PREDICTED PROBABILITY OF DEFAULT

Default of Credit Card Clients in Taiwan

© Katie Lashko
Kateryna.Lashko@gmail.com

SUMMARY

During this phase of the project, the main goal was to build a machine learning model, which would derive predicted probabilities of default of credit card clients. Due to extreme project time constraints, only a logistic regression was built using a stepwise selection method and a p-value of 0.03. The model was able to achieve a c-statistic of 0.76 on the training dataset and 0.74 on the validation dataset (as illustrated in Figure 1 and Figure 2 below). The optimal cut-off point for the final model, as determined by K-S statistic, was 0.23.

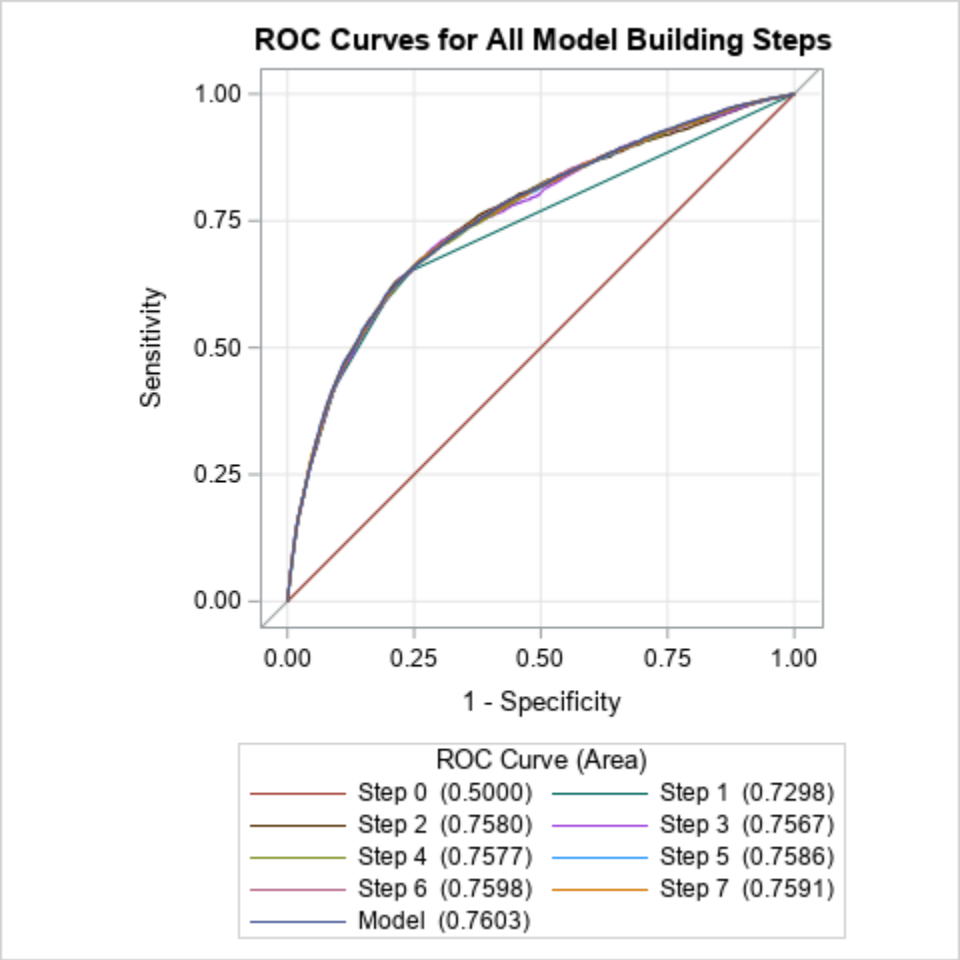

Figure 1: Training data ROC Curve

Figure 2: Validation data ROC Curve

RESULTS and RECOMMENDATIONS

The final model included eight variables, as outlined in Table 1. Moreover, there are several main takeaways from odds ratios of the main effects in the model.

