Springboard Data Science Career Track
Jesse Mailhot
Mentored by Shmuel Naaman

Capstone Project 1

UCI Credit Card Default (Kaggle.com)

Introduction

Addressing credit card default is a major issue for banks, both small and large. In order to best investigate this issue, a bank would need to analyze data on their credit card clients and identify which features indicate probable default. In this project I will explore how the given features relate to default and build a machine learning model to predict which clients will default. If employed in a business setting, this model could be run on a monthly or weekly basis to identify clients in danger of defaulting, and alert the bank to take appropriate action.

Goals

1. Identify which features best predict default
2. Explore relationship between demographic features and default
3. Create machine learning model to predict default

Dataset

The dataset for this project was obtained from the UCI Machine Learning Repository, via kaggle.com. It contains records of 30,000 credit card clients in Taiwan from April to September, 2005. Below is a list of features included in the dataset.
(Note: all dollar amounts given in New Taiwanese Dollar, NT$)

- ❑ Default (0 - No / 1 - Yes)
- ❑ Credit Limit (NT$)
- ❑ Sex (Male, Female)
- ❑ Education Level (Graduate School, Bachelor's Degree, High School, Other)
- ❑ Marital Status (Married, Single, Other)
- ❑ Age
- ❑ Payment Delay (0 - 9 months)
- ❑ Bill Amount (NT$)
- ❑ Payment Amount (NT$)

Data Wrangling

The dataset does not contain any missing values; however, both the marriage and education features contain unlabeled values which need to be addressed. My solution was to add these to the ‘other’ column. In addition to marriage and education, the payment delay columns had low sample sizes for most values over 2 months. Instead of including the month delay, my solution was to make the the payment delay column binary (0 - payment, 1 - no payment). I also changed the values of the sex column from 1 (Male) and 2 (Female) to 0 and 1, respectively. Additionally, I also changed the labels of the bill, pay, and pay status columns to represent the months.

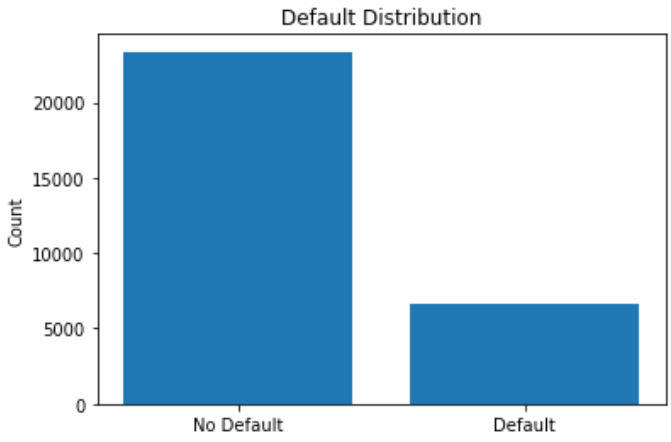
Data Analysis

In this section I looked at the categorical distribution for each variable, as well as how each variable’s categories relate to default.

Default

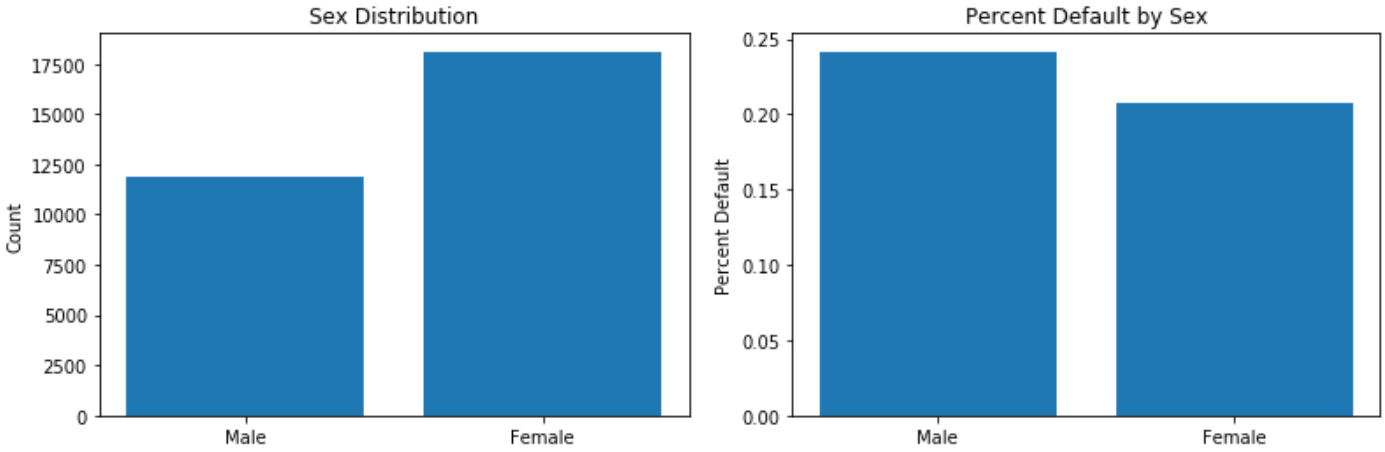
Visualizing the default data, it is clear that the dataset is unbalanced. This will be important once we start constructing our machine learning model since this is our target variable.

