DSCI 310 Group 4: Default payments in Taiwan and comparison of the predictive accuracy of probability of default

Shravan Chaniara, Jordan Yu, Diana Liang, Hannah Martin

Contents

Abstract

Financial institutions incur monetary loss when a client or borrower is unable to pay their interest or their initial principal on time. Thus, it is necessary for such institutions to assess the risk that potential borrowers cannot repay their loan in determining their eligibility for the loan in the first place. The present study endeavors to answer the question “Is there a way to effectively predict whether or not a client will default on their credit card payment?” and uncover the most significant features that contribute to the higher likelihood of defaulting. The result of predictive accuracy of the projected likelihood of default will be more beneficial than the binary result of categorization - credible or not credible customers - from the standpoint of risk management.

Introduction

Amidst growing financial insecurity during the pandemic, unsecured debt has continued to rise (Frech (2021)). Consequently, the consumer credit market and risk prediction has been a matter of great speculation and fear, lest there be a repeat of the financial crises that rocked the economic world in the late 2000s: In 2006, Taiwan was rocked by a credit card debt crisis with debt from credit cards and cash cards reaching $268 billion USD and over half a million people unable to repay their loans (Yeh (2009)). As many could barely afford to pay the minimum credit card debt balance every month or continued to default on their payments, significant societal problems consequently plagued the country, many banks incurred heavy losses and the government eventually needed to step in to stabilize the financial system (Yeh (2009)). This situation arose because many banks in Taiwan had lowered the requirements for credit card approval in order to gain more customers within the increasingly competitive industry (Tsai (2010)). Such examples indicate that a strict assessment of an applicant’s capability to make their card payment is critical to a well-developed financial system and a business’s survivability in the banking industry.

Table 1: Summary Stats

X	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	
mean	14976.622	166849.3	1.6046667	1.8533750	1.5550417	35.41963	-0.0139167	-0.1339167	-0
std	8653.231	129755.4	0.4889324	0.7868091	0.5224766	9.18461	1.1263995	1.2045235	1
min	1.000	10000.0	1.0000000	0.0000000	0.0000000	21.00000	-2.0000000	-2.0000000	-2
max	30000.000	1000000.0	2.0000000	6.0000000	3.0000000	75.00000	8.0000000	8.0000000	8

This project focuses on the case of customers default payments in Taiwan and finds the predictive accuracy of the probability of the customers to default. The purpose of this study is to assess the true probability of default because the real probability of default is unknown.

Dataset Information

This project used the data Dua (2016) from UCI Machine learning repository. As the response variable, this project used data from a binary variable, `default payment` (Yes = 1, No = 0). The following 23 factors were considered as explanatory variables in this study, which was based on a review of the literature:

- Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- Gender (1 = male; 2 = female).
- Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- Marital status (1 = married; 2 = single; 3 = others).
- Age (year).
- History of past payment. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months . . . 8 = payment delay for eight months; 9 = payment delay for nine months and above.
- Amount of bill statement (NT dollar).
- Amount of previous payment (NT dollar).

The objective of this project is to maintain ease of interpretation for the average reader. In line with this goal, we will simplify the models and methods of analysis we choose to use as well as exclude some features in the data set in favor of greater readability.

Methods and Results

Exploratory Data Analysis

First, we load and tidy the data. The dataset was split into a 80% training and 20% testing set. The model will be built using only the training data. This gives the ability to compute a final performance metric for our model by evaluating it on the testing data. train_test_split() function shuffles the data to ensure the data ending up in the training and test sets is randomized.

Before any data analysis can be done let's check if the dataset has null/missing values that may affect further analysis: The dataset is super clean, and no missing value is found.

Statistics information

This part we have a look at some basic statistics of the training set:

We can see the following information:

- The average amount of given credit in NT dollars is 166849.320000;
- The average age of all clients is 35.419625, etc.

