IDS 702: Final Project
Predict the Default Probability of Credit Card Client

Ziming Huang

November 22, 2020

Motivation & Research Question

- Compared with traditional credit risk analysis, machine learning enable people to analysis large data and capture the nonlinear complex relationship between customer behavior and default probability.

- I will try to compare machine learning and traditional regression (logistic regression) in analyzing customer behavior based on their demographic characters and payment history to predict their probability of default for next month.

Data Description

- The data set includes the payment data in October, 2005, from an important bank in Taiwan. Among the total 23,150 credit card holders, 5,362 observations (23.16%) are cardholders with default payment.
- Card holder's demography information (age, marriage status, gender, and education level) and the history of past six month's payment information (bill statement, payment amount, and payment status) are used as explanatory variables.
- This research employed a binary variable – default payment (Default = 1, No Default = 0), as the response variable.

Data Exploration

Demography Feature

- Age: cardholders around 35 years old have the lowest default rate and young cardholders under 25 are more likely to default
- Gender: the average default rate for male and female cardholder are 25.25% and 21.79%
- Marriage: the average default rate for married and single cardholder are 24.76% and 21.75%
- Education: the average default rate for cardholder with graduate and high school degree are 19.55% and 25.86%

Male cardholder whose age is under 25 and with lower education level is more likely to default

Data Exploration

Payment History

- Payment Status: the average default rate increases with the number of payment delay month, this relationship still holds for the payment status 6 months ago (April, 2005)
- Bill statement: On average, defaulted cardholders have lower bill statement (the difference is significant)
- Payment amount: On average, defaulted cardholders pay about 44% less in the past six month (the difference is significant)

Payment history might be a good indicator for future default

Logistic Regression

Model Selection

- From EDA, no evidence that the interaction term should be included, so I try to include all these 23 variables
- However, I have the issue of multicollinearity of bill statement amount,I revise to include only September bill amount
- Step-wise method gives a similar model, my final model is:

$$logit(default_i) = \beta_0 + \beta_1 Age_i + \beta_2 Marriage_i + \beta_3 Educ_i + \beta_4 Female_i$$

$$+ \sum_{k=1}^{6} \lambda_k L_k(Payment_i^{Status}) + \sum_{k=1}^{6} \gamma_k L_k(Payment_i^{Amount}) + \delta L_1(Bill_i^{Statement}) + \epsilon_i$$

Model Assessment

- VIF: scores all below 3
- Residual plot: only few points (<3) outside the 95% band
- Outlier and influential point: 3 points are potential outliers, but result is robust after throwing them away

Logistic Regression

Table 1: Regression Result (partial)

Variable	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-1.0284	0.0999	-10.294	0.0000 ***
LIMIT_BAL	-0.0015	0.0002	-7.518	0.0000 ***
SEX2	-0.1272	0.0353	-3.601	0.0003 ***
EDUCATION2	-0.0821	0.0415	-1.980	0.0477 *
EDUCATION3	-0.1395	0.0550	-2.536	0.0112 *
EDUCATION4	-1.4946	0.5973	-2.502	0.0123 *
MARRIAGE2	-0.2310	0.0399	-5.786	0.0000 ***
AGE	0.0025	0.0021	1.176	0.2396
PAY_0	0.6749	0.0239	28.199	0.0000 ***
PAY_2	0.0859	0.0251	3.418	0.0006 ***
PAY_3	0.1029	0.0264	3.896	0.0001 ***
PAY_4	0.0521	0.0291	1.791	0.0732 .
PAY_5	0.0781	0.0312	2.505	0.0122 *
PAY_6	0.0913	0.0266	3.434	0.0006 ***
BILL_AMT1	-0.0008	0.0003	-2.778	0.0055 **
PAY_AMT1	-0.0064	0.0020	-3.225	0.0013 **
PAY_AMT2	-0.0061	0.0020	-3.099	0.0019 **
Accuracy Rate:	0.7718			
Sensitivity:	0.6084	Specificity:	0.8211	(thres.mean)
Area under curve:	0.7492	(thres.best)		

Compare with Machine Learning Algorithm

Machine Learning Algorithm

- Random forest, Decision Tree, and Gaussian Bayesian
- Separate data into train and test sub-sample with split rate 0.25
- Repeat training for 10 times

Table 2: Model Prediction Result

	Sensitivity	Specificity	Accuracy_{In}	Accuracy_{out}	AUC
Random Forest	0.8277	0.6155	0.7780	0.7927	0.7336
Decision Tree	0.9999	1	0.9999	0.7319	0.6276
Gaussian Bayesian	0.7597	0.6701	0.7389	0.7425	0.7762
Logistic Regression	0.5015	0.6228	0.8241	0.7586	0.7509

Result Comparison

- Random forest has the highest out-of-sample accuracy, but this model might suffer under-fitting problem
- Decision tree performs the best for in sample test, but the model might suffer over-fitting problem
- Overall, linear logistic regression still has a good performance

Conclusion

Most variables in my data show predict power

- Demography
  - Male, lower education level, and married cardholders are riskier
  - Age plays no role in predicting default
- Account Information
  - Cardholder with higher amount of the given credit is less riskier
  - History payment status is important for default prediction, cardholders who delayed pay previous bill are more likely to default

Algorithm Comparison

- None of these three machine learning algorithms significantly outperforms the simple logistic regression
- When apply complex algorithm, more things need to consideration, such as avoid overfitting and under-fitting