The Business Question of Credit One

Use the data to predict the probability of default

Business Questions

- The number of customers who have defaulted on loans is increasing. Credit One could risk losing business.
- What are the key factors to decide one's probability of default? Can we get a better scoring method from the historical data?
- Is there a better way to decide how much credit to allow the customers to use?

Data Science Process

1. Collect the Data	• Pull out the data set of customers from the open sauce mySQL. • Extract data into Python dataframe and Excel format.
2. Process the Data	• Examine data: Understand the column and data type. • Clean the data: Drop the duplicate, missing values, errors.
3. Explore the Data	• Split and plot the data in different ways. • Test the correlation of variables.
4. Perform In-Depth Analysis	• Create the predictive models • Evaluate the models. • Revise the models and go back to explore the data if needed.
5. Communicate Results	• Answer the business problems. • Visualize the findings in simple way. • Communicate the results to non-technical people.

Process the Data

- Number of data : 29965
- PAY_0-PAY_6: payment history (-2: no consumption-1: paid in full; 0: use of revolving credit; 1: payment delay 1 month)

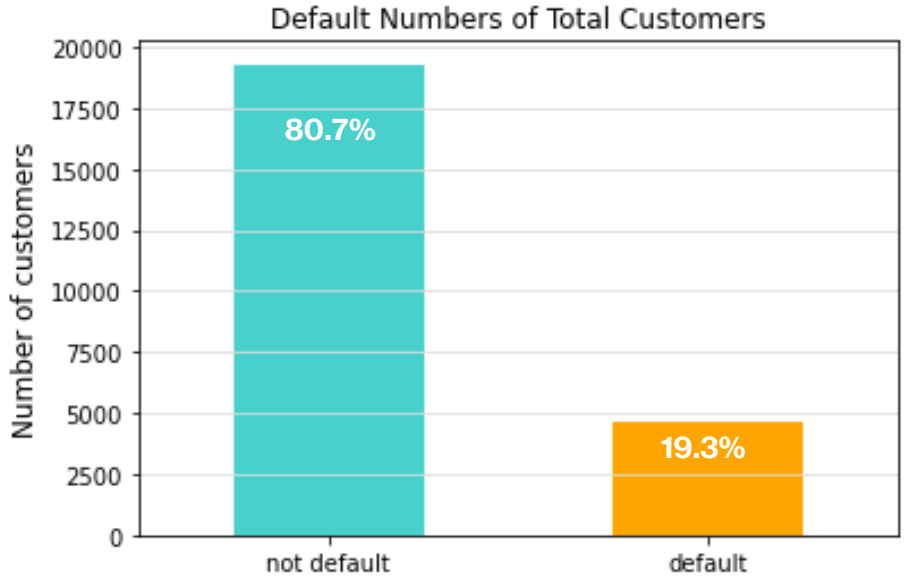
LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1
440000	Male	graduate school	Married	79	0	0	0	0	0	0	429309
250000	Female	university	Married	75	0	-1	-1	-1	-1	-1	52874
180000	male	graduate school	Married	75	1	-2	-2	-2	-2	-2	0
210000	Male	university	married	75	0	0	0	0	0	0	205601

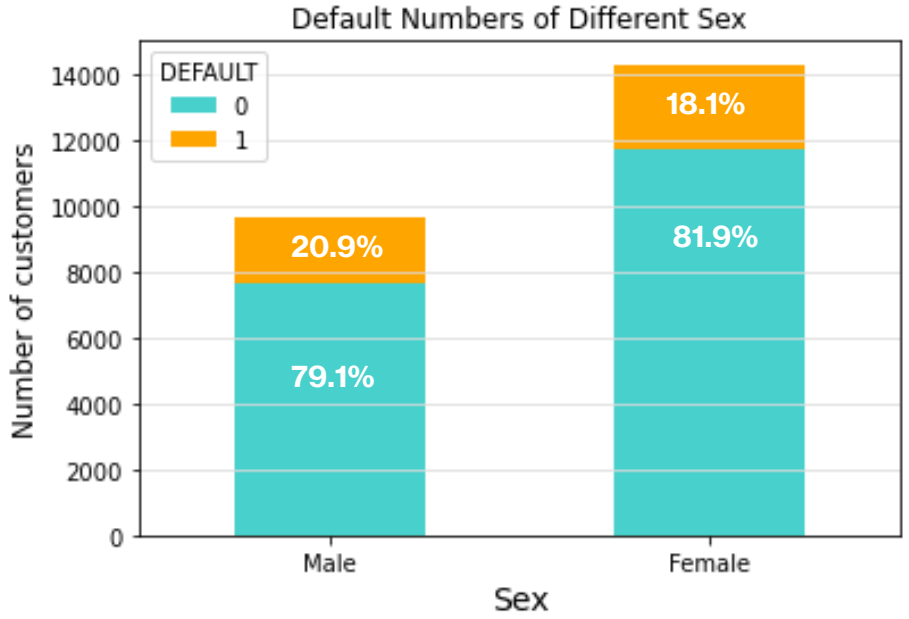
BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	DEFAULT
437906	447326	447112	438187	447543	15715	16519	16513	15800	16531	15677	No default
1631	1536	1010	5572	794	1631	1536	1010	5572	794	1184	No default
0	0	0	0	0	0	0	0	0	0	0	default
203957	199882	203776	205901	210006	9700	8810	9000	7300	7500	7600	No default