To better understand the correlation between variables, we would like to compute and visualize the correlations:

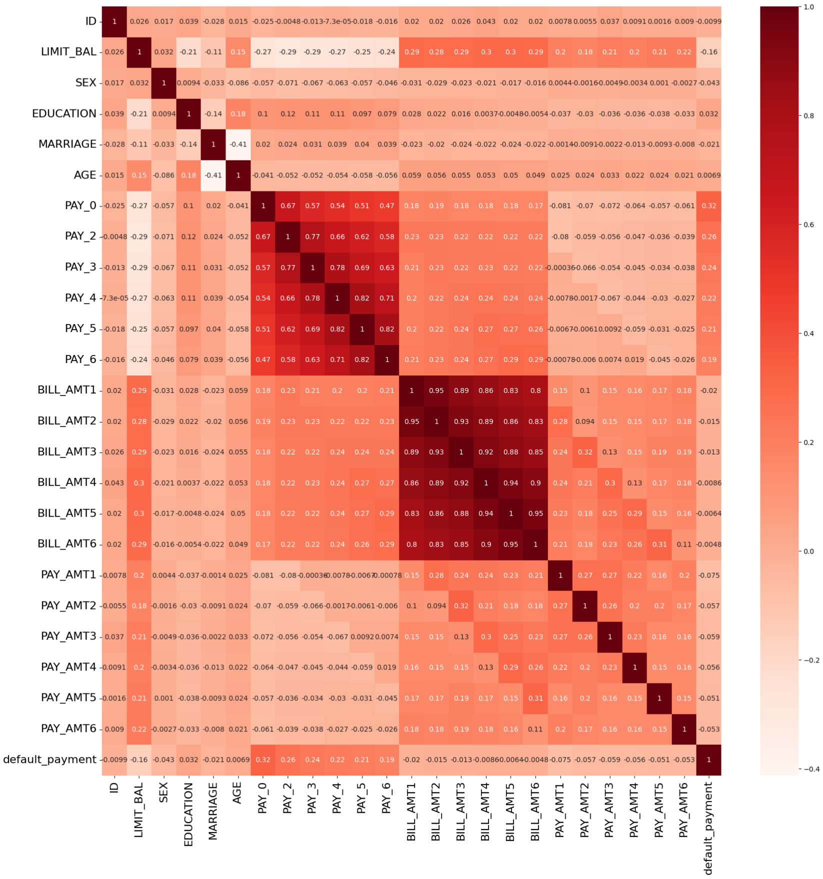

Figure 1: Heat Map

The heatmap shows some positive/negative correlations:

Positive correlations:

- `Default payment` - `PAY_0` to `PAY_6` (Repayment status from April to September, 2005);
- `Limit balance` - `BILL_AMT1` to `BILL_AMT6` (Amount of bill statement from April to September, 2005), etc.

Negative correlations:

- `Limit balance` - `PAY_0` to `PAY_6` (Repayment status from April to September, 2005), etc.

Specifically, `PAY_0` has the highest correlation with `default_payment`. This will give us a signal that `PAY_0` plays an important role for predicting `default_payment`.

Exploring variables

LIMIT_BAL

First, we look at the amount of given credit (in NT dollars). Credit card limits are likely an indicator of how wealthy someone is since banks tend to give higher limits to clients that have more money with them. Thus, this may be an important feature when predicting if someone is able to pay the bill on time.

