Springboard Capstone Project 1 Milestone Report

In 2006, Taiwan faced a credit crisis that came to a head in the third quarter of 2006. This all started with banks over-issuing cash and credit to unqualified applicants and was catalyzed by many cardholders, irrespective of their repayment ability, overusing their credit cards for consumption, accruing heavy credit and cash-card debts. This crisis warranted more thorough investigation into improved techniques for assessing the probability of default by each individual consumer.

We have a lot of interesting data from this time period, as detailed below.

**Attribute Information:**

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

* X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
* X2: Gender (1 = male; 2 = female).
* X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
* X4: Marital status (1 = married; 2 = single; 3 = others).
* X5: Age (year).
* X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
* X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
* X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

In summary, there are 23 features of the dataset that can be explored to get a better understanding of the information provided. Some of them are time-series related information, such as the history of past payment, amount of bill statement in NT dollars, and the amount of previous payments. These are all very abstract concepts and, while they may be informative on their own for a machine learning model, they are quite hard to understand as they are right now. To rectify this issue, I would like to construct visualizations that show where the state of the Taiwan economy is from April to September of 2015, by each customer.

From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. With the real probability of default as the response variable (Y), and the predictive probability of default as the independent variable (X), the simple linear regression result (Y = A + BX) shows that the forecasting model produced by artificial neural network has the highest coefficient of determination; its regression intercept (A) is close to zero, and regression coefficient (B) to one. Therefore, among the six data mining techniques, artificial neural network is the only one that can accurately estimate the real probability of default.

## Data Wrangling

The data was fairly straightforward to wrangle and understand. While it had a duplicate row for column names, that was an easy fix. Since this is a dataset from a UCI repository, there were no null values present in any of the columns. However, for some categorical data, there was data that was not described in the dataset description. For example, education should just range from 1-4, but there are values such as 0, 5, and 6 present in the dataset. There were other unexplained variables in the payment status columns, such as -2 and 0, but without any further explanation of the meaning, I can only move forward by examining their relationships with the data and how it relates to the model I am trying to make.

### One-Hot Encoding

Since there were various columns with categorical data, I needed to employ one-hot encoding on this dataset. One-hot encoding is useful in python to train any type of model that contains categorical features. In this case, there are 9 different features that contain categorical values. Within those columns, there are varying options for distinction, such as sex having 2 options and history of past payment having 10 different values. One-hot encoding will split each of these columns into columns that have one category in the column, and whether that value is present in the row. Going back to the 'sex' category, this will split it into two columns called: "SEX\_male" and "SEX\_female", each of which will only contain binary variables.

## Data Story

This dataset is rich with information and I wanted to better understand the variables that I will be using by visualizing various features of the dataset. To do this, I mainly employed parallel coordinate plots of payment history for each customer. This resulted in semi-time series analysis of the macro-trends of credit consumer behavior. While there are more examples in the notebook provided online, here are two examples of the plots created.

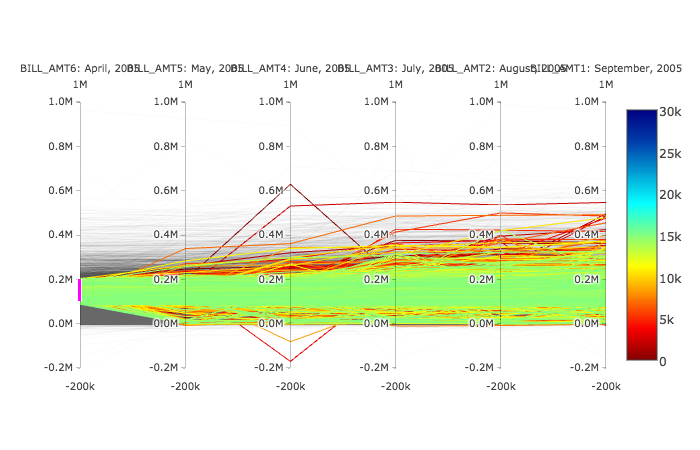


Figure : Trend of total bill amount for each customer from April, 2005 to September, 2005. The plot focuses on customers who had between 100k-200k of debt in NT Dollars.

