ECN190 Term Project: Predicting Credit Card Default Risk

Jatin Mohanty

3/14/2020

## Introduction and Literature

**Question: Can we predict credit card payment default?**

This is an important question especially regarding finance. With the impending recession as a response to the Coronavirus, it is becoming more important to know who is reliable customer and who is not[Wall Street Journal, 2020].Banks use this kind of analysis to determine whether to approve someone for a credit card or loan[Asia Pacific University of Innovation and Technology, 2019]. In this report, only logistic regression is used. However, there are many machine learning methods being utilized such as KNN, SVM, and decision trees[International Journal of Scientific Research, 2018]. This type of analysis can range farther than just predicting default risk for bank and credit card loans. A similar type of analysis could be used for analysis of risk for patients and consumers in healthcare and insurance companies as well. It is also important to get this type of analysis to be correct so people are not accepted when they should not be, possibly leading to worse lives[Hartman, 2018]. Or rejecting them when the should be accepted, having them take their business to other companies. An additional benefit to these analysis is that more responsible lending could lead to a softened impact on people and economies during economic downturns as less people default on loans and are able to endure during recession[Pruden, 2019]

## Data

This dataset named “Default of Credit Card Clients Data Set” was taken from the UCI Machine Learning Repository. The data set was donated by the Department of Information Management at Chung Hua University, Taiwan as well as the Department of Civil Engineering at Tamkang University, Taiwan. This data set shows the bank customers credit card payments in Taiwan as well as basic information about each customer(detailed later). There are 30,001 instances with 24 beginning variables. Some limitations to this data set are that for one, it does not completely encapsulate the spending and saving habits of each customer. Additionally, the history of payment is relatively short (6 months) which can be a hinderance when trying to determine a reliable customer. Lastly, this data is time series, so for the sake of this project, it must be transformed to be static data. While time series in itself is very interesting, it is not what I am looking to observe for this project. Lastly, the payment cycles recorded are from 2005, which while still useful, are not as useful as a more recent statement would be. Below is a brief view of the original dataset.

## LIMIT\_BAL SEX EDUCATION MARRIAGE AGE PAY\_0 PAY\_2 PAY\_3 PAY\_4 PAY\_5 PAY\_6  
## ID 20000 2 2 1 24 2 2 -1 -1 -2 -2  
## 1 120000 2 2 2 26 -1 2 0 0 0 2  
## 2 90000 2 2 2 34 0 0 0 0 0 0  
## 3 50000 2 2 1 37 0 0 0 0 0 0  
## 4 50000 1 2 1 57 -1 0 -1 0 0 0  
## 5 50000 1 1 2 37 0 0 0 0 0 0  
## BILL\_AMT1 BILL\_AMT2 BILL\_AMT3 BILL\_AMT4 BILL\_AMT5 BILL\_AMT6 PAY\_AMT1  
## ID 3913 3102 689 0 0 0 0  
## 1 2682 1725 2682 3272 3455 3261 0  
## 2 29239 14027 13559 14331 14948 15549 1518  
## 3 46990 48233 49291 28314 28959 29547 2000  
## 4 8617 5670 35835 20940 19146 19131 2000  
## 5 64400 57069 57608 19394 19619 20024 2500  
## PAY\_AMT2 PAY\_AMT3 PAY\_AMT4 PAY\_AMT5 PAY\_AMT6 default.payment.next.month  
## ID 689 0 0 0 0 1  
## 1 1000 1000 1000 0 2000 1  
## 2 1500 1000 1000 1000 5000 0  
## 3 2019 1200 1100 1069 1000 0  
## 4 36681 10000 9000 689 679 0  
## 5 1815 657 1000 1000 800 0

### Original Variables

Below is the explanation of the original variables of the dataset:

**ID**: ID of each customer

**LIMIT\_BAL**: This integer is the amount of credit in NT dollars given to customer and his/her family

**SEX**: Dummy variable, 1 for male, 2 for female

**EDUCATION**: Indicates level of education, 1 = Graduate School. 2=University, 3=High School, 4=Other

**MARRIAGE**: Indicates marital status, 1= Married, 2 = Single, 3= Other

**AGE**: Age in years

**PAY\_0-PAY\_6**: History of past payments, -2= No consumption, -1= Pay Duly(Full balance paid off), 0 = Use of revolving credit, 1 = Payment delayed for one month, 2 = payment delayed for 2 months … and so on until 9 = payment delayed for 9 months.