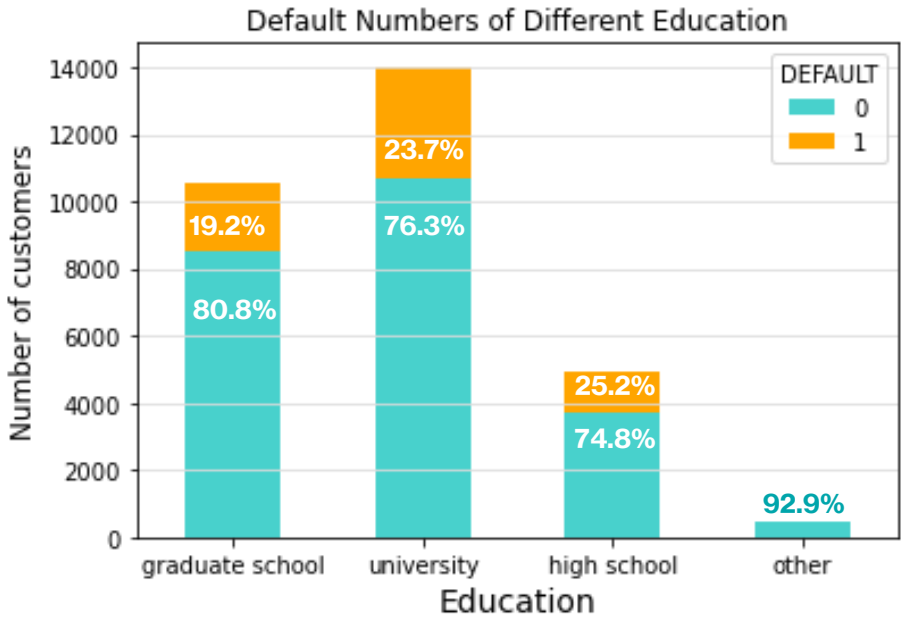
Exploratory Data Analysis (EDA)

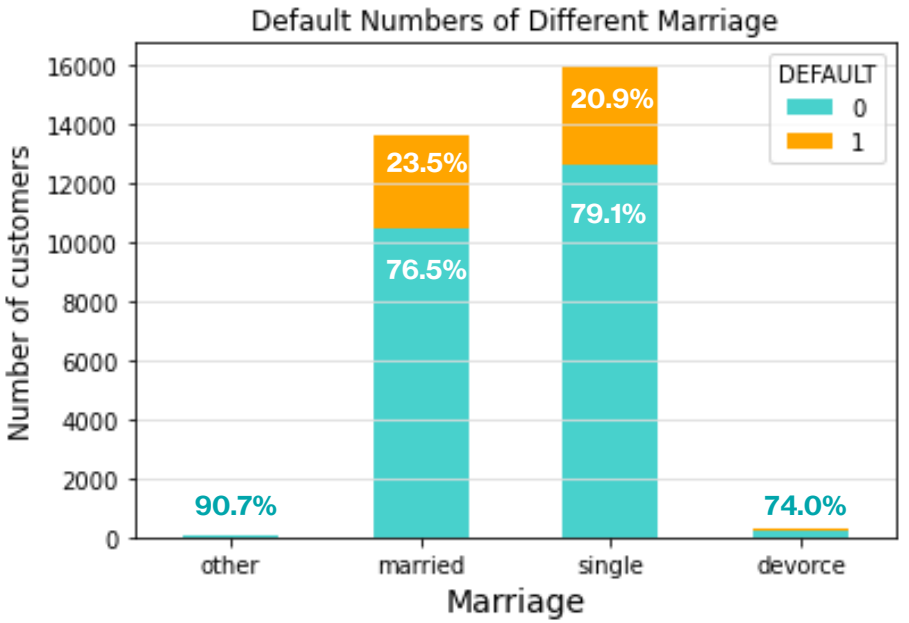
0: not default
1 : default

- The proportion of default customers is lower with higher education level.
- Customers with other and single status have the lowest proportion of default. Customers with divorce status have the highest proportion of default.





