732541062971551

Confusion Matrix:
[[3817  726]
 [ 558  746]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.87      0.84      0.86      4543
         1.0       0.51      0.57      0.54      1304

    accuracy                           0.78      5847
   macro avg       0.69      0.71      0.70      5847
weighted avg       0.79      0.78      0.78      5847


Random Forest CV ROC-AUC: 0.7648 (+/- 0.0234)
```

3. Random Forest

We extract the feature importance and classifier after training the random forest pipeline. And we get feature names by combining numeric features with one-hot encoded categorical features in order to create the importance dataframe. The visualization shows that the most significant predictive feature is credit_limit, followed by age, bill_statement200509, bill_statement200508. The high-risk client cluster is Cluster 3 with default payment rate of 25.54%, and the overall ROC-AUC score for model is 0.7733, validating its reliability for prioritizing risk assessments and tailoring client management strategy.





```
Critical Business Insights Summary:
1. Most Important Predictive Feature: credit_limit
2. High-risk Client Cluster: Cluster 3 (Default Rate: 25.54%)
3. Model Performance: ROC-AUC = 0.7733
```

Evaluation

When selecting the optimal clustering number for Principle Component Analysis, we generally choose the elbow point. After generating the scatter plot and cluster profile, we can compare the data among different clusters intuitively, in order to identify the variable characteristic of target groups, as well as the most direct influence factors on the dependent variable.

In terms of K-Fold Cross Validation, the ROC-AUC Score can directly reflect the discrimination accuracy of model's prediction results, and a lower standard error represents strong stability. Meanwhile, the confusion matrix displays the number of TN(True Negatives), FP(False Positives), FN(False Negatives) and TP(True Positives). The precision and recall rate, as well as macro average accuracy and weighted average accuracy can all be calculated based on this to show the model does better on which classification.

Finally, the Random Forest Model reflects the influence weights of different feature variables, which directly guides the businesses with strategic resource allocation.

Summary

Through our data analysis and modeling results, we know that the most highly-correlated influence factor for default payment rate is credit limit. The higher the client's credit limit, the less likely he or she is to default on payment. The following influence factors are bill statement, previous payment and repayment status. Thus we reach the final conclusion that the clients with larger credit limit, bill statement, previous payment and past on-time repayment records are least likely to default on payment. Conversely, those who own lower credit limit, bill statement, previous payment and leave default payment records in the past are the high-risk population.

Deployment

For businesses to project expected improvement, I recommend Method 1: K-Means Clustering with PCA visualization or Method 2: K-Fold Cross Validation. They can integrate statistical model into credit decisions and portfolio management by estimating reduced default lose for high-risk clusters and increased revenue from low-risk client growth. The precise calculation of ROI maybe a little bit challenging due to uncontrollable external factors, so A/B Testing and incremental rollout can both serve as a viable alternative for measurement. Ethical issues like fairness across demographics and data privacy regulations should be taken into account to avoid the disproportionate risk measure that may unfairly impact certain groups.

The limitation of our work is that the forecast accuracy for default-payment group is lower than the non-default group, which may potentially result in the failure of default payment detection. To address such problem, in the future maybe we can generate more synthetic default cases using techniques like SMOTE, adjust model weights to prioritize default cases and lower the classification threshold in order to catch more true defaults.

Appendix:

Contribution of Team Members:

Data Cleaning: Shicheng Li (sl975)

Data Analyzing: Shicheng Li (sl975), Shifei Ruan (sr678), Alaina Hu (ayh16), Weigao Tan (wt99)

Data Modeling: Shifei Ruan (sr678), Ricky Han (rh396), Samiha Jain (sj472)

Report Writing: Shifei Ruan (sr678), Shicheng Li (sl975)

Slides Making: Samiha Jain (sj472), Alaina Hu (ayh16), Weigao Tan (wt99), Shifei Ruan (sr678), Shicheng Li (sl975), Ricky Han (rh396)