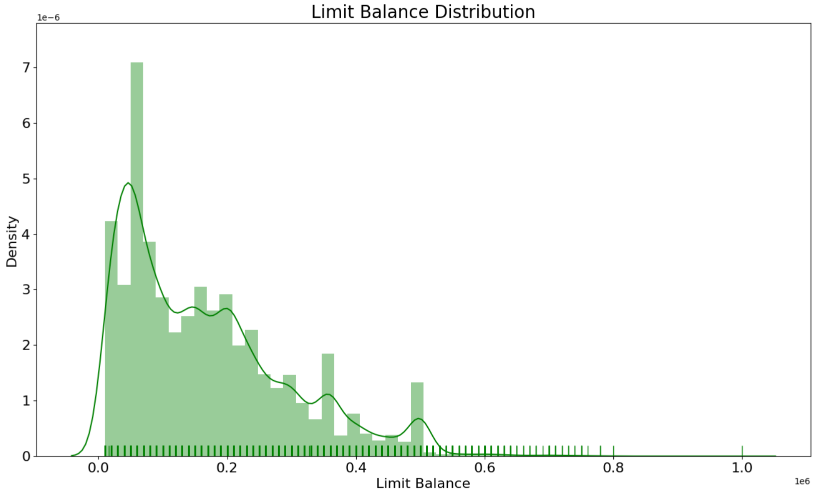

Figure 2: Limit Balance Distribution

Repayment Status

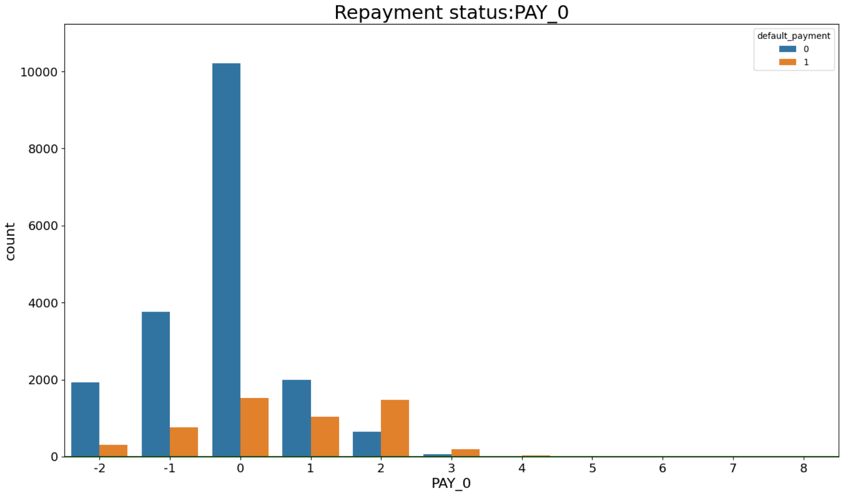

Figure 3: Repayment Status PAY 0

Looking at the above plots on repayment status shows that if a client defaults on their payment for 2 months (e.g PAY_X = 2), it is a indicator to predict that default_payment = 1. It looks like repayment status will be an important feature in the model.