(Note: 22.12% default rate)



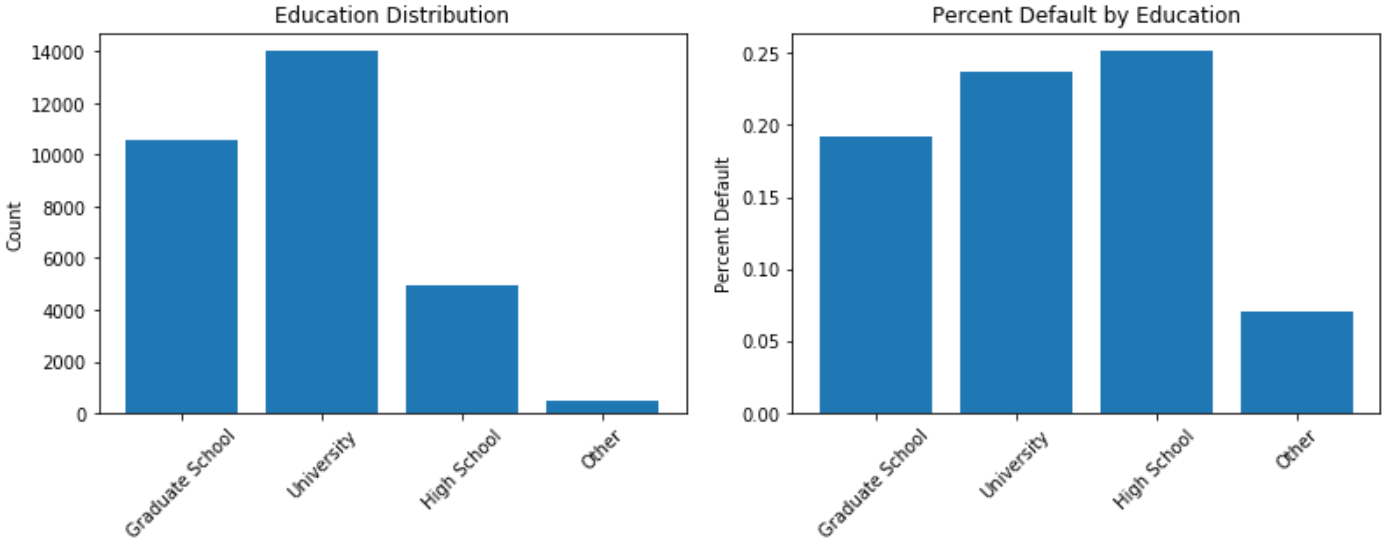
Sex

The dataset contains approximately 60% female clients, but male clients defaulted approximately 3% more than female clients. A chi squared test for sex produces a test statistic of *47.71* and a p-value of $4.94e^{-12}$, indicating that mean default for males and females is not the same.