Table 1: Main Effects Model

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age Category	6	18.46	0.0052
Bill Amount (month 4)	1	12.7687	0.0004
Education	3	18.233	0.0004
Credit Limit	9	38.5393	<.0001
Marital Status	2	19.1513	<.0001
Cumulative Payment Amount	1	57.2675	<.0001
Payment History	6	2216.2628	<.0001
Gender	1	13.9735	0.0002

More specifically, customers who missed more than 12 payments are almost 16 times more likely to default than those customers who missed no payments. Moreover, customers who missed at least one payment are 2.6 times more likely to default than those customers who had a positive payment history. For a complete list of odds ratios, please refer to Appendix Table 1.

Furthermore, the model shows that borrowers over the age of 60 have 1.24 times the odds of defaulting on a loan, as compared to 50 to 60-year-olds. Also, the odds of incurring a default is about 1.4 times higher for loan amounts between $700,000-$800,000 as compared to $10,000 to $50,000 loans.

Lastly, although the final model achieved a relatively high concordance on both train and validation datasets (76% and 74%, respectively); there are two main recommendations to improve the performance of a future model-- further explore potential data inconsistencies and include additional possible predictors, such as income level, credit score, interest rate on the account, etc.

Furthermore, with a longer timeline, another machine learning model, such as a KNN or a Random Forest, can be fitted to the data to achieve a potentially higher AUC.

METHODOLOGY and ANALYSIS

Data Used

The client provided data for 30,000 customers and 23 predictor variables, as well as a binary variable of customer default. Predictor variables included: gender, age, marital status, education level, history of payment amounts, historical repayment status, historical bill statements, and amount of credit given.

Potential Data Issues

An assumption was made that negative bill amounts can be attributed to customers paying extra on a specified amount due. However, to ensure accuracy, these data points should be investigated and verified for the validity of that assumption.

It would also be helpful to verify customers' education level coding, as customers with graduate degrees and university degrees have higher rates of default as compared to customers with high school degrees or 'others' (Table 2). This phenomenon is further illustrated by the odds ratio of 'others' vs. graduate degree (0.381), which suggests that customers with a graduate degree are about 2.6 times more likely to default as compared to customers with "other" degree type. Although this is possible, these results seem counter-intuitive, as higher degree attainment generally leads to higher incomes, and thus, anticipated payment ability. Therefore, further verification of the validity of the variable coding is necessary.

Additionally, better instruction needs to be provided regarding the calculation of each bill amount as compared to the payment amount, and the inclusion of calculation of interest rate on each account. The raw interest rate could be informative in the model building process as well.

Lastly, the client should ensure that all main effects variables are permitted to utilize to make decisions on future loans. More specifically, gender and age variables might violate the United States' non-discrimination laws, and therefore, might have to be omitted from a model that is used in the customer acceptance/rejection decision-making process.

Table 2: Default distribution by Education Attainment

Defaulted	Education	Frequency	Percent
No	Graduate School	8549	36.59
No	University	10700	45.8
No	High School	3680	15.75
No	Others	435	1.86
Yes	Graduate School	2036	30.68
Yes	University	3330	50.18
Yes	High School	1237	18.64
Yes	Others	33	0.5

Data Re-coding and Feature Engineering

During the exploratory stage, some inconsistencies in the data were found. Namely, education level included out of range values of 0, 5, 6, which were grouped with value 4, "other." Marital status also included a value of 0, which, again, was consolidated into "other" category. Also, the payment status variable had negative values, which were turned into 0 and grouped with "true" zeroes to indicate a positive payment history.

Several variables were also categorized—age, payment status, and credit line limit (please refer to Appendix Tables 2, 3, and 4). Each monthly payment was also summed into one value, Payment Amount, and only the sum version of the variable was used in the model building process. The reason for this decision was to achieve relative model simplicity while still capturing the potential importance of the predictor in the model.

In the next stage of the project, we can also address potential variable scale issues with standardizing some of the continuous variables (payment amount, bill amount) to improve model accuracy.

