STAT 538: PREDICT CREDIT CARD DEFAULTS

SHUYI TAN

1. **Introduction**

During the pandemic, many people lost their financial sources because of unemployment. As a result, some people are strapping for cash and defaulting on their credit card payments by continuously applying for new credit cards. This imposes severe risk on banks and credit card companies. Therefore, in this project, I am going to evaluate whether a credit card applicant will be able to pay the bill based on their demographic characteristic and historical bank data including past payment and bill amount. To classify applicants into two classes: defaulters and non-defaulters, two models: logistic regression and Naive Bayes classifier will be utilized in this project.

2. **Data Overview**

The data I used is the ”default of credit card clients dataset” from the UCI Machine Learning Repository[1]. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. There are 30,000 observations in the dataset with 25 variables. A binary variable “default” will be used as the response variable, in which 1 represents default, and 0 represents not default. For the explanatory variables, they are displayed in the following table:

Explanatory Vairables			
Variable Name	Description	Type	Levels
SEX	Gender	Factor	1 = Male 2 = Female
EDUCATION	Educated Level	Factor	1 = Grad School 2 = College 3 = High School 4 = Others
MARRIAGE	Marital Status	Factor	0 = Unknown 1 = Married 2 = Single 3 = Others
AGE	Measured in Year	Numeric	N/A
LIMIT_BAL	Amount of Given Credit	Numeric	N/A
PAY0,...,PAY6	# of Months Payment Delayed	Numeric	N/A
PAID1,...,PAID6	Percentage of Bill Paid Per Month	Numeric	N/A

Date: December 04, 2020.

3. Exploratory Data Analysis

3.1. **Missing Data Analysis.** Checking each column, it is great to see that there is not any NA/Nan/Inf data existing in the data. However, with levels $1, 2, 3$ being specified as "married","single", and "others" respectively, the level "0" was not mentioned by the data author. There are 323 "0" values in the variable marriage. We cannot recklessly replace the "0"s with a pre-specified marriage level because of insufficient knowledge of credit card applicants. Therefore, I am going to set up a new level: "unknown" to represents the "0"s in the marriage variable.

3.2. **Demographic Variables.** We start by examining the distribution of the our response variable, the response variable is what we are asked to predict: either a 0 for the bill was paid on time, or a 1 indicating the applicant had payment difficulties. We can first check the number of cases falling into each category:

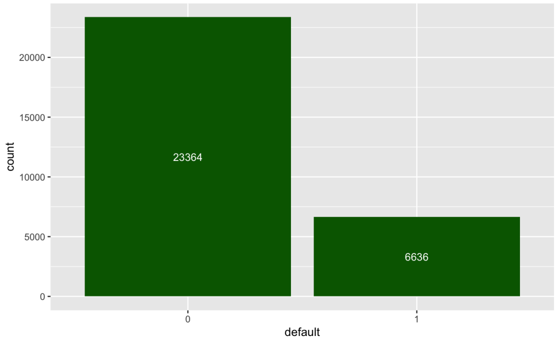

Figure 1

We can see that there are more default cases than not default cases. In order to take a closer look at our predictors, I am going to a conduct bivariate analysis of demographic variables with the response variable. Integrating the proportion of defaults within each level of categorical variables, there are some interesting insights from the plots in figure 2. In terms of gender, We can see that females have a slightly lower proportion of defaults than males. As for education, within the