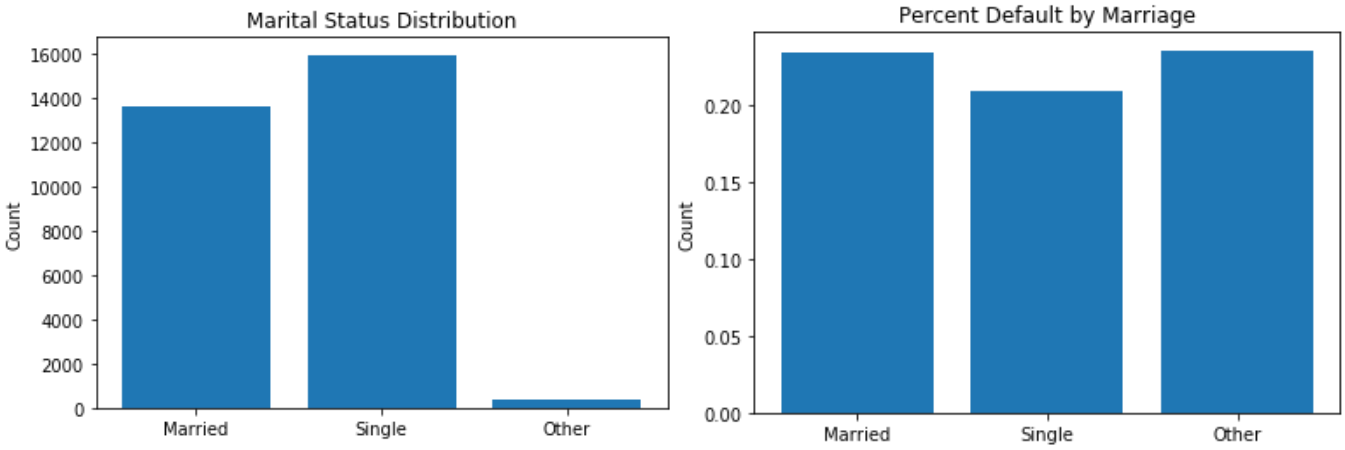
Education Level

There is clear negative correlation between level of education and probability of default. The 'other' category is an outlier; however, it also has a low sample size and can be disregarded in our analysis as it represents only 1.56% of the population. A chi squared test for education produces a test statistic of *160.41* and a p-value of $1.50e^{-34}$, indicating that mean default for level of education is not the same.



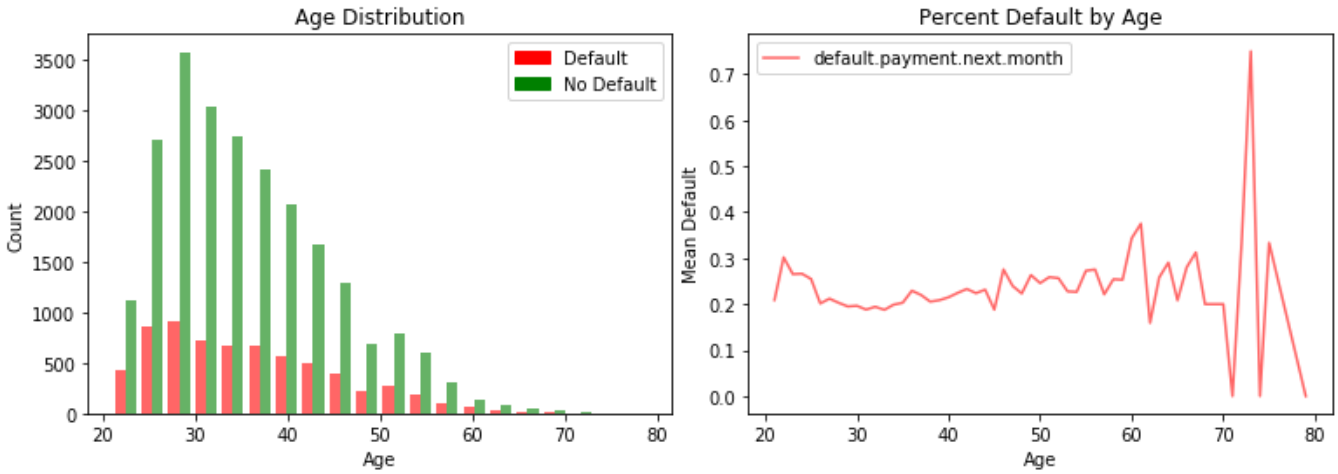
Marital Status

Based on these plots, it is clear that married clients are more likely to default than single clients. Further analysis could be done to investigate the relationship between sex, marital status, and default. The 'other' category is again an outlier representing only 1.26% of the population, this time with a high default rate. A chi squared test for marital status produces a test statistic of *28.13* and a p-value of $7.79e^{-7}$, indicating that mean default for marital status is not the same.



Age

Based on the percent default by age plot, the group most likely to default is age 20-25. Default is lowest at 25-35, staying around the population mean of 22%, then increases as age passes 50+. Due to the low sample size of clients over 50, it is not possible to gain any useful insights into the default rate of this age group for this dataset. From the age distribution, it is clear that the majority of the clients in this dataset are between age 25 and 40. A t-test for default by age produces a test statistic of *-2.41* and a p-value of *0.016*, indicating that mean age for defaulting clients is not significantly different than the mean age of non-defaulting clients.



