Predict the Default Probability of Credit Card Client

Ziming Huang, November 22, 2020

Summary

This project compares logistic regression and other machine learning algorithms in analyzing credit card client's next month's default probability based on their demographic features and payment history over the past six months. Specifically, the project uses the default of credit card clients data in Taiwan and compared the logistic regression predictive accuracy with four machine learning methods. According to the model result, the client's default probability varies among different demographic features including gender, marriage status, and education level. Married male clients with lower education levels are expected to default at a higher probability. As for the algorithm performance, none of these algorithms outperforms the other four from all aspects.

1 Introduction

With larger data and more computing power becoming available, machine learning is possibly being applied in credit risk analysis. Compared with traditional credit risk analysis, machine learning enables people to analyze large data and capture the non-linear complex relationship between customer behavior and default probability. Some major analysis institutions, such as Moody's as well as Morgan Stanley have applied machine learning to create a credit risk model to help their analysts make decisions.

However, it is unwise to conclude that these complicated machine learning algorithms must outperform simple methods, such as logistic regression, in analyzing a client's default probability. Sometimes complicated algorithms might cost more time to deal with model problems and also make it difficult even impossible to do the model interpretation. Under this background, it will be an interesting problem to think of the trade-off of using a machine learning algorithm. Therefore, this project focuses on predicting the default probability of credit cardholders by using the card holder's demographic feature and their payment history over the past six months. Specifically, the demographic attributes including the client's age, gender, marriage status, and education level. Payment history data includes payment status, bill statement, and previous payment amount. The data set this project uses includes a total of 23,150 credit card holders from a major bank in Taiwan, China. Data set can be obtained from UCL Machine Learning Repository[1]. The rest of the report will follow this structure: Section 2 includes data description and exploratory and section 3 is the logistic model specification and machine learning algorithm comparison. The last section concludes.

2 Data

The project uses the default of credit card clients' data set in Taiwan from the UCL Machine Learning Repository. This data set includes 23,150 clients and total of 24 attributes. Among the total 23,150 credit card holders, 5,362 observations (23.16%) are cardholders with default payment.

[1]Default of credit card clients Data Set: http://archive.ics.uci.edu/ml/datasets

2.1 Data Description

This data set includes the client's gender, age, marital status, and education level. Besides, the data set also provides the client's past six months (from April 2005 to September 2005) account bill statement, payment delay status, previous payment amount, and initially given credit. This data set also includes an extra column to indicate whether this client defaults for the next month. Data variables definitions are shown in table 3 in appendix. For the rest of the paper, this binary-value variable is the model response, and models are applied to predict the client's default status based on the client's attributes.

2.2 Data Exploratory

To further explore the data set, a summary table of these key attributes is provided. According to this summary data, the client's age has a wide range from 21 years old to 79 years old. Besides age, this table provides some information on the client's payment record over the past two months. Same as age, both the previous payment amount and bill statement have a large range and standard deviation, indicating that this data set has a diversified sample.

The next step compares the client's demographic feature and payment record between the default group and the non-default group. Consider the default probability at different ages first. According to the data set, for clients whose age is between 25 years old to 60 years old, the default probability is within the range of 20% to 25%. However, for younger clients whose age is around 20, the default probability increases to about 30%. For clients whose age larger than 60 years old, the default probability varies in a larger range. This variation might comes from the lacking of observations for these age groups. For gender, the probability of default for male clients is around 25.15%, which is 3.3% higher than female clients. A married client has a 2.5% higher probability of default. The difference of default probability becomes larger when it comes to the client's education level: the default probability for clients with high school education level is 25.86%, which is about 6% higher than clients with graduate education level.

Besides the demographic feature, the client's payment record also provides useful information in prediction default probability. Take the payment status as an example. For clients who default in October, the average month that these clients delay pay is about one month, while the data for the non-default client is zero. The average initial given credit for non-default clients is 168426.2 NT dollars, while the average credit given for default clients is 115928.3 NT dollars. However, the billing statement and previous payment amount do not show a significant difference between the default and non-default groups.