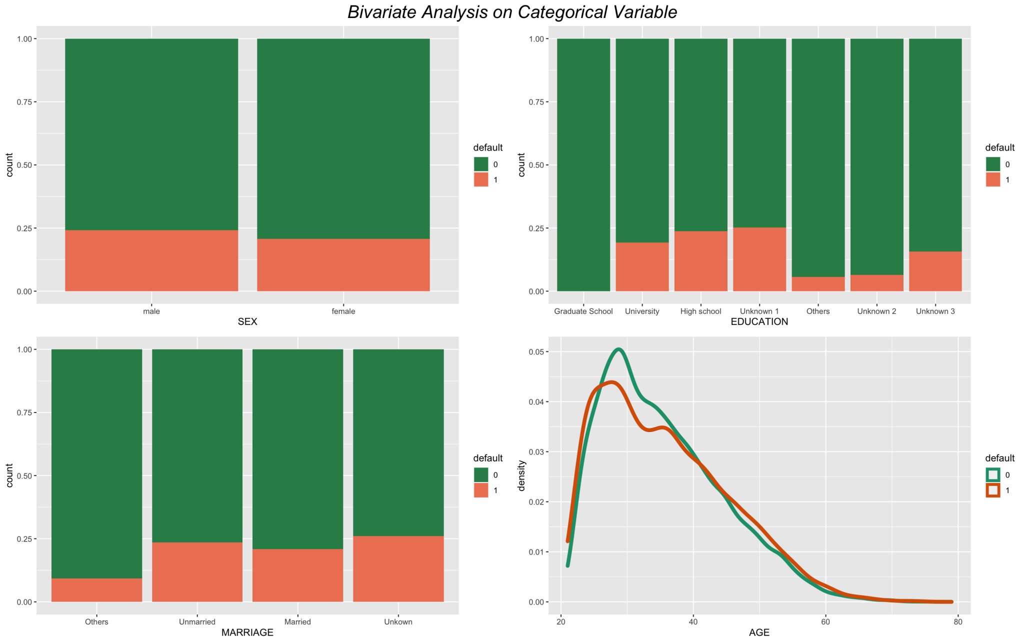

Figure 2

known levels, it seems that people are less likely to default as they are getting more educated. It is understandable because usually people are paid more with a higher education degree. When it come to marital status, comparing unmarried and married applicants, fewer married people defaults. For age, the distributions are almost uniformly distributed but a little bit right-skewed for the defaulters. It reveals that old age people are almost non-defaulters. At this point, we may roughly infer that these demographic variables have some effects on the response variables. But this is to be verified in the modeling section.

3.3. **Numeric Variables.** We have a great amount of numeric variable, following the same logic as that of demographic variables, we visualize the distribution of the amount of the given credit in NT dollar as well as the percentage an applicant paid for his/her bill in the past 6 months. The plot on the left-hand-side shows displays the density of applications' given amounts of credit separated by default. We observe that more defaulters have been given credits that are lower than $125,000$ NT dollars, and non-defaults are more likely to be given a higher amount of credits, which is reasonable. The plot on the right-hand side shows the distribution of the log-transformed percentage of the paid amount per month, which is calculated by diving the monthly paid amount by the bill amount. It can be observed from the box plot that few people pay in full for the bill, but non-defaulters tend to pay more than the defaulter.

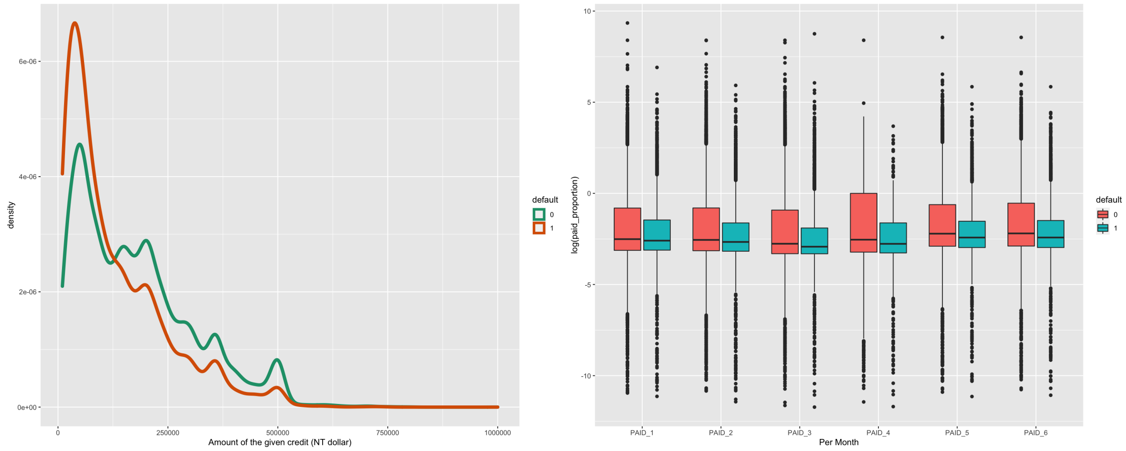

Figure 3

3.4. **Initial Feature Selection.** Before we dive into the modelling part, we would like to check the collinearity of our numeric predictors. From the heatmap, we can see that there are severe collinearity among variables $PAY_0...PAY_6$, $PAID_3$, and $PAID_4$. Generally, we would like to remove variables with an absolute correlation higher than 0.75. Setting 0.75 as the cutoff, we obtained 4 variables with high correlation that exceeds the cut off, which are PAY_4,PAY_5,PAY_3, and $PAID_4$. These variables will be removed from the list of predictors. Now we will start the modelling.