Summary of Demographics

From the demographic data, the following trends have been identified:

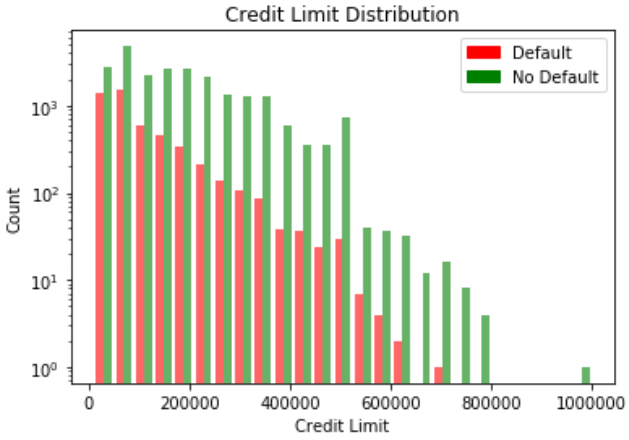
- ❏ Majority of the clients are between the ages of 25 and 40
- ❏ Limited sample of clients age 50+
- ❏ Approximately 60% of the clients are female
- ❏ Female clients defaulted approximately 3% less than male clients
- ❏ Negative correlation between default and level of education
- ❏ Married clients defaulted approximately 2% more than single clients
- ❏ 'Other' categories for both marriage and education represent less than 2% of the population

With this information in mind, I will now continue the exploratory data analysis of the payment features.

Credit Limit

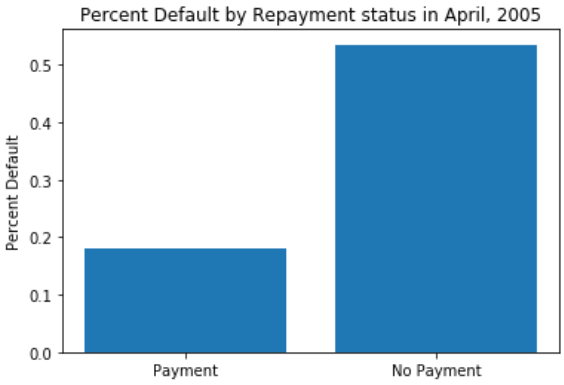
Clients that defaulted have a noticeably lower credit limit than clients who did not default. There are some outliers with very high credit limits, but these did not default so we will ignore them as this is consistent with the trend of the data. A t-test for default by credit limit produces a test statistic of *26.91* and a p-value of $1.30e^{-157}$, indicating that mean credit limit for defaulting clients is significantly different than the mean credit limit of non-defaulting clients.

(Note: customers who defaulted had a mean balance of NT$47,990 less than clients that did not default)