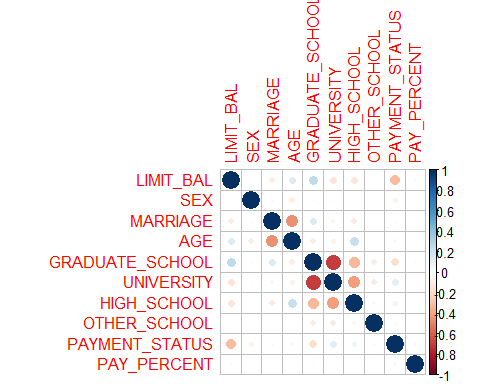
**BILL\_AMT1-6**: Amount of bill statement from September 2005 -April 2005

**PAY\_AMT1-6**: Amount of previous payment in September - April 2005

**default.payment.next.month** = If customer defaults on next payment, 1=yes, 0 =no, This is the response variable

### Data Transformations

## corrplot 0.84 loaded



## LIMIT\_BAL SEX MARRIAGE AGE GRADUATE\_SCHOOL UNIVERSITY HIGH\_SCHOOL  
## ID 20000 1 0 24 0 1 0  
## 1 120000 1 1 26 0 1 0  
## 2 90000 1 1 34 0 1 0  
## 3 50000 1 0 37 0 1 0  
## 4 50000 0 0 57 0 1 0  
## 5 50000 0 1 37 1 0 0  
## OTHER\_SCHOOL PAYMENT\_STATUS PAY\_PERCENT DEFAULT  
## ID 0 -2 8.943406 1  
## 1 0 3 29.279147 1  
## 2 0 0 10.838834 0  
## 3 0 0 3.625926 0  
## 4 0 -2 54.005433 0  
## 5 0 0 3.263983 0

In the end we end up with 12 variables, if there were changes to variables, they are explained below:

**LIMIT\_BAL**: Same

**SEX**: The dummy variable was changed to 0 and 1 instead of 1 and 2. This was done for simplicity and preference.

**MARRIAGE**: I assinged marraige as 1 as well as assigning other to single. So I essentially changed the variable to be a dummy to represent if a person is married or not married. I chose this method because although I recognize being in a relationship with someone is not being single, there are the most financial implications for being married. Additionally, those who are married share finances with their spouse.

**GRADUATE\_SCHOOL, UNIVERSITY, HIGH\_SCHOOL, OTHER\_SCHOOL**: This set of variables was the most difficult to decide on. Assigning a number value to the level of education is tricky. While implying a higher number could mean more educated, for example test results, in this case I could not make that distinction. As especially with the other category, measuring education level becomes more vague and not quantifiable. Therefore I decided to put these variables as dummies as the highest level of education acheived. I decided to only put a 1 for the highest level of education instead of multiple, for example, someone who went to Graduate School also went to University and High School. I made this choice to avoid multicollinearity as it seems inevitable that each column would be too related. As seen from the correlation matrix, High School and University seem to be correlated, but not too much so I let them stay. In addition to this, the dummies are more accurate because they portray each education level as being equl “distance” to eachother. This is important because if we made the education level continuous, a 4 for Graduate school and a 1 for other school would imply that graduate school is 3 “better” than other school, and although that may be the case, for this project I decided that each level of education is different(like the orange juice brands) and not better or worse than another.Lastly, from the correlation plot above, the negative correlation between graduate school and university can be explained because those who went to graduate school “did not” go to university(in my dataset), so it follows that they are negatively correlated.

**PAYMENT\_STATUS**: For this variable, I took all the PAY variables and summed them. I did this mostly to make the data static instead of time series. With this, the value of this shows timeliness the customer pays their credit card bills. The lower the value the better as each postive number represents how many months late the customer was paying in a period. These values can also be negative, so negative represents a better customer in this case as it represents a customer who uses the credit card and pays it off.

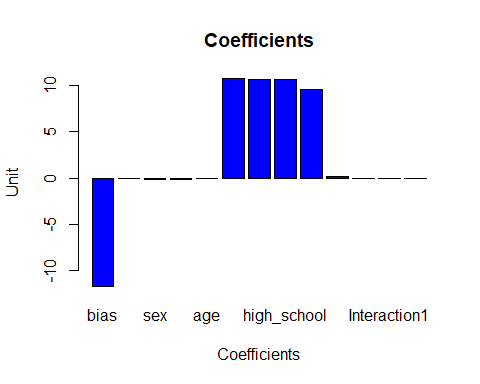
**PAY\_PERCENT**: This variable was again made as a way to condense the time series data to static. For this, I added the toal bill accumulated in the six month period and the total amount paid off in the six month period and divided the amount paid by the amount billed. This again is used to see the quality of a customer in paying off their bills. Hopefully this metric can be used as a way to predict the dependability of a customer.