4. **Logistic Regression**

The objective of my project is to predict whether or not a credit card applicant will be a defaulter based on his demographic information and past payment record. Therefore, I would like to choose a logistic regression model as the baseline model. The idea of a logistic regression model is to model the probabilities for classification problems with two possible outcomes. Unsure about what combination of predictors will have better performance, we will perform model selection with various criteria on a set of candidate logistic regression models.

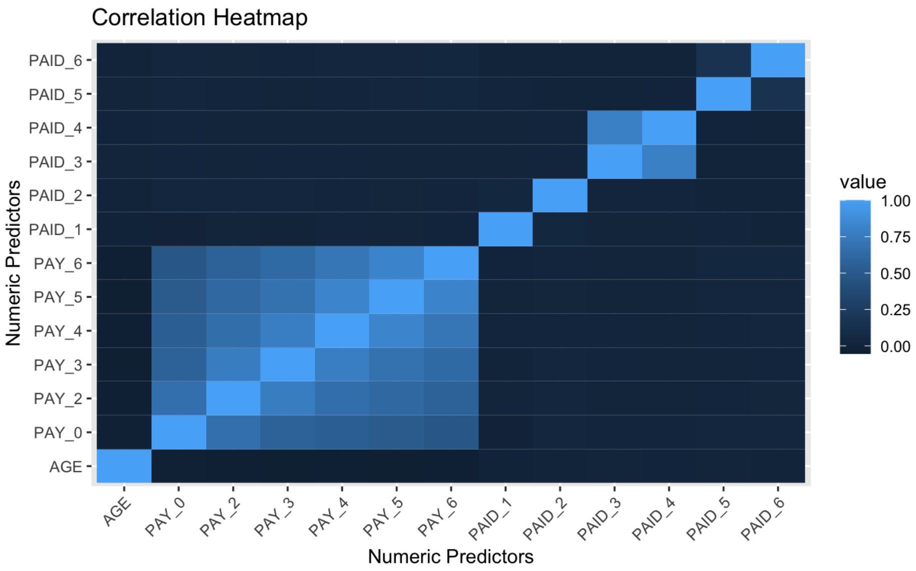

Figure 4

4.1. **Model Selection.** A general multiple logistic regression model can be expressed as :

$$ln(\frac{\hat{p}}{(1-\hat{p}})=b_0+b_1X_1+...+b_pX_p$$

where $\hat{p}$ is the expected probability that the outcome is present. We will start by setting up a full model with all the predictor we have. A full model will have 13 predictors:

LIMIT-BAL	SEX	EDUCATION
MARRIAGE	AGE	PAY-0
PAY-2	PAID-3	PAID-5
PAID-6		

The methods that we will use for model selection are forward selection and backward selection. The forward approach uses a sequence of steps to allow features to enter or leave the regression model one-at-a-time. [2], and the backward approach begins with a full models and starts removing the least significant variables one after the other until a pre-specified stopping rule is reached. For the stopping rule, we will use AIC and BIC respectively as the selection criteria.

Randomly splitting the data into a training set and test set, it's necessary to check whether the distribution of the training set and setting is similar to that of the whole data. A simple measure specifically for binary response variable is to check the ratio of either level through some proportion tables:

Proportion Table

	0	1
Full Data	0.7788	0.2212
Training Set	0.7772889	0.2227111
Test Set	0.7833333	0.2166667

We can see that the ratios of two levels in sampling sets are similar to that of the full data. Therefore, we can trustingly train our model on the train set and preform predictions on the test set. After performing model selection using two methods with AIC/BIC as stopping criteria, we obtained four models. In additional to AIC and BIC, the classification accuracies are also provided, which is the proportion of correctly classified responses based on the threshold of probabilities 0.5.