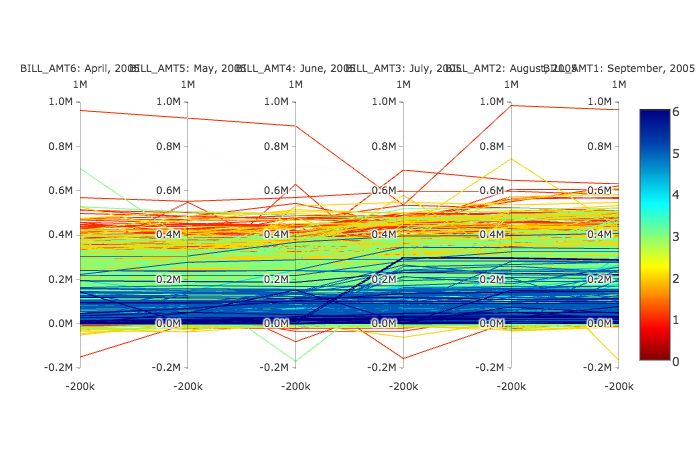


Figure : Trend of the total bill amount segmented by education population. This plot will form the basis for inferential statistics later performed.

Both of the figures observed the amount of debt that each Taiwanese credit customer had from April, 2005 to September, 2005. Figure 1 highlights an arbitrarily selected group of customers from the dataset and tracks their bill amount over time. The only real trend that can be seen is that the mean and variance of this debt increase over time.

Figure 2 looks at debt over time by consumers in different ranges of education. There is a lot of noise here, but it appears the higher educated classes of people have more debt, on average than their lower-educated counterparts. This observation formed the basis for hypothesis testing of the amount of debt had by different groups of people, binned by level of education.

## Inferential Statistics of Taiwan Credit Card Users

So far, we have cleaned and assessed the data within the Taiwan Credit dataset. We examined the data for fatal flaws as well as trends among all of the customers, discovering the data was relatively clean enough to use to create a prediction model and that there were interesting trends in credit user behavior over time. Now I'll use some summary statistics to test some hypotheses among certain groups of interests.

I will focus on education level and start by using an ANOVA-type of analysis of the average payment amount and bill amount to assess whether education makes a difference in the average level of debt and payment among each of the groups. I will then see if the highest education has a noticeably higher amount of debt than the other groups combined.

To perform the anova-type analysis, I will need to compute the summary statistics of the bill amount and pay amount of each Taiwan customer over time and then use those statistics to compute the distributions for each person. I'll create four new columns: BILL\_AMT\_AVG, BILL\_AMT\_STD, PAY\_AMT\_AVG, PAY\_AMT\_STD.

## Testing Normality!

Before I get ahead of myself and perform ANOVA testing, I need to test one of the main assumptions, that the datasets are normally distributed. I'll get to the point by plotting the Q-Q plots of each average amount.

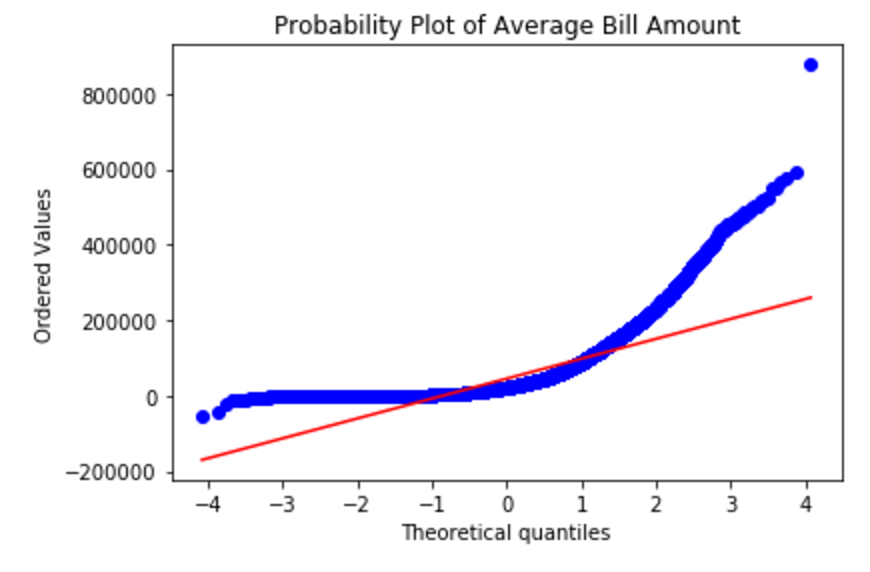


Figure : The average bill amount of each customer is obviously skewed. There is a semi-straight line until the -1 quantile, where the graph starts to behave exponentially.