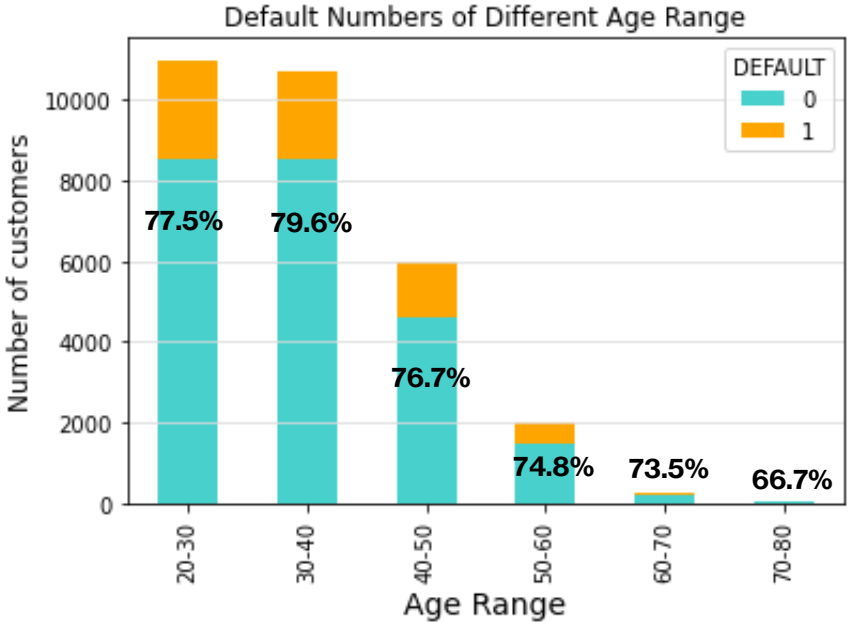


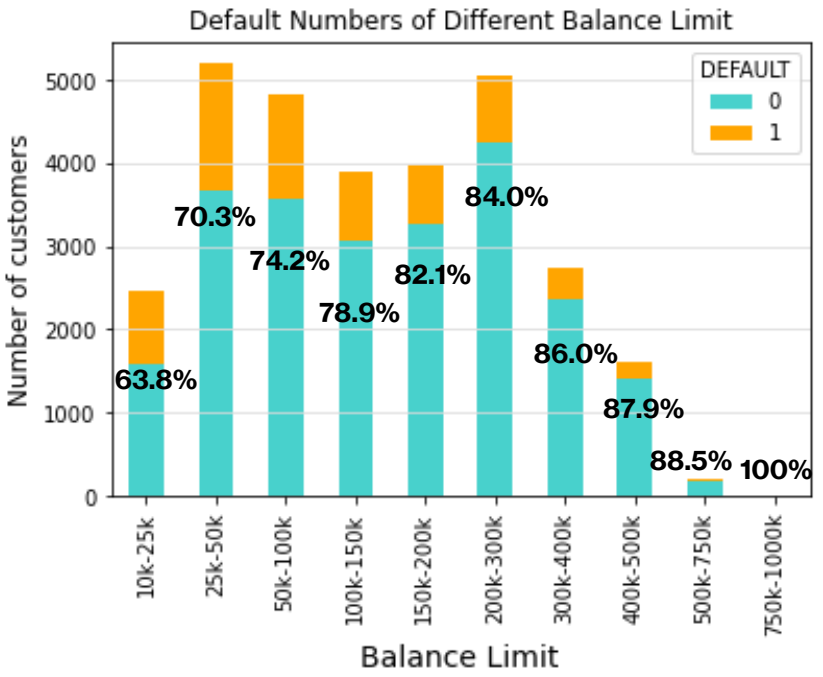


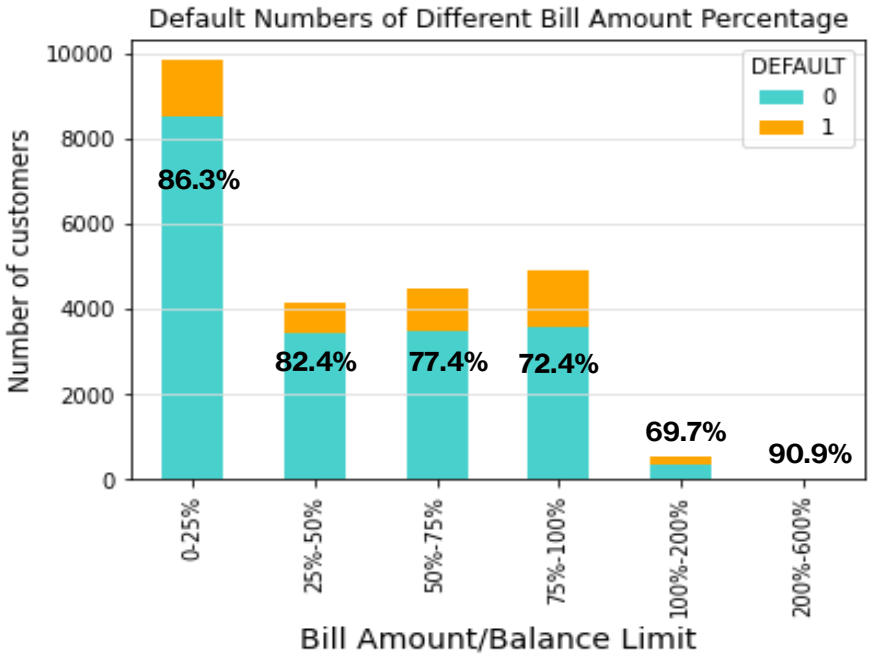
Exploratory Data Analysis (EDA)

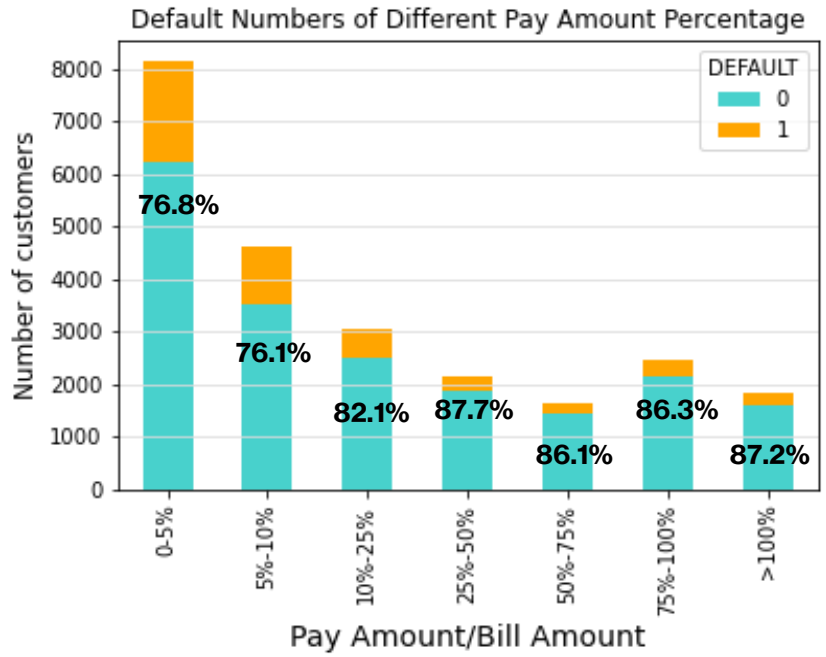
0: not default
1 : default

- The proportion of default customers decrease when the balance limit increase.
- The proportion of default customers increase when the ratio of bill amount to balance limit increase.
- The proportion of default customers is higher when the ratio of pay amount to bill amount is lower than 10%.









Perform In-Depth Analysis

Create and Evaluate the Models.

The prediction accuracy of three different model is range from 77.98% to 82.14%.

Gradient Boosting Classifier Model has the highest prediction accuracy.