Another measure is the AUC, and it calculates the area under an ROC Curve. The ROC plots of logistic regression models are provided in section 6.

4.2. **Model Comparison.** Below it is a comparison table displaying different measures of these model:

	# of Predictors	AIC	BIC	Accuracy	AUC
Full Model	20	21263.26	21263.26	0.8113	0.6059
AIC Forward	16	21088.56	21224.92	0.8116	0.6059
BIC Forward	7	21141.31	21205.48	0.8107	0.6059
AIC Backward	16	21088.56	21224.92	0.8116	0.6059
BIC Backward	7	21141.31	21205.48	0.8107	0.6059

There are a number of insights from this table. Firstly, comparing the number of predictors in each model, we can see that models that are selected by BIC have far fewer predictors than those selected by AIC. This is reasonable because BIC penalizes model complexity more heavily, which results in the outcome that AIC has a preference for a larger model than BIC. Comparing the predictive accuracy of each model, actually, they are very close, but the models selected by BIC are slightly better. If we take account of the computational cost, the third and fourth models have the highest accuracy while with the lowest computational cost. Besides, fewer predictors can also effectively reduce the risk the overfitting. As for the AUC value, they are identical for each model. In summary, using the logistic regression model selected by the BIC criterion (in either direction) will be a relatively proper choice if we decide on using logistic regression to predict the default.

5. **Naive Bayes Classifier**

In this section, we are going to use the Naive Bayes Classifier to predict whether a credit application will default. Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The algorithm depends on Bayes' Theorem, which is stated as [3]

$$p(h|d) = (P(d|h) * P(h))/P(d)$$

where

- $P(h|d)$ is the probability of hypothesis h given the data d. This is called the posterior probability.
- $P(d|h)$ is the probability of data d given that the hypothesis h was true.
- $P(h)$ is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- $P(d)$ is the probability of the data (regardless of the hypothesis).

we are interested in obtaining the posterior probability of $P(h|d)$ from the prior probability $p(h)$ with $P(D)$ and $P(d|h)$. When applying the Bayes' Theorem to the our classification problem, the idea is to calculate the values of each attribute value e.g. $P(d1, d2, d3|h)$, in which they are assumed to be conditionally independent given the target value and calculated as $P(d1|h) * P(d2|H)$ and so on. Before we training a Naive Bayes Model on the training data, we will first conduct feature selection on the predictors to acquire an optimal set.

5.1. **Feature Selection.** The method we used for feature selection is the Recursive feature elimination (RFE), which offers a rigorous way to determine the important variable. Using Naive Bayes model and 2-fold cross-validation to generate a control object, the RFE algorithm returns an optimal set of predictors with the highest accuracy in the training data. Plotting accuracy against the number of predictors, you can observe from the figure 5 that there is an “elbow” in the accuracy when there are five predictors, which mean the model have the “best” performance here. These optimal predictors includes:

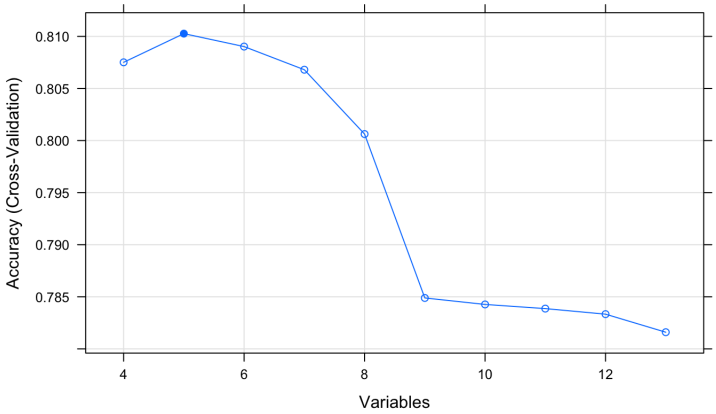

FIGURE 5

Predictors	Description
PAY_0	The repayment status of current month
PAY_2	The repayment status 2 months ago
LIMIT_BAL	Amount of the given credit (NT dollar)
PAY_6	The repayment status 6 months ago
PAID_5	The paid proportion of bill 5 months ago

It is surprising to see that all the demographic variables are removed from the model, and only historical bank data are utilized.