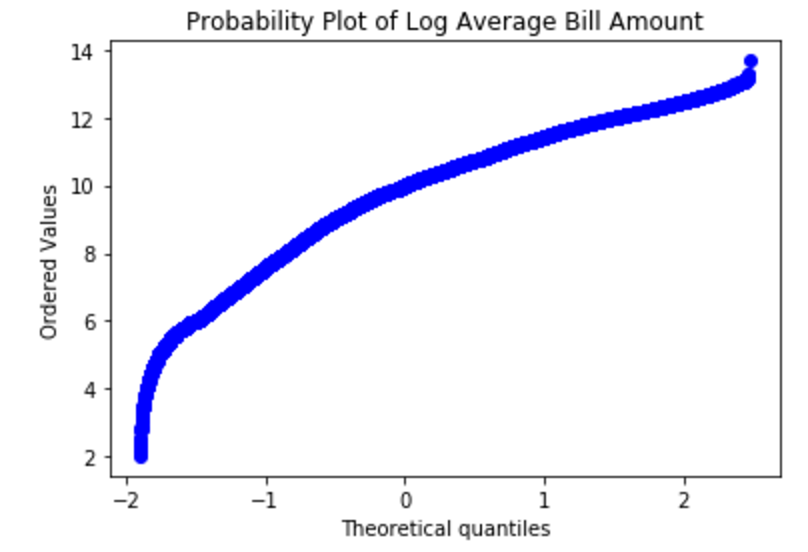


Figure : An attempt to remedy the skewed data, this graph depicts a log-normal transform of the previous figure. Due to the separate functions, the log transform made the linear part logarithmic, and the exponential part straight, leaving skewed data.

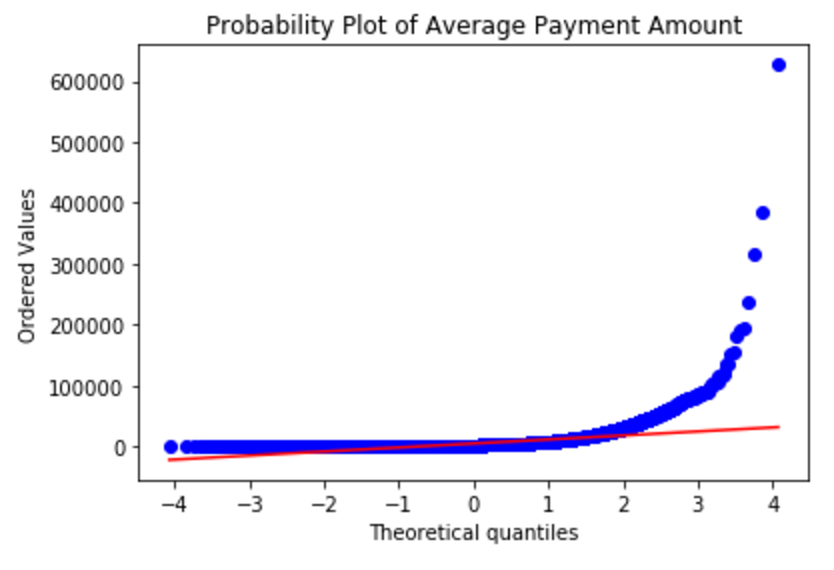


Figure : The average payment amount of each individual has a wider range of normality than the bill amount, but still encounters skew with the top 90% of consumers having exponential average payments, compared to the average.

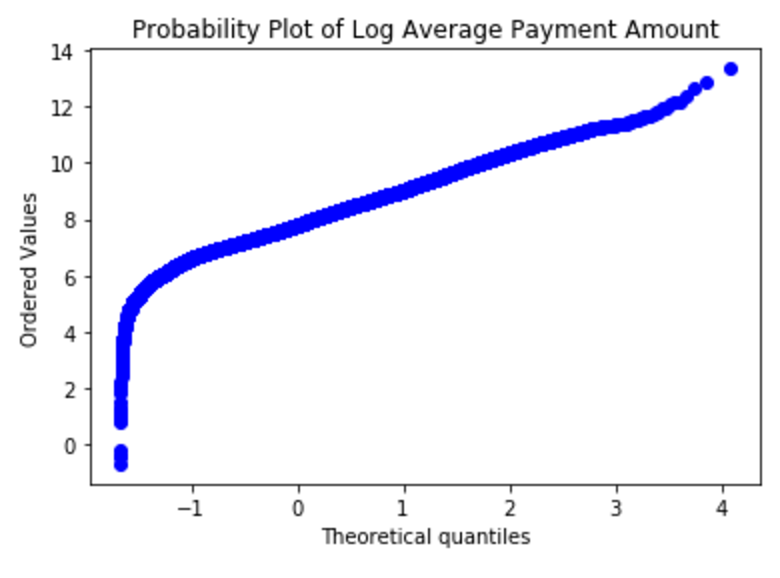


Figure : Again, a log-transform did not help rectify the skewness of the data in the Q-Q plot in Figure 5.