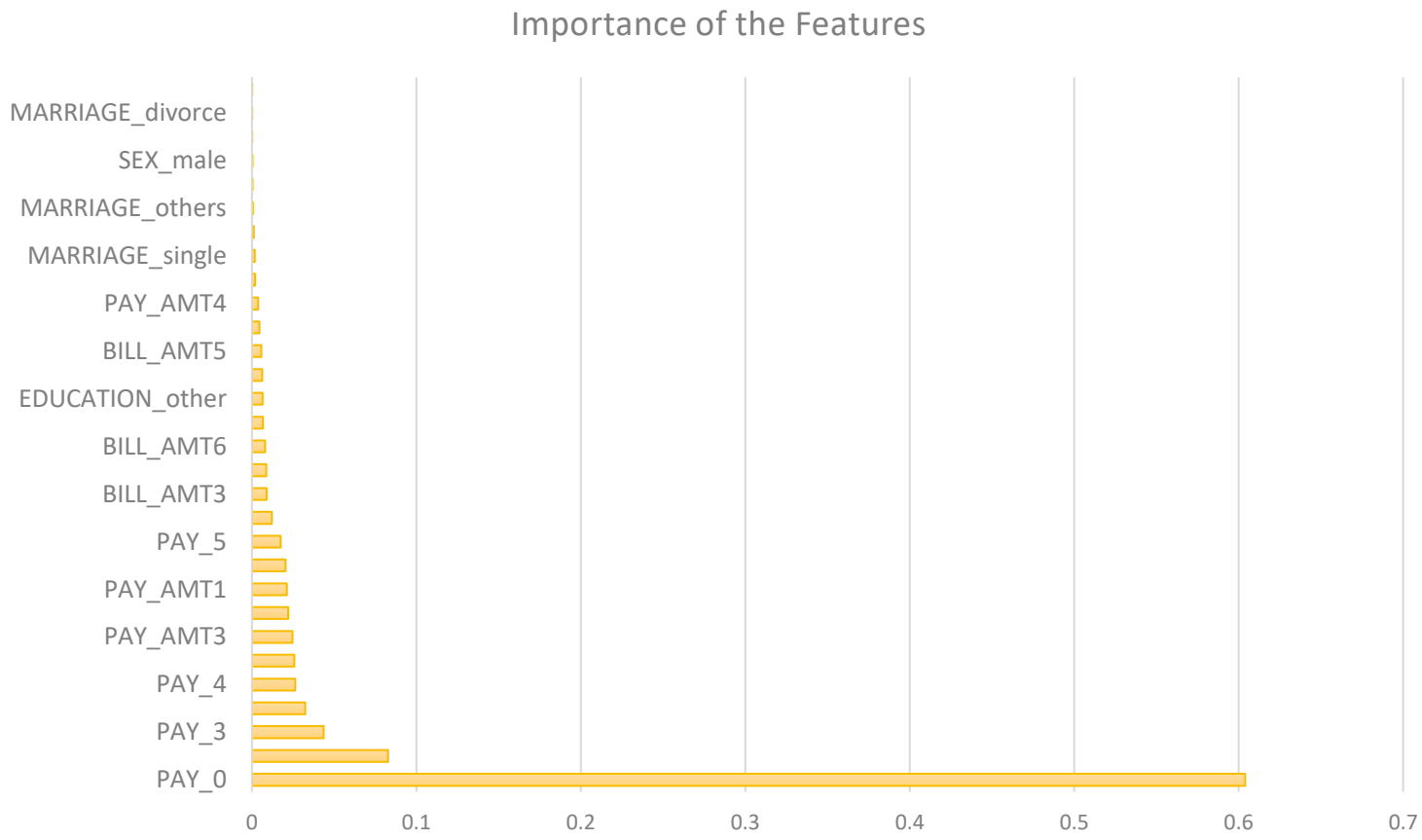
Models' Name	Accuracy of Prediction
Random Forest Classifier	81.38%
Gradient Boosting Classifier	82.14%
SVC	77.98%

Perform In-Depth Analysis

The Importance of the Features.

The importance of the features for predicting the default possibility of customers.

PAY_0 is the most importance feature.