5.2. **Model Fitting.** Now we fit a Naive Bayes model to the training data and perform predictions on the test set, we may do an initial evaluation of the model by looking at the confusion matrix:

	Reference	
Prediction	0	1
0	5078	812
1	762	848

and obtain the Accuracy $= (TP + TN)/total = 79.01$. If we visualize the prediction results by a ROC plot shown below: we will see that indeed the model have a fair performance with an area

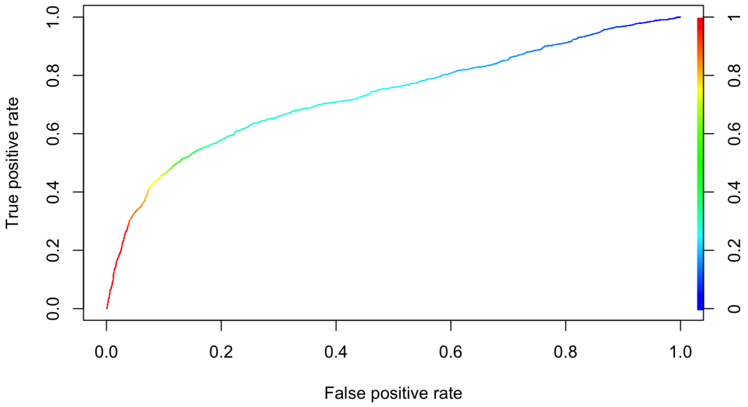

FIGURE 6

	Reference
.90-1	excellent (A)
.80-.90	good (B)
.70-.80	fair (C)
.60-.70	poor (D)
.50-.60	fail (F)

under an ROC curve (AUC) of 0.7297127 based on traditional the diagnostic rule above.[4]

6. **Comparing Logistic Regression and Naive Bayes Classifier**

Comparing the performances of logistic regression and naive Bayes Classifier, if we simply look at the classification accuracy, it seems that the logistic regression models slightly outperform the Naive Bayes Classifier. However, there is a limitation of this measure. Since my data does not have an even number of classes in the response variable “default”, we may get high accuracy. However, this is not a good score if most records for every 100 belong to one class, and I can achieve the high accuracy by always predicting the most common level value. Therefore, in our situation, we can instead pay attention to the AUC values of two kinds of models, and we can observe that the AUC of Naive Bayes Classifier is higher than that of logistic regression models. Furthermore, the performances of logistic regression models will be judged as ”poor” based on the traditional diagnostic rule (69.97), which can be verified by the ROC plots of logistic regression below: It can