Correlations

During the exploratory stage, some predictor variables were found to be highly correlated. More specifically, Bill Amount variables. Due to high collinearity in these variables, only two of them were used in the modeling process, Bill Amount 4 and Bill Amount 6, to prevent further destabilization of the model coefficients. However, in the future, other methods can be applied to address the issue of multicollinearity, such as PCA or Lasso regression.

Descriptive Statistics

As outlined in Table 4, out of the group that did not default, females had a higher non-default rate as compared to males; however, there is also a higher representation of females as compared to males in the default group, and in the dataset overall (Table 3). Despite women constituting a higher proportion in the data, this customer group exhibits lower odds of default as compared to men customers (Appendix, Table 1).

Table 3: Gender Distribution

Gender	Frequency	Percent
Male	11888	39.63
Female	18112	60.37

Table 4: Default distribution by Gender

Defaulted	Gender	Frequency	Percent
No	Male	9015	38.59
No	Female	14349	61.41
Yes	Male	2873	43.29
Yes	Female	3763	56.71

Furthermore, similar percentages of married and single people defaulted on their payments, 48.32%, and 50.35%, respectively. However, the proportion of single non-defaulters is slightly higher in the dataset (Table 5).

Table 5: Default distribution by Marriage Status

Defaulted	Marriage	Frequency	Percent
No	Married	10453	44.74
No	Single	12623	54.03
No	Other	288	1.23
Yes	Married	3206	48.31
Yes	Single	3341	50.35
Yes	Other	89	1.34

Analysis/Model Building

The data was somewhat imbalanced, as it included ~ 22% percent of default accounts and ~78% of non-default accounts. However, considering that the rare event rate was at a reasonable 22% in the dataset, a re-sampling technique was not performed.

The data was split into training and validation datasets, with a 70/30 split. The training dataset was subsequently fitted with a logistic regression model via a stepwise selection method with a p-value of 0.03 and reference coding. As previously mentioned, the final model included eight main effect variables, achieving concordance of 76% on the training dataset. The final logistic regression model was then applied to the validation dataset (Table 6). The concordance on the validation dataset was 74%.

Table 6: Final Model Parameter Estimates

Parameter Estimates and Profile-Likelihood Confidence Intervals				
Parameter		Estimate	95% Confidence Limits	
Intercept		-1.4889	-1.8414	-1.1466
agec	26 to 30	-0.1759	-0.3413	-0.0092
agec	30 to 35	-0.142	-0.3069	0.0241
agec	35 to 40	-0.00275	-0.1667	0.1625
agec	40 to 50	0.0237	-0.1332	0.182
agec	greater than 60	0.2158	-0.1583	0.579
agec	le 25	-0.0366	-0.2165	0.1443
BILL_AMT4		1.28E-06	5.74E-07	1.97E-06
Education	2	0.00239	-0.0831	0.088
Education	3	-0.0537	-0.1687	0.0608
Education	4	-0.9646	-1.4511	-0.5321
Limit	100 to 200K	-0.1845	-0.2909	-0.0781
Limit	200 to 300K	-0.3066	-0.4377	-0.1762
Limit	300 to 400K	-0.4236	-0.5939	-0.256
Limit	400 to 500K	-0.3385	-0.5576	-0.125
Limit	50 to 100K	-0.0595	-0.1689	0.0496
Limit	500 to 600K	-0.2712	-0.9836	0.3624
Limit	600 to 700K	-0.8624	-2.3098	0.1964
Limit	700 to 800K	0.316	-1.2226	1.566
Limit	900K to 1 MIL	-3.5083	.	8.1297
Marriage	1	0.000962	-0.3062	0.3197
Marriage	2	-0.1886	-0.4993	0.1335
Pay_Amt		-4.96E-06	-6.28E-06	-3.71E-06
paycat	1 to 3 missed	1.1993	1.0973	1.301
paycat	3 to 6 missed	1.8001	1.6853	1.915
paycat	6 to 9 missed	2.0468	1.8925	2.2018
paycat	9 to 12 missed	2.5124	2.3507	2.6762
paycat	at least one missed payment	0.9418	0.7917	1.0897
paycat	more than 12 missed	2.7489	2.5152	2.9896
SEX	2	-0.1424	-0.217	-0.0677