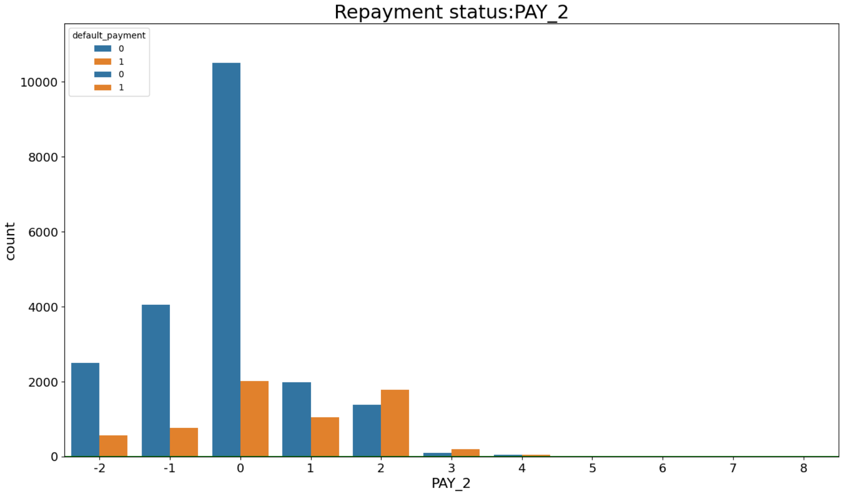

Figure 4: Repayment Status PAY 2

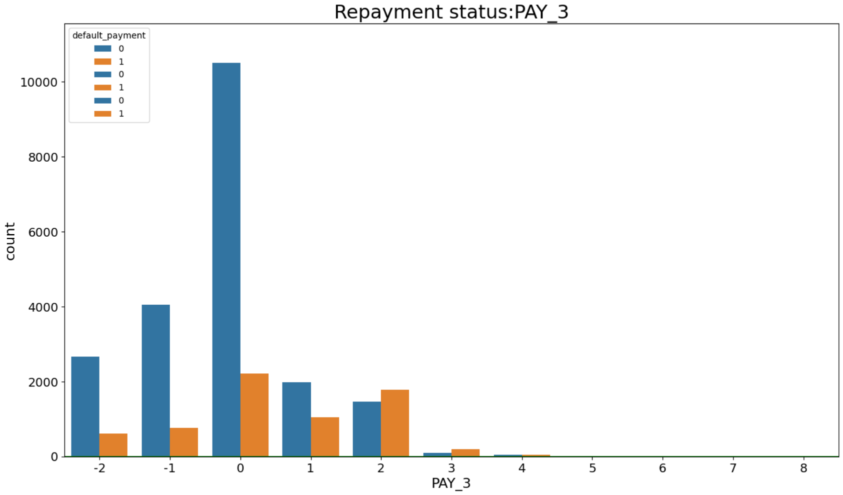

Figure 5: Repayment Status PAY 3

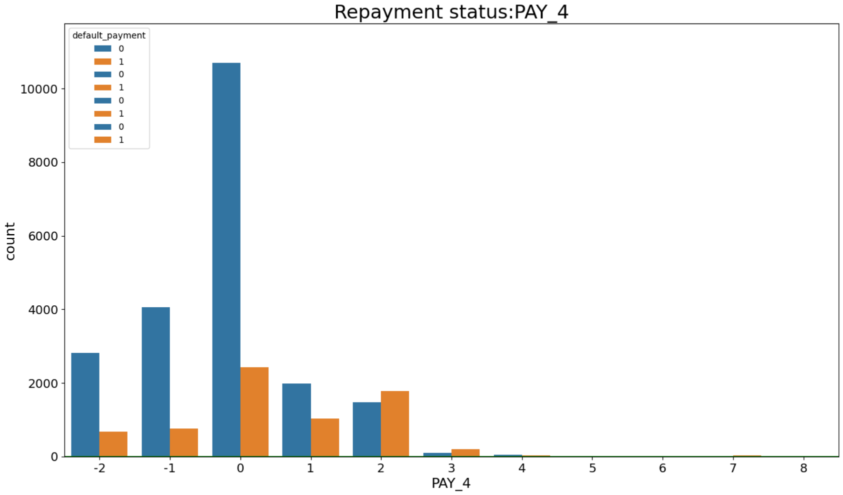

Figure 6: Repayment Status PAY 4

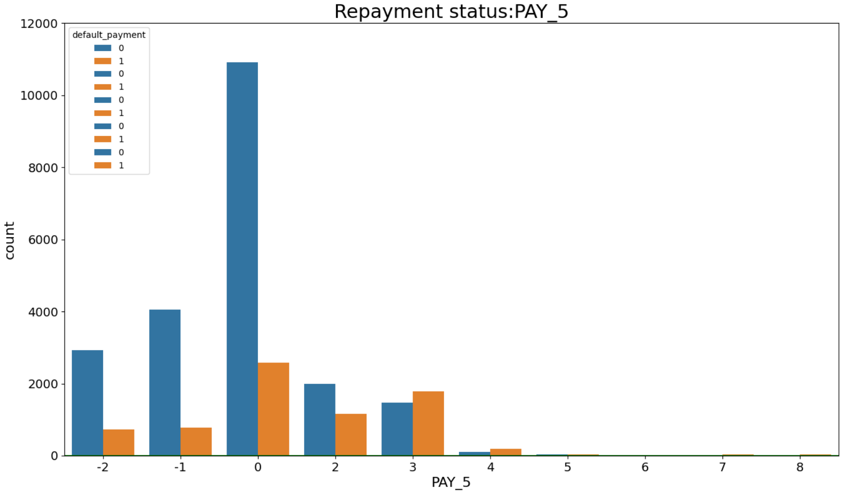

Figure 7: Repayment Status PAY 5

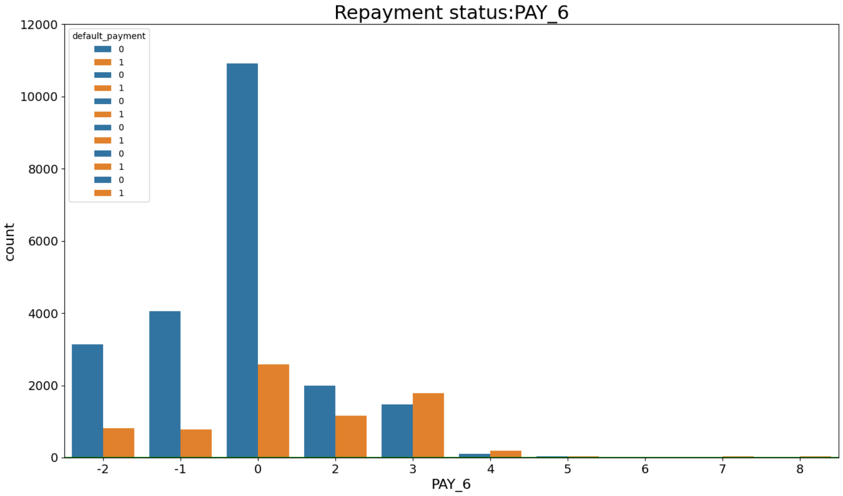

Figure 8: Repayment Status PAY 6

Analysis

Class Imbalance

The above plot shows the percentage of rows with default payment = 0 versus default payment = 1. Clearly there is class imbalance in this dataset. Because of this, we will use the area under the Receiver Operating Characteristic curve (ROC AUC) as our primary metric to evaluate our model instead of accuracy, which tends to be misleading in cases of class imbalance. ROC AUC evaluates how good the model is as distinguishing between classes, and gives a more accurate sense of how well our model generalizes when dealing with class imbalance.

Preprocessing

We apply scaling to numeric features to ensure the model built will be robust and not sensitive to the scale of each individual feature. To do this, we use the StandardScaler() function to set the sample mean to 0 and standard deviation to 1.