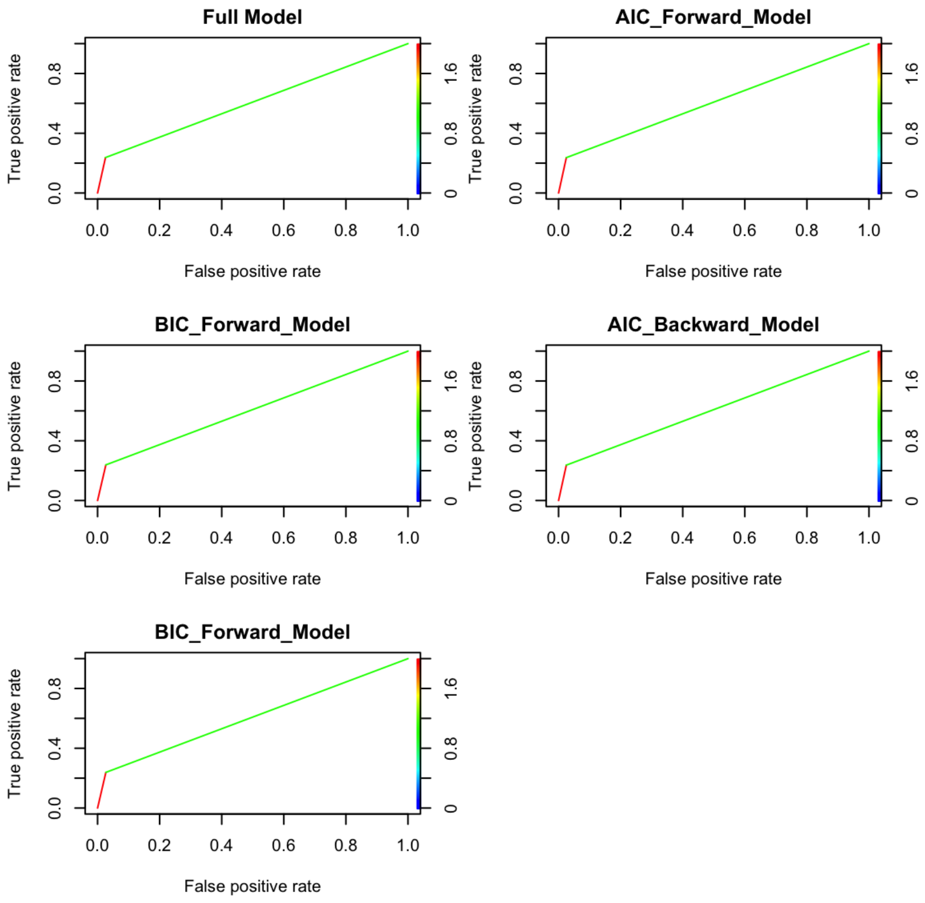

Figure 7

be seen that the ROC curves are far from left-hand border and the top border, and they are close the curve comes to the 45-degree diagonal, which indicates bad performance in this data.

Looking at the predictors of two kinds of models, an interesting phenomenon is that the logistic regression all values the significance of demographic variables like age, gender, and marriage, while the Naive Bayes use only some of historical bank records to reach close classification accuracy and higher AUC value. It's an enlightening view to the banks and companies, because if they adopt the Naive Bayes Classifier, using only a few of their internal data can help to accomplish the predictive tasks without the need to collect demographic data.

Theoretically, referring to a paper written by Professor Andrew Ng and Professor Michael I Jordan, they gave an insightful conclusion: "when the training size reaches infinity, logistic regression (discriminative model) performs better than the generative model Naive Bayes."[**?**, **?**] We did not see stand-out performances of logistic regression models perhaps because of our size of the training set. Big banks and credit card companies need to consider which model to adopt based on how much data they have in hand.

7. Conclusion and Further Improvement

In this analysis, logistic regression models selected by different methods and a Naive Bayes Classifier are used to predict the eligibility of credit card application through the probabilities of default. Different conclusions are acquired: for logistic regression, predicting default need the combination of demographic data and historical data; while for the Naive Bayes Classifier, using a few of the payment-related variables is sufficient. When it comes to their similarity, some variables like the given credit, recently paid percentage, and repayment status play important roles in both of these models. Considering their performance, our current results indicate that the Naive Bayes Classifier has better predictive accuracy measure by the AUC value. While this is not a consolidated conclusion, because the problem of imbalanced classes has not been solved in this project. As we discussed above, it may lead to a misleading outcome that we achieve the high accuracy by always predicting the most common level value. Some potential techniques will be used to tackle this problem, such as random Oversampling, random undersampling, and Synthetic Minority Oversampling Technique (SMOTE).

References

[1] Yeh,I-Cheng UCI Machine Learning Repository.2016. http://gim.unmc.edu/dxtests/roc3.htm

[2] Kuhn, Max. Johnson,Kjell. *Feature Engineering and Selection: A Practical Approach for Predictive Models.* Section 11.4. bookdown.org. 2019.

[3] Brownlee, Jason. *Naive Bayes for Machine Learning.* Machine Learning Mastery. 2019. https://machinelearningmastery.com/naive-bayes-for-machine-learning/

[4] Tape, Thomas G. *Interpreting Diagnostic Tests.* The University of Nebraska Medical Center. http://gim.unmc.edu/dxtests/roc3.htm

[5] Ng, Andrew. Jordan, Michael I. *On Descriminative vs. Generactive classifiers: A comparison of logistic regression and naive Bayes.* Standford University. http://ai.stanford.edu/ ang/papers/nips01-discriminativegenerative.pdf

Department of Statistics, University of British Columnbia