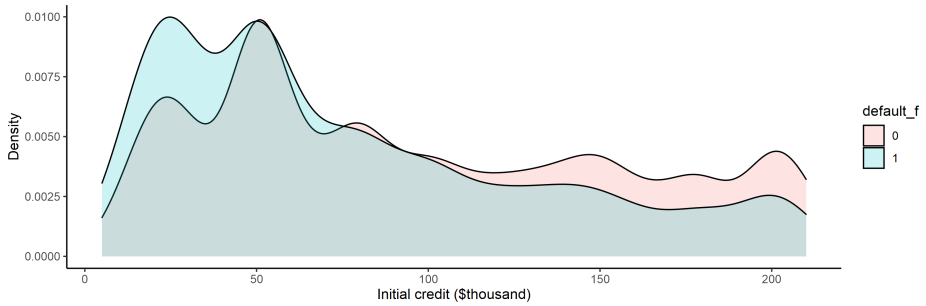

Figure 1: Age vs Default probability

3 Model

In this section, logistic regression and other four machine learning algorithms are applied to analyze the relationship between the client's attributes and default probability. Specifically, model performance will be graded based on their prediction accuracy.

3.1 Logistic Regression

According to data exploratory analysis, no strong evidence shows that the model should include interaction terms among these 24 attributes, thus only the main effect variables are considered being included in the model. Since the payment status variable in this data set does not contain the value 9, so these variables will be used as numeric variables instead.

The initial model includes all 23 attributes (exclude the default indicator). However, when computes the variation inflation score (VIF), the scores for bill statement variables are larger than 16, indicating that the model might have the multicollinearity problem with these six variables. To address this issue and include useful information as much as possible, the bill statement variables for April 2005 to August 2005 are removed from the model, that is, only the bill statement amount of September 2005 is included in the model. After dropping problematic variables, the VIF score for all variables are below 3, thus the multicollinearity problem might be solved and the baseline model for logistic regression is:

$$\begin{aligned}\text{logit}(\text{default}_\text{i}) = {} & \beta_0 + \beta_1 \text{Age}_\text{i} + \beta_2 \text{Marriage}_\text{i} + \beta_3 \text{Educ}_\text{i} + \beta_4 \text{Female}_\text{i} + \beta_5 \text{Init\_credit} \\ & + \sum_{\text{k}=1}^{6} \lambda_\text{k} \text{L}_\text{k}(\text{Payment}_\text{i}^{\text{Status}}) + \sum_{\text{k}=1}^{6} \gamma_\text{k} \text{L}_\text{k}(\text{Payment}_\text{i}^{\text{Amount}}) + \delta \text{L}_1(\text{Bill}_\text{i}^{\text{Stataement}}) + \epsilon_\text{i}\end{aligned}$$

where $\text{L}(\cdot)$ is the lag operation and $\text{L}_\text{k}(\cdot)$ is lagged k term.

To check if there are any redundant variables in this baseline model, a step-wise method is used to do further model selection work. The models are given by step-wise with AIC and BIC are[2]:

$$\begin{aligned}(\text{AIC})\text{logit}(\text{default}_\text{i}) = {} & \beta_0 + \beta_1 \text{Marriage}_\text{i} + \beta_2 \text{Educ}_\text{i} + \beta_3 \text{Female}_\text{i} + \beta_4 \text{Init\_credit} \\ & + \sum_{\text{k}=1}^{6} \lambda_\text{k} \text{L}_\text{k}(\text{Payment}_\text{i}^{\text{Status}}) + \sum_{\text{k}=1}^{2} \gamma_\text{k} \text{L}_\text{k}(\text{Payment}_\text{i}^{\text{Amount}}) + \delta \text{L}_1(\text{Bill}_\text{i}^{\text{Stataement}}) + \epsilon_\text{i} \\ (\text{BIC})\text{logit}(\text{default}_\text{i}) = {} & \beta_0 + \beta_1 \text{Marriage}_\text{i} + \beta_2 \text{Female}_\text{i} + \beta_3 \text{Init\_credit} \\ & + \sum_{\text{k}=1,3,5,6} \lambda_\text{k} \text{L}_\text{k}(\text{Payment}_\text{i}^{\text{Status}}) + \sum_{\text{k}=1}^{2} \gamma_\text{k} \text{L}_\text{k}(\text{Payment}_\text{i}^{\text{Amount}}) + \epsilon_\text{i}\end{aligned}$$

To decide the final model, an ANOVA test is applied. The test result is shown in the appendix. According to the test result, the model selected by AIC criteria is the most promising one, therefore, the rest of the report will focus on AIC model. The final model results are shown in table 1.