To handle categorical values, we will apply one-hot-encoding. This creates binary dummy variables for each category. Next, we apply scaling and one-hot-encoding through using a column transformer, which applies the transformations to each column specified.

Model 0: Dummy Classifier

Firstly, we will try a baseline model to act as a comparison measure for the final model built. Using a pipeline to do this ensures that the model is built using just the training data, and that the testing data has no influence on the model. 5-fold Cross validation is used to give a more robust measure of performance error. K-fold Cross-validation spilts the training data into k folds, and each time one fold is the validation set. Each fold fits the model on the training portion and uses one fold as a validation set to calculate a performance metric, which can be averaged to get a overall score of how well the model does. This ensures that outliers don't negatively influence the performance metric.

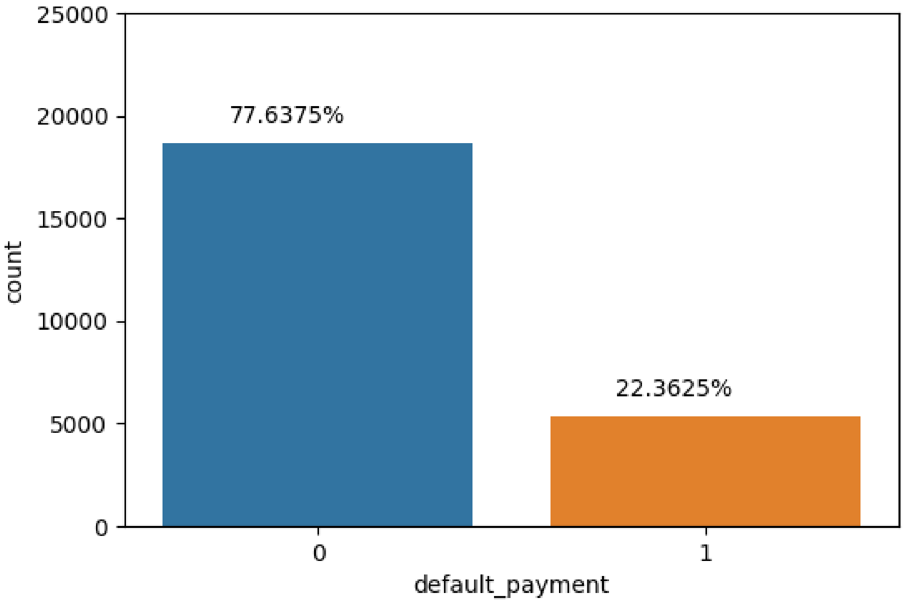

Figure 9: Count Plot

fit_time	score_time	test_score	train_score
0.0559318	0.0190411	0.5	0.5
0.0513411	0.0239758	0.5	0.5
0.0591900	0.0192940	0.5	0.5
0.0494611	0.0184088	0.5	0.5
0.0515850	0.0185170	0.5	0.5