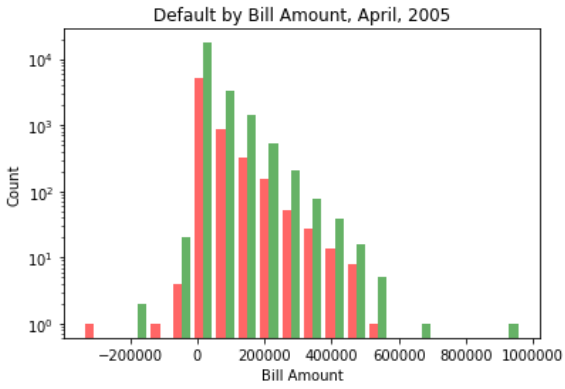


Pay Data - April, 2005

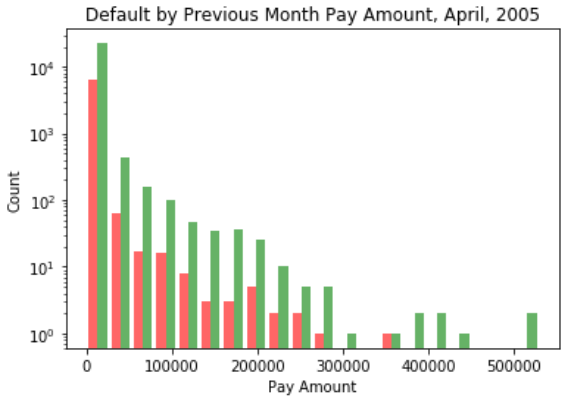
Looking at the plot of repayment status for April, it is interesting to note that clients who made payments still defaulted at a rate of 18.67%. This seems a bit high for clients who are paying, but perhaps they are spending more than they paying, thus resulting in eventual default. This will be investigated further in a later section of the report. From the bill amount plot we can see that there are some negative bill amounts, as well as some very high bill amounts. Particularly, take note of the negative bills that are classified as default. This leads me to believe that these data points may be incorrectly classified. As we will see, this trend of negative bills defaulting, as well as paying clients defaulting at a relatively high rate continues throughout the 5 months of data in our dataset. As expected, the clients who made no payment defaulted the most compared to clients who paid more. A t-test for default by bill amount and pay amount produces test statistics of *0.930* and *9.22,* and p-values of *0.352* and $3.03e^{-20}$ respectively, indicating that the mean bill amount for defaulting clients is not significantly different than the mean bill amount of non-defaulting clients, but the mean pay amount is significantly different.



Default rate: 18.67%



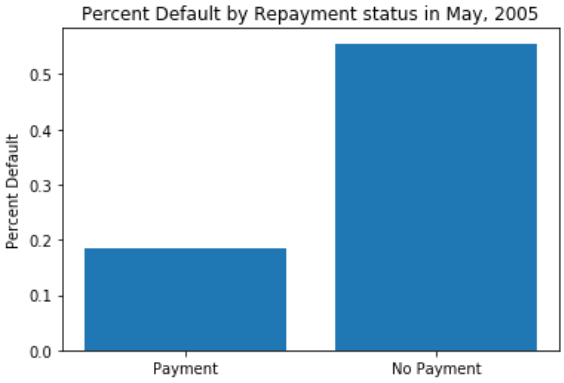
Average default bill amount: NT$38,271.44



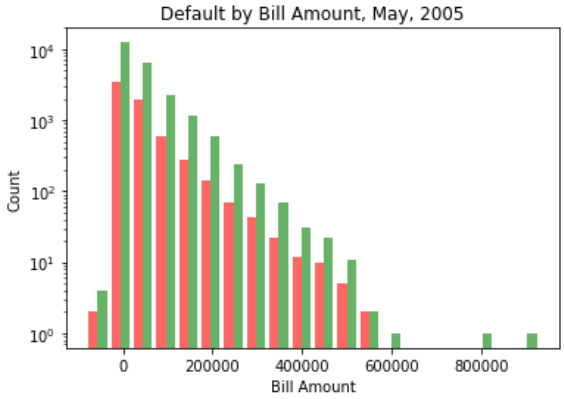
Average default payment amount NT$3,441.48

Pay Data - May, 2005

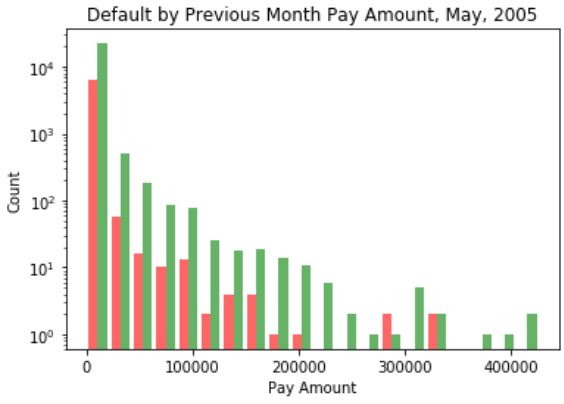
In May the percent of paying clients who defaulted is 18.45%; while this is 0.22% lower than April, it is still higher than expected. There are no outliers in terms of clients with negative bills who defaulted, however, there are still clients with negative bills who are classified as default. The average bill amount of clients who defaulted increased by $1,265.75, while the average payment amount of clients who defaulted decreased by $222.34. A t-test for default by bill amount and pay amount produces test statistics of *1.17* and *9.56,* and p-values of *0.241* and $1.24e^{-21}$ respectively, indicating that the mean bill amount for defaulting clients is not significantly different than the mean bill amount of non-defaulting clients, but the mean pay amount is significantly different.