For model assessment, the model bin residual plot is used to investigate potential outliers and how the model fits the data. First, the residuals against predicted probabilities and age are generally random and within the 95% (with only two points outside the band). The VIF score for all variables is below 3, indicating no serious multicollinearity problem. The overall accuracy, sensitivity, and specificity are 77.29%, 60.82%, and 82.25% respectively (threshold=0.2316). The AUC value is 0.749, demonstrating the effectiveness of the final model to some degree (Figure 2).

[2]Small changes have been made to avoid multicollinearity problem

Table 1: Logistic regression result

Variable	Estimate	Std. Error	z	value Pr(>|z|)
(Intercept)	-0.9336	0.0557	-16.755	< 2e-16 ***
PAY_1	0.6750	0.0239	28.206	< 2e-16 ***
Init_credit	-0.0015	0.0002	-7.829	4.92e-15 ***
PAY_3	0.1022	0.0264	3.870	0.0001 ***
PAY_5	0.0804	0.0306	2.623	0.0087 **
Marriage2	-0.2523	0.0359	-7.018	2.25e-12 ***
Marriage3	-0.0085	0.1517	-0.056	0.9554
PAY_AMT1	-0.0066	0.0020	-3.355	0.0008 ***
female	-0.1327	0.0349	-3.798	0.0001 ***
Educ2	-0.0816	0.0415	-1.968	0.0490 *
Educ3	-0.1262	0.0539	-2.340	0.0193 *
Educ4	-1.500	0.5971	-2.512	0.0120 *
PAY_AMT2	-0.0064	0.0020	-3.222	0.0013 **
PAY_6	0.0913	0.0263	3.474	0.0005 ***
PAY_2	0.0853	0.0251	3.395	0.0007 ***
BILL_AMT1	-0.0009	0.0003	-2.961	0.0031 **
PAY_4	0.0534	0.0287	1.863	0.0625 .

Model interpretation:
(intercept): For a married male client with a graduate education level, the average odds of default is 0.3931, keeping other variables the same.
Init_credit: For a client with an additional one thousand increments for the initially given credit, the average odds of default will decrease 0.0015 ($1 - e^{-0.0015}$), keeping other variables constant.
PAY_1 (Sept.): For clients who delays paying September's bill for one additional month, the average odds of default will increase 0.9640 ($e^{-0.6750} - 1$), keeping other variables constant.
Educ2(university): For a client whose education level is a university, his/her average odds of default will increase 0.0784 ($1 - e^{-0.082}$) compared to clients whose education level is graduate school, keeping other variables constant.

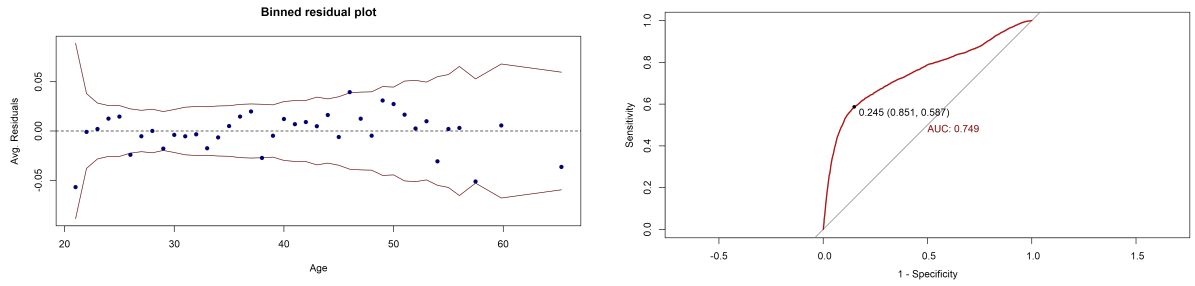

Figure 2: Model Assessment

3.2 Machine Learning Algorithm

Four machine algorithms are applied to compare the predictive ability with logistic regression. These machine algorithms are random forest, decision Tree, and Gaussian Bayesian. Specifically, K-Nearest Neighbors uses 5 neighbors as a parameter as it performs best in many trials. The

original data set is separated into train and test sub-sample with a split rate of 0.25. All models are repeat training 10 times for model cross-validation. Models results are shown in table 2.