Model 1: Logistic Regression

Next, a logistic regression model is fitted to the training data. Logistic regression uses the training data to learn coefficients, which then can be used to calculate prediction probabilities of the each class using the sigmoid function. This allows us to calculate the probability of a client defaulting on their credit card based on their data. The hyperparameter C is used to control the fundamental tradeoff of bias and variance, to reduce the likeliness of the model overfitting or underfitting. We chose logistic regression over more complex classifiers such as ensemble trees because it is easier to interpret and efficient to train. The predicted coefficients give information about feature importance and direction of association, making the model easily interpreted. Moreover, since the dataset is significantly larger than the number of features, logistic regression is less likely to overfit because it is a low variance model.

Hyperparameter optimization is carried out to find the best value of C for the data in hopes to reduce bias and variance. Cross-validation is used to test how well the model performs on unseen data during hyperparemter opimization, enabling performance metrics to be calculated to compare different values of C.

C	Train.Scores	CV.Scores
0.0316228	0.7724754	0.7689002
0.1000000	0.7727518	0.7686736
0.3162278	0.7728306	0.7684365
1.0000000	0.7729194	0.7682234
3.1622777	0.7729977	0.7678970
10.0000000	0.7730728	0.7674601
31.6227766	0.7731342	0.7671300

Next, the logistic regression model with the optimized C value is fitted to the training data. The model is built using pipelines to ensure all data preprocessing is constant and that no information from the testing set leaks into the training of the model.

Feature Importance

Next, we will look at feature importance. In logistic regression, the magnitude and direction of the learned coefficients explain the relationship between a explanatory feature and the response variable.

coefficient	absolute_value
1.0613115	1.0613115
-1.0045704	1.0045704
0.7219127	0.7219127
-0.6688425	0.6688425
-0.3219945	0.3219945
-0.3176690	0.3176690
-0.3148366	0.3148366
-0.2489065	0.2489065
0.2488997	0.2488997
0.2450337	0.2450337
-0.2250865	0.2250865
0.2146013	0.2146013
-0.1767683	0.1767683
-0.1764516	0.1764516
-0.1757003	0.1757003
-0.1721139	0.1721139
0.1665569	0.1665569
-0.1638805	0.1638805
-0.1632858	0.1632858
0.1605075	0.1605075
-0.1488593	0.1488593
-0.1480731	0.1480731
0.1464547	0.1464547
-0.1462427	0.1462427
0.1458649	0.1458649
0.1420812	0.1420812
0.1373272	0.1373272
0.1355065	0.1355065
0.1346587	0.1346587
0.1325932	0.1325932
0.1277497	0.1277497
0.1256826	0.1256826
-0.1227187	0.1227187
0.1209245	0.1209245
0.1199460	0.1199460
0.1010633	0.1010633
-0.0964676	0.0964676
-0.0955501	0.0955501
-0.0954486	0.0954486
0.0942384	0.0942384
0.0928493	0.0928493
0.0898831	0.0898831
-0.0848013	0.0848013
-0.0839578	0.0839578
0.0808301	0.0808301
-0.0800034	0.0800034
-0.0766308	0.0766308
-0.0752523	0.0752523
0.0677422	0.0677422
-0.0666298	0.0666298
0.0648119	0.0648119
-0.0618421	0.0618421
-0.0585649	0.0585649
-0.0558367	0.0558367
0.0548550	0.0548550
-0.0491085	0.0491085
-0.0457311	0.0457311
0.0457020	0.0457020

The table above shows the most important features according to the model. Positive coefficients indicate that an increase in the feature increases the probability that the response variable is class 1, and negative coefficients indicate that a increases in the feature decreases the probability that the response variable is class 1.

The most important feature is repayment status in April with a 2 month delay.