Default rate: 18.45%



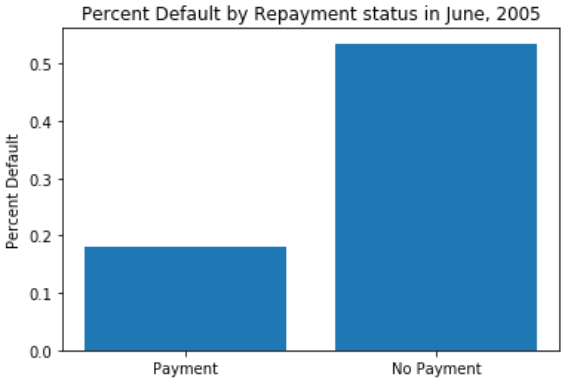
Average default bill amount: NT$39,540.19



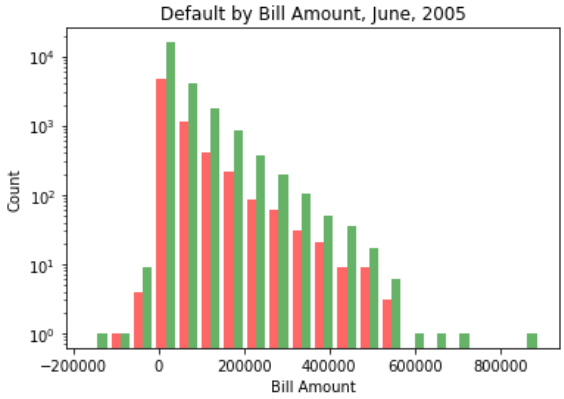
Average default payment amount NT$3,219.14

Pay Data - June, 2005

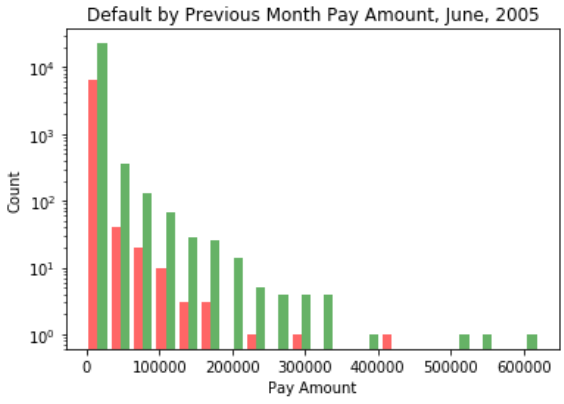
In June the percent of paying clients who defaulted is 17.95%; again decreasing from the previous month, this time by 0.5%. The trend of clients with negative bills defaulting continues, as does the trend of clients who pay nothing being the group with the most defaults. The average bill amount of clients who defaulted increased by $2,495.76, while the average payment amount of clients who defaulted decreased by $63.51. A t-test for default by bill amount and pay amount produces test statistics of *1.75* and *9.85,* and p-values of *0.079* and $6.83e^{-23}$ respectively, indicating that the mean bill amount for defaulting clients is not significantly different than the mean bill amount of non-defaulting clients, but the mean pay amount is significantly different.



Default rate: 17.95%



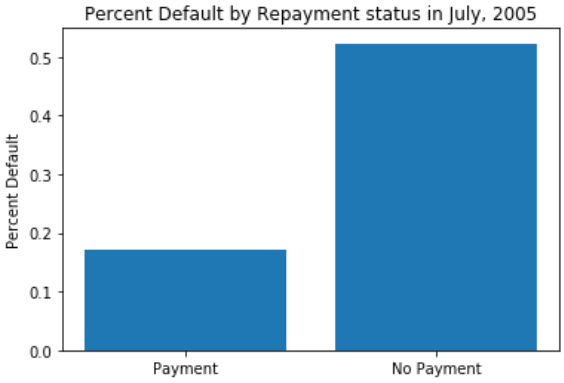
Average default bill amount: NT$42,035.94



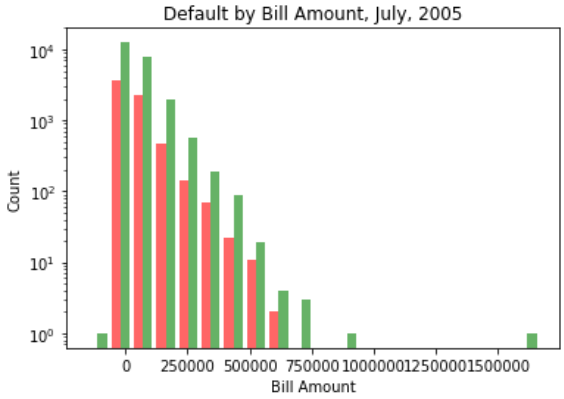
Average default payment amount NT$3,155.63