**Q-Q Plot Interpretation**

We can see that for both the average bill payment and the average bill amount, there is a high degree of non-linearity. When applying a logarithmic transformation to the data, there is still a discrepancy. The plots seem to be linear up until a point, and then is exponential. That said, I cannot perform any parametric hypothesis testing on this dataset.

This also gives a hint about the type of modeling method I should use. For example, a linear model may struggle to predict the outcomes of whether a person will default or not in the next month, whereas a neural network or CART model may do better.

**Alternative to Anova: Non-Parametric Hypothesis Testing**

Since The data is clearly skewed and not normal, and log-normalization did not help, I will use a Kruskal Wallis test to determine if there is a difference in the amount of debt accrued by each education category provided. The Kruskal Wallis tests a difference in median instead of mean, and ranks the data by value to perform the analysis. While this ranking helps the test work, it also loses some information related to the data, so as usual, take the result with skepticism unless it is very strong or weak.

**Nonparametric Hypothesis Testing Takeaways**[**¶**](http://localhost:8888/notebooks/Springboard_Assignments/Capstone_Proj_1/New_Proj/Inferential%20Statistics.ipynb#Noparametric-Hypothesis-Testing-Takeaways)

All of the groups differ dramatically from one another. I checked the number of samples for each group to see if there was a dramatic difference between the groups, which there was not. Then I performed the test between groups that I thought wouldn't be statistically different, such as graduate student and college student, and found that they are different. The only not significantly different group was between the graduate credit card holders and the 'others' category. More details can be seen in the jupyter notebook link.

**Estimating Effect Size**

I'll find the mean of each group and compare the effect size of each by computing Cohen's d, as well as overlap and superiority.

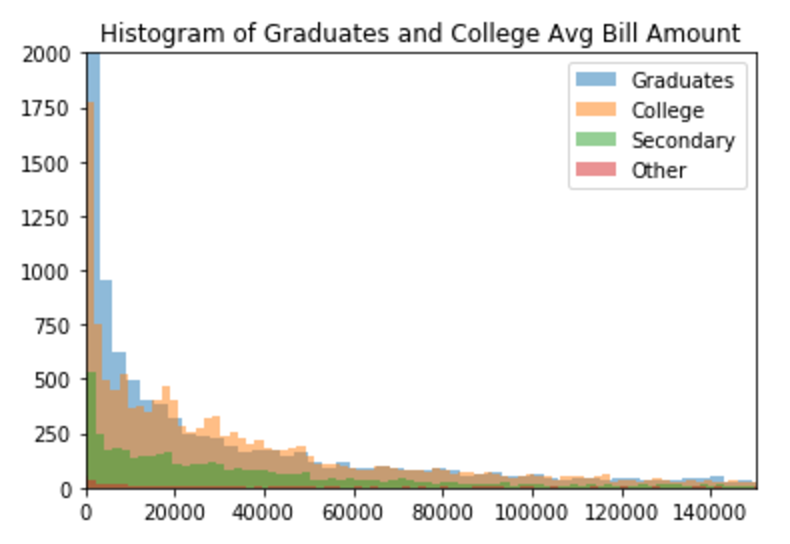


Figure : Distributions of average debt amounts across all the educational categories. Kruskal-Wallis hypothesis testing revealed that each group is significantly different than the other. This figure confirms that statement.

Effect size seems to be very small. A graph of the distributions clearly demonstrates why this would happen.

## Inferential Statistics Conclusion

After seeing a trend among the parallel coordinate plots in the data story notebook, hypothesis testing has not revealed a difference in the amount of debt between each education class. We can see a long-tail distribution among debt holders, and it's my hypothesis that the most damage is done to credit companies by defaults in the longer tail. The modeling and machine learing portion of this project will explore that hypothesis.

## Modelling the Data

We have seen that there is a great deal of nonlinear information. To model this information, I will start with basic logistic regression and measure its accuracy, precision, and recall to classify whether an individual will default in the next month. I will also consider non-linear algorithms, such as CART and neural network models.