To further interpret this, the odds ratio can be calculated. $OR = e^{\beta}$, where β is a model coefficient.

The odds ratio for x0_2 is 2.8901590141210964 can be interpreted as clients who delayed payment for 2 months in April have about 2.9 times the odds of defaulting their credit card payment next month than those who did not delay for 2 months of, controlling for the other features since there is some correlation.

It makes logical sense that this feature is important since it was highly correlated with the response variable during the exploratory data analysis, and because it makes sense that if someone has delayed payment in the past they may not be able to pay in the future as well.

Another important feature is x0_0, corresponding to repayment status in April with a revolving credit. The magnitude is negative here meaning that a value of 1 for this binary variable is negatively associated with defaulting next month. This makes sense because revolving credit lets clients pay a minimum balance instead of the full bill, making them less likely to default. The odds ratio is calculated below.

The odds ratio is 0.3662019170982376 shows that clients who had a repayment status of using a revolving credit in April decreases the odds of class 1 versus class 0 by about 63%, controlling for the other features.

Due to the large number of features, we only looked at the first 2 most important features, however this method can be done for the remaining features as well.

Testing the Model

Finally, the model will be evaluated by predicting on the test dataset. Performance metrics including area under the ROC curve, f1-score, precision and recall will be calculated. Precision is the ratio of true postive and total positives. Recall is the measure of the model correctly identifying true positives. F1-score is the harmonic mean of precision and recall.

The test ROC AUC is almost the same as the cross-validation ROC AUC so we can conclude that there is little optimization bias and we are not overfitting. Area under the ROC curve is the classifier's ability to distinguish between classes. So 78% of the time, the model is able to correctly distinguish between class 0 and class 1.

X	recall	precision	f1.score
0	0.359	0.669	0.467

It appears that the model does significantly better for the non-default class, since precision, recall, and f1-score are very high. Recall is particularly low for the default class, and very high for the non-default class. Thus this model is very good at identifying clients who will pay their bill on time. Precision can be interpreted as when the model predicts class 0 (non-default) it is correct 84% of the time, and when it predicts class 1 it is correct 64% of the time, which is not as good.

Discussion

Looking at the results above, the analysis has a few limitations. One of the main limitations is the extrapolation of data. The dataset used was from Taiwan so the model may not apply properly for financial institutions and people outside of Taiwan. The other limitation is that logistic regression favors interpretability over prediction accuracy. The coefficients are easy to understand while the logistic regression is not as accurate as other algorithms.

Another limitation of the analysis is that logistic regression is a higher bias model than other classifiers such as ensemble trees. This means that the model pays less attention to the training data and may oversimplify