Pay Data - July, 2005

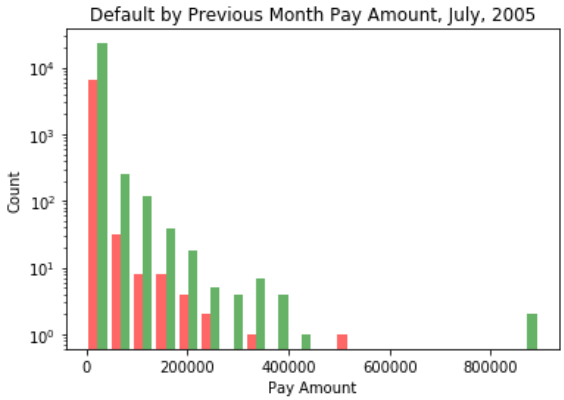
In July the percent of paying clients who defaulted is 17.19%, 0.76% lower than June. The trend of paying clients defaulting continues to trend slowly downward as time goes on. There is one clear outlier in the bill amount column, but since they did not default we do not need to be concerned with it; however there are still customers with bills of $0 or less defaulting. The average bill amount of clients who defaulted increased from June by $3,145.66; however, the average payment amount of clients who defaulted also increased from June by $211.72. A t-test for default by bill amount and pay amount produces test statistics of *2.43* and *9.75,* and p-values of *0.015* and $1.84e^{-22}$ respectively, indicating that the mean bill amount for defaulting clients is not significantly different than the mean bill amount of non-defaulting clients, but the mean pay amount is significantly different.



Default rate: 17.19%



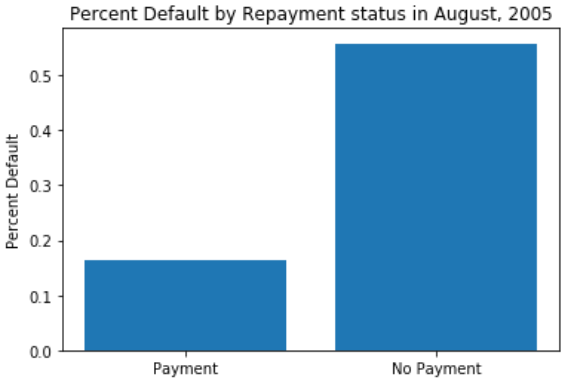
Average default bill amount: NT$45,181.60



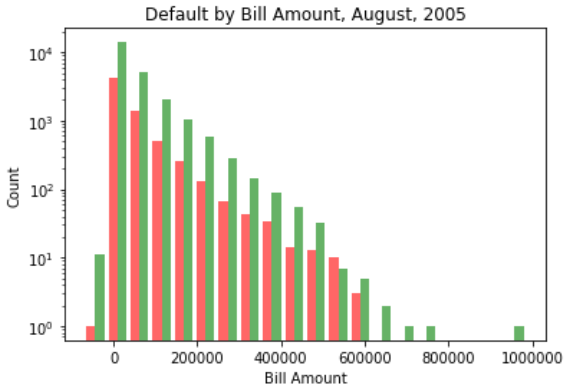
Average default payment amount NT$3,367.35

Pay Data - August, 2005

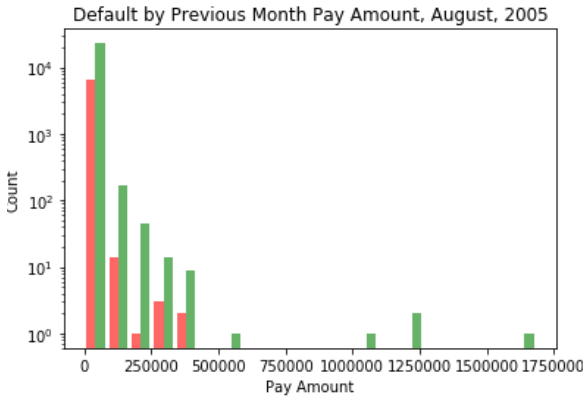
In August the percent of paying clients who defaulted is 16.27%, down 0.91% from July. The trend of customers with bills of $0 or less defaulting continues. In August it is the pay amount column that has outliers. Again, since these outliers did not default we will ignore them since we are trying to predict default. The average bill amount of defaulting clients increased from July by $2102.02, and the average payment amount of defaulting clients increased from July by $21.30. A t-test for default by bill amount and pay amount produces test statistics of *2.46* and *10.16,* and p-values of *0.014* and $3.17e^{-24}$ respectively, indicating that the mean bill amount for defaulting clients is not significantly different than the mean bill amount of non-defaulting clients, but the mean pay amount is significantly different.