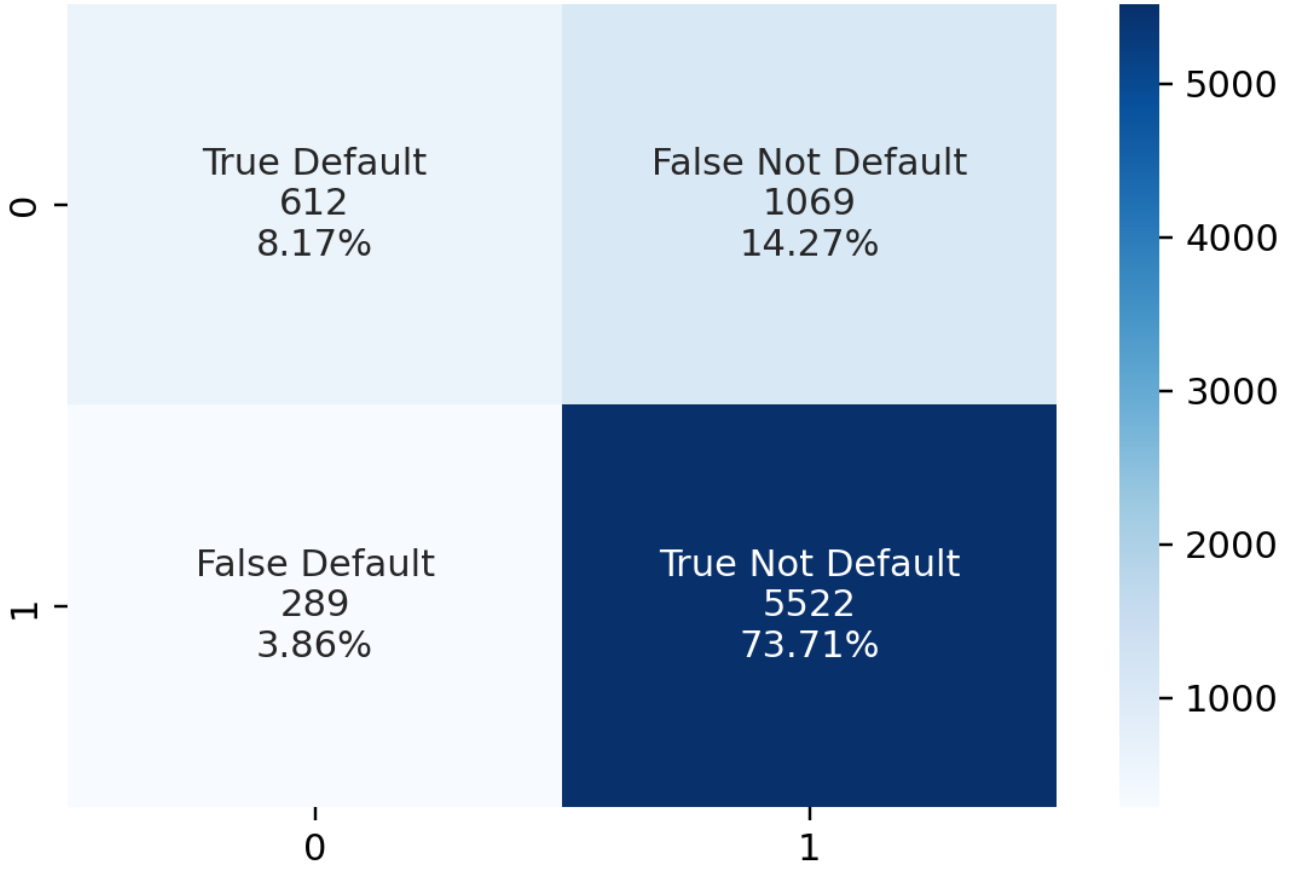
Perform In-Depth Analysis

Confusion Matrix

- The total test samples is 7492.
- The Gradient Boosting Classifier Model predict result:

  Predict 6591 customers will not default.
  Only 5522 customers will truly not default.
  The prediction accuracy is 83.78%

  Predict 901 customers will default.
  Only 612 customers will truly default.
  The prediction accuracy is 67.92%



Answer the Questions

What are the key factors to decide one's probability of default? Can we get a better scoring method from the historical data?

We can use the historical data to predict one's probability of default with 82% certainty. The most important factor of the prediction is the most recent payment history.

The proportion of default customers varies with education, marriage, bill amount/credit limit, pay amount/bill amount.

Thank you!

Presented by Wanyun Ho
howanyun@gmail.com

Observations and Recommendations

- I used three different classification algorithms to build the machine learning models. The highest accuracy of prediction the model can get is 82.14%. However, from the confusion matrix, the prediction of the default customers can only reach 67.92%. For improving the prediction accuracy, I may need to try more different models, or I need more features of the dataset.
- For the other business question, “Is there a better way to decide how much credit to allow the customers to use?” I can’t answer it with the data I have, I may need more customers’ information. For example, the salary, the current loan status, credit history etc. may be useful features for building up the model to decide one’s credit limit.
- For the visualizations, I found there’s relationship between the default proportion and education, marriage, balance limit, the ratio of bill amount to balance limit, and the ratio of pay amount to bill amount. However, from the importance of the features analysis, I found the importance of these features is relatively low. Since I don’t know how the importance of features is calculated, I can’t figure out the reason.