Table 2: Model prediction performance

	Sensitivity	Specificity	$\text{Accuracy}_{\text{In}}$	$\text{Accuracy}_{\text{out}}$	AUC
Random Forest	0.8275	0.6213	0.7780	0.7927	0.7336
Decision Tree	0.8062	0.4550	0.9999	0.7319	0.6276
K-Nearest Neighbors	0.8101	0.7091	0.7865	0.6707	0.5973
Gaussian Bayesian	0.7623	0.6553	0.7389	0.7425	0.7762
Logistic Regression	0.5015	0.6228	0.8241	0.7586	0.7509

According to the model prediction performance table, the random forest has the largest sensitivity and K-Nearest Neighbors has the largest specificity, which one is more important depends on the preference of the problem. The random forest has the highest out-of-sample accuracy, but this model might suffer an under-fitting problem.The decision tree has the largest in-sample accuracy, indicating that this model performs the best for the in-sample test, however, the large difference between in-sample and out-of-sample accuracy shows that this model has the over-fitting problem.

4 Conclusion

According to the logistic model result, most of the client's attributes are statistically significant at a 5% significant level, indicating that including these attributes into the model can help improve the predicting power of future default. Specifically: (1) for demographic features: a male married client with lower education level are riskier and with higher expected future default probability, however, age seems to play no role in predicting default; (2) for payment history record: a client with higher amount of the initially given credit (Init_creadit) is less risky. Besides, history payment status is important for default prediction, a client who delayed pay his/her previous bill is more likely to default.

When doing the performance comparisons among logistics regression and the other four machine learning algorithms, those complicated machine learning models might suffer the over-fitting and under-fitting problems and the simple logistics regression model provided a satisfactory result. According to the comparison result, it is worth thinking of whether a more complicated model should replace the simple one. In fact, a simple model might work probably better than expected and it can help judge whether a complex model is even justified. Unless a complicated model can provide large enough improvement in performance, simple models should always be taken into consideration.

However, this project has some limitations. First, the data set is imbalanced. This data set does not include observations of clients whose age is larger than 60 years old, with this limitation, the model is not able to give reliable statistical inference to these client groups. Second, other variables that might impact the client's default probability are not included. The alternative data set that includes a real GDP growth rate, GNP, and gross national income . Given that the main data set only contain six-month payment history data, including these low frequency-updated data does not have enough explained power. With these limitations, further study will think of including more observations from enough diversified group and macroeconomic indexes with higher updated frequency.

Appendix

Table 3: Variable Definition

Variable	Definition	Notation
Default(binary)	Indicator of default status for October, 2009 (Default = 1, No Default = 0)	defualt
Age(discrete)	Client's age	Age
Marriage(category)	Client's marriage status (1=married; 2: single)	Marriage
Gender(binary)	Client's gender (1: male, 2:female)	Gender (male/female)
Education(category)	Client's education level (1= graduate school; 2=university; 3=high school; 4=others)	Educ
Payment Status(category)	Client's history payment status from April to September, 2005 (-1: pay one month advanced, 0: pay duly; 1: payment delay for one month;...; 8: payment delay for eight months; 9: payment delay for nice month and above	PAY_1 (Sept.), $\cdots$, PAY_6 (Apri.)
Previous payment amount (continuous)	Amount of previous payment from April to September, 2005 (NT thousand dollar)	PAY_AMT1 (Sept.) $\cdots$, PAY_AMT6 (Apri.)
Bill statement (continuous)	Amount of bill statement from April to September, 2005 (NT thousand dollar)	BILL_AMT1, $\cdots$, BILL_AMT6 (Apri.)
Credit given (continuous)	Amount of the given credit, this amount includes both the individual consumer credit and his/her family (supplementary) credit (NT thousand dollar)	Init_credit

Table 4: Variable Statistic Summary

Variable	Mean	Min.	Max.	Std.
AGE	35.24	21.00	79.00	9.291339
LIMIT_BAL	156.3	10.0	1000.0	127.5823
BILL_AMT1	61.298	-165.580	964.511	77.6107
BILL_AMT2	32.77	-67.53	983.93	75.2012
PAY_AMT1	6.075	0.000	873.552	16.90432
PAY_AMT2	6.119	0.000	1227.082	20.30216

Table 5: ANOVA test for model selection

Model	p-value	Favor Model
Baseline vs AIC	0.598	AIC model
Baseline vs BIC	0.000	Baseline
AIC vs BIC	0.000	AIC model