CONCLUSION

During the initial modeling process of credit card customers' predicted probabilities of default, a logistic model was built, which succeeded in achieving AUC of 0.76 on training and 0.74 on validation datasets. Furthermore, the optimal cut-off point was determined to be 0.23; however, the lender is more equipped to determine the cost associated with a higher/lower cut-off threshold of predicted probabilities. Additionally, in the future, a more accurate model can potentially be achieved via addition of informative predictor variables, exploration of the accuracy of several already included predictor variables, or testing of several other models, such as a Random Forrest or a Decision Tree Model, as well as employing a standardization method to several continuous variables.

APPENDIX

Table 1: Odds Ratios for Main Effect

Effect	Odds Ratio Estimate
paycat more than 12 missed vs. positive payment history	15.625
paycat 9 to 12 missed vs. positive payment history	12.334
paycat 6 to 9 missed vs positive payment history	7.743
paycat 3 to 6 missed vs positive payment history	6.051
paycat 1 to 3 missed vs positive payment history	3.318
paycat at least one missed payment vs positive payment history	2.565
Limit 700 to 800K vs 10 to 50 K	1.372
agec greater than 60 vs 50 to 60	1.241
agec 40 to 50 vs 50 to 60	1.024
Education 2 vs 1	1.002
Marriage 1 vs 3	1.001
BILL_AMT4	1
Pay_Amt	1
agec 35 to 40 vs 50 to 60	0.997
agec le 25 vs 50 to 60	0.964
Education 3 vs 1	0.948
Limit 50 to 100K vs 10 to 50 K	0.942
agec 30 to 35 vs 50 to 60	0.868
SEX 2 vs 1	0.867
agec 26 to 30 vs 50 to 60	0.839
Limit 100 to 200K vs 10 to 50 K	0.832
Marriage 2 vs 3	0.828
Limit 500 to 600K vs 10 to 50 K	0.762
Limit 200 to 300K vs 10 to 50 K	0.736
Limit 400 to 500K vs 10 to 50 K	0.713
Limit 300 to 400K vs 10 to 50 K	0.655
Limit 600 to 700K vs 10 to 50 K	0.422
Education 4 vs 1	0.381
Limit 900K to 1 MIL vs 10 to 50 K	0.03

Table 2: Age Categories

Age Category	Frequency	Percent	Cumulative Frequency	Cumulative Percent
26 to 30	7142	23.81	7142	23.81
30 to 35	5796	19.32	12938	43.13
35 to 40	4917	16.39	17855	59.52
40 to 50	6005	20.02	23860	79.53
50 to 60	1997	6.66	25857	86.19
greater than 60	272	0.91	26129	87.1
le 25	3871	12.9	30000	100

Table 3: Credit Limit Categories

Credit Limit	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10 to 50 K	7676	25.59	7676	25.59
100 to 200K	7880	26.27	15556	51.85
200 to 300K	5059	16.86	20615	68.72
300 to 400K	2759	9.2	23374	77.91
400 to 500K	1598	5.33	24972	83.24
50 to 100K	4822	16.07	29794	99.31
500 to 600K	127	0.42	29921	99.74
600 to 700K	56	0.19	29977	99.92
700 to 800K	22	0.07	29999	100
900K to 1 MIL	1	0	30000	100

Table 4: Payment Status Categories (missed payments)

Payment Categories	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1 to 3 missed	3511	11.7	3511	11.7
3 to 6 missed	2144	7.15	5655	18.85
6 to 9 missed	1103	3.68	6758	22.53
9 to 12 missed	1110	3.7	7868	26.23
at least one missed payment	1666	5.55	9534	31.78
more than 12 missed	535	1.78	10069	33.56
positive payment history	19931	66.44	30000	100