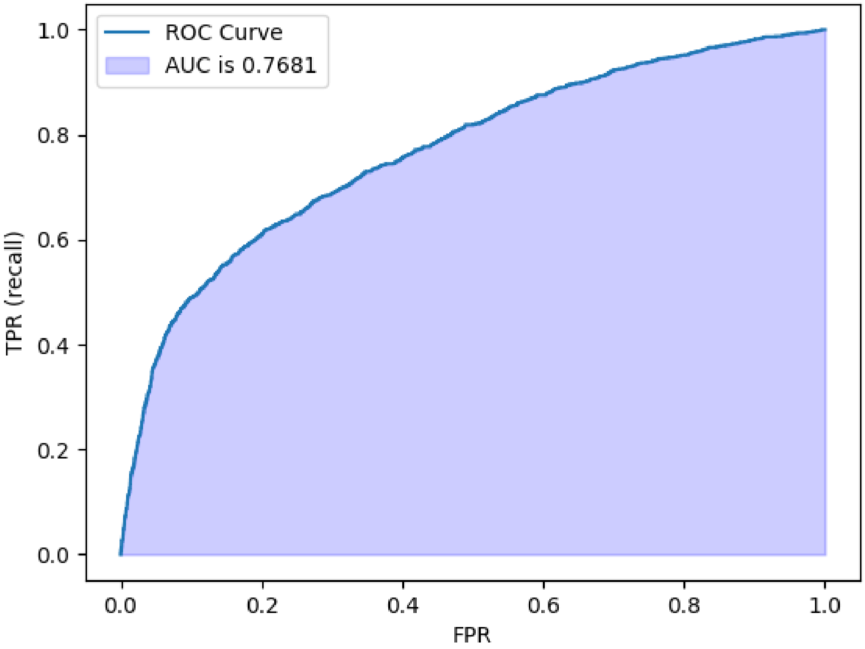

Figure 10: ROC

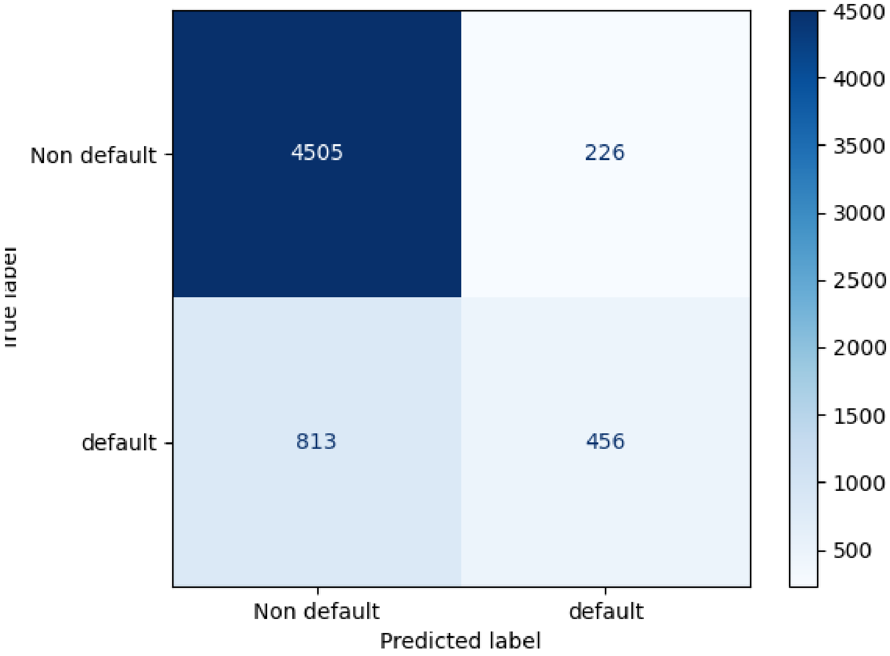

Figure 11: Confusion matrix

it, resulting in less accuracy. However, the results show that the model did very well when predicting clients who did not default next month (class 0). The high recall and precision score for class 0 suggest this model can be very useful when trying to distinguish if a client will pay the bill on time. Thus, this model can be used to help financial institutions make various risk decisions about a client including whether to offer someone a loan or whether to raise the credit card monthly limit. A prediction probability can be calculated based on the input features of each individual client. A prediction probability close to 0 indicates that the client is likely to pay the credit card bill.

According to the model, the most important feature was repayment status. Specifically, whether or not someone delayed payment for 2 months, and whether or not someone used a revolving credit card. This provides insight into what features financial institutions should look for when assessing risk for individual clients. Also, the model suggests that giving someone a revolving credit card where they only have to pay a minimum balance instead of the whole bill each month reduces the probability of the client defaulting. This information could be used to prevent the institute from losing money due to people defaulting.

Overall, the model provides insight into important features that may benefit financial institutions when managing risk, and allows an institute to calculate the probability that an individual client will default.

References

Dua, & Graff, D. 2016. *UCI Machine Learning Repository: Default of Credit Card Clients Data Set [Data Set]*. University of California, Irvine, School of Information; Computer Science. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

Frech, Houle, A. 2021. *Trajectories of Unsecured Debt and Health at Midlife. SSM - Population Health.* Vol. 15. https://doi.org/10.1016/j.ssmph.2021.100846.

Tsai, B.-H. 2010. *Gauging Bank Efficiency During Card Insolvency Crisis: The Case of the Taiwanese Banks. The Journal of Developing Areas.* Vol. 44. https://doi.org/10.1353/jda.0.0087.

Yeh, & Lien, I-Cheng. 2009. *The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. Expert Systems with Applications.* Vol. 36(2). https://doi.org/10.1016/j.eswa.2007.12.020.