**INTERACTION VARIABLES**

I plan to use interaction variables between the **PAY\_PERCENT and the PAYMENT\_STATUS** as it seems reasonable that the percentage paid of a customer relates to the lateness that they pay with as a customer who pays less is probably more behind on payments. I also want to use a second interaction variable between **LIMIT\_BAL and PAYMENT\_STATUS**. I believe that these are related as a person with a higher limit balance is most likely more reliable on payments as they would not get a raise in their limit by the credit card issuer if they did not make payments on time.

**Additional**: I changed binary variables including my response variable to factors for logistic regression. Additionally, I removed the last element, 30001 because it was causing trouble in my k fold cross validation. Lastly, I made the correlation plot to get a better idea of the dataset visually.

## Model

My base model is as follows below with all the variables from the altered data. 

### Interpreting Model Results

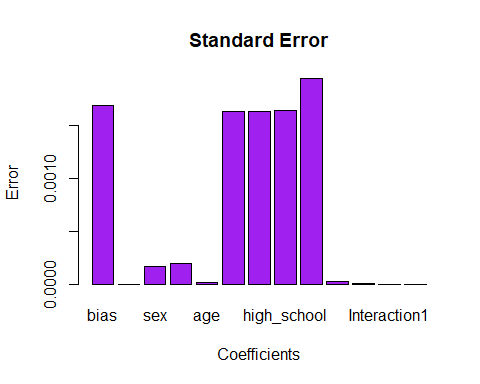
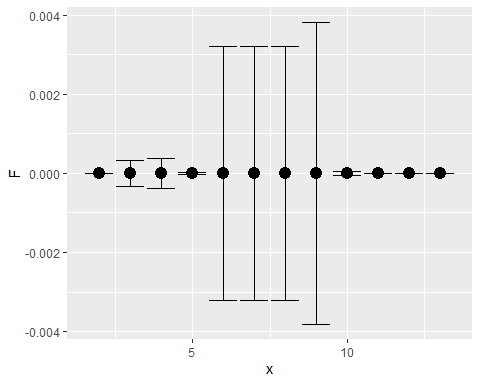
Looking at the bar chart for the results of the logistic regression, it is visible that the bias and 6th through 9th and the bias almost cancel out. This makes sense because every sample has at least one level of education. Looking past these larger coefficients, it seems that marriage has abouta -0.2 effect on the odds of the outcome, and payment\_status has 0.146 increase on odds. This makes sense as those who are married are typically less likely to default. Additionally, from our scoring system as detailed before for payment\_status, the higher the score, the more delinquent payments the customer has; meaning those with higher score are more likely to default. Lastly for the interaction terms, they had little impact as we might have predicted by the lack of correlation shown from the correlation plot.

## Results(Bootstrap, CV, Lasso)

### Bootstrap

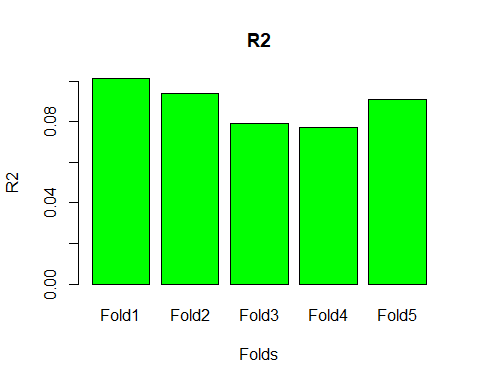
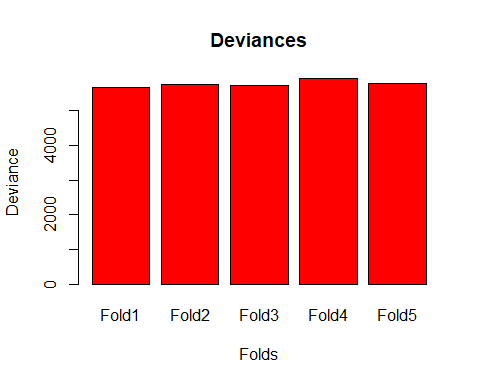
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Loading required package: ggplot2



The purpose of bootstrap is to estimate the standard errors of each coefficient. This gives us the a better idea of the the range that coefficeints can exist in. Additionally, we can make confidence intervals for each predictor to show the spread for across (in this case) 500 outcomes for the same predictor across 500 draws of n with replacement. I did not make histograms for the distribution of the 500 coefficients to save some space; with that being said, they are present and are useful figure to see how the coefficients are distributed. I opted to more breifly plot the standard error measured from each parameter. I also made a 95% confidence interval plot as to show the spread the coefficients covered. While I recognize I put the mean at zero, I did this in order to get the plots more comprehensible as I was more focused on the ranges rather than the real confidence interval values. When doing this I also considered standardizing the data, but with so many binary variables, I decided against it.

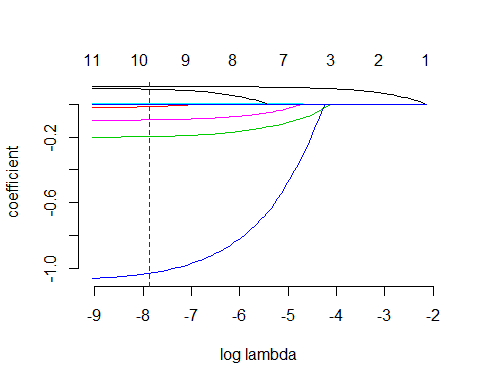
### Cross-Validation



For the purpose of my project, cross-validation would be used to test the accuracy of the model to see if it should actually be used to approve credit cards or increase limit balances. It is an assurance that the model is working how it should be. Since we only have a given dataset, we do not know how our model will respond to new data. In an attempt to simulate this, we divide the dataset into k sections, in this case 5. From this, we can train on 4 of the data sets, and see how our model responds to “unseen” data. We repeat this across all the different folds. This gives us an idea of how will our model will perform with real unseen data even though there are imperfections still. This can also be used with Lasso to choose which variables to maintain. To find the measurement for out of sample deviance for a logistic model as we did not cover it as much in class. However, I found the slides measuring deviance as -2 times the log likelihood + Constant and R2 as 1 - deviance(Beta\_hat)/deviance(Beta=0). With this I was able to measure my out of sample errors as I saved them across each fold and graphed them above. It can be seen by the graphs that the model did a poor job of predicted unseen data as the highest R2 was 0.1.

### Lasso

## Loading required package: Matrix



For my Lasso, I used the gamlr library, unfortunately, I was not able to run cross validated Lasso from the gamlr library as it did not work for me with logistic regression. I attempted to look to other libraries that could help and found one called glmnet, although it also produced errors when fed my data. So I was not able to conduct cross-validated Lasso, but I was able to do each separately.

**Lasso**: Based on a tuning parameter lambda, we can assign a penalty for more coefficients as to prevent overfitting. This penalty is added to the deviance to punish models with many regressors. This is good as it allows us to supposedly predict more accurately because the model is not as dependent on the training data. Additionally, this allows us to limit the amount of predictors we have to provide some more clarity in causal inference.

**Cross-Validated Lasso**: This method is one in which we can find the optimal value for lambda. In this, we cross-validate as before into k folds. From this we test the model with a given lambda on each fold, saving the out of sample deviance of each set of lambda tests. We go through all the lambdas and choose the lambda that results in the lowest out of sample deviance from all the iterations.

Why are Lasso and Cross-Validated Lasso appropriate methods for my question? Since my model is trying to predict if a customer is going to default, it is important for the predictions it makes to be correct. With this, implementing Lasso can decrease my out of sample deviance by preventing overfitting of my model. We would like to have the cross-validation method in order to find the optimal lambda to use for the model.

The figure above shows my lasso path plot for my model. As the model goes from right to left more coefficients are added. The dotted vertical line depicts optimal lambda.

## Reference List

**1.** Aslam, Uzair & Aziz, Hafiz Ilyas Tariq & Sohail, Asim & Batcha, Nowshath. (2019). An Empirical Study on Loan Default Prediction Models. Journal of Computational and Theoretical Nanoscience. 16. 3483-3488. 10.1166/jctn.2019.8312.

**2.** Husejinovic, Admel & Keco, Dino & Masetic, Zerina. (2018). Application of Machine Learning Algorithms in Credit Card Default Payment Prediction. International Journal of Scientific Research. 7. 425. 10.15373/22778179#husejinovic.

**3.** Kayla Hartman. (2018). Towards Ethical Machine Learning. Towards Data Science. <https://towardsdatascience.com/towards-ethical-machine-learning-302e580f5815>

**4.** John Hilsenrath & Kate Davidson. (2020). U.S. Braces for Sharp Economic Downturn as Coronavirus Bears Down. Wall Street Journal. <https://www.wsj.com/articles/u-s-braces-for-sharp-economic-downturn-as-coronavirus-bears-down-11584386207?ns=prod/accounts-wsj>

**5.** Adam Pruden. (2019). Can Artifical Intelligence Prevent the Next Financial Crash. Frog Design. <https://www.frogdesign.com/designmind/can-artificial-intelligence-prevent-the-next-financial-crash/>

## Conclusion

While my particular model was unable to predict defaults of credit card payments by customers successfully, this is still an important topic worth revisiting. In the future I would like to do this model again with different polynomial values and different interaction terms as well. Additionally, my alterations to the data which I eventually regressed on were probably the source of my errors as I feel I need to research more on how to deal with the education level as well an alternate way to accurately represent the time series data statically.