Default rate: 16.27%



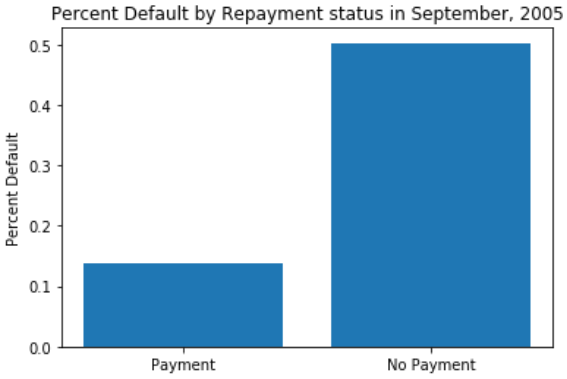
Average default bill amount: NT$47,283.62



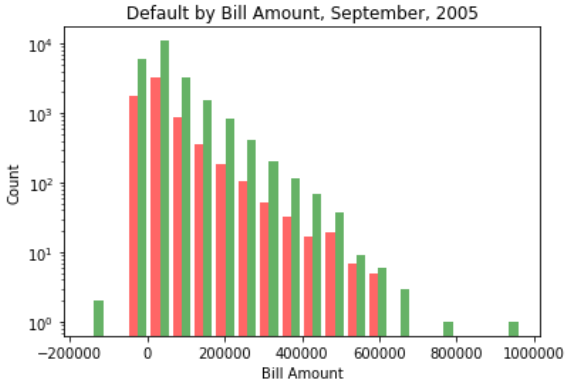
Average default payment amount NT$3,367.35

Pay Data - September, 2005

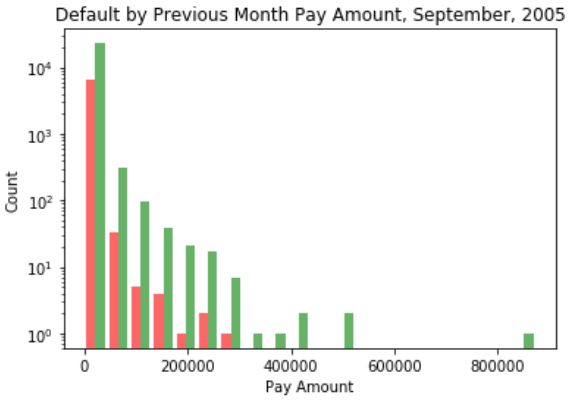
In September the percent of paying clients who defaulted is 13.83%, down 2.44% from August. The trend of customers with bills of $0 or less defaulting continues. The average bill amount of defaulting clients increased from August by $1,225.54 and the average payment amount of defaulting clients increased from July by $8.39. A t-test for default by bill amount and pay amount produces test statistics of *3.40* and *12.66,* and p-values of *0.001* and $1.15e^{-36}$ respectively, indicating that both the mean bill amount and pay amount for defaulting clients is significantly different than the mean bill amount and pay amount of non-defaulting clients.



Default rate: 13.83%



Average default bill amount: NT$48,509.16



Average default payment amount NT$3397.04

Summary of Payment Insights

This concludes the exploratory data analysis for the demographic features of our dataset. To recap, we learned:

- ❏ Percent of paying clients that defaulted decreased over time
- ❏ Clients with balance of 0 or less defaulting
- ❏ Average bill amount of defaulting clients increased over time
- ❏ Large difference between average bill amount and average payment amount for defaulting clients
- ❏ Based on the statistical tests, the mean bill amounts of defaulting and non-defaulting clients are not significantly different, with the exception of September, while the mean pay amount is significantly different for all months

In regards to the percent of paying clients that defaulted decreasing over time, this is may be due to clients who defaulted making no further payments. Clients with a balance of $0 or less defaulting is concerning, as this will prove difficult for our model to identify and classify correctly. I will investigate this further in the next section. The last two points make sense in terms of credit card clients who default: high spending, low payment, eventually default.

Data Modeling

The first step of this phase of the project is to engineer a new feature to describe clients who paid their bills in full. This will enhance the ability of the model to identify responsibly paying clients. This will be followed by feature selection, model and metric selection, and finish with model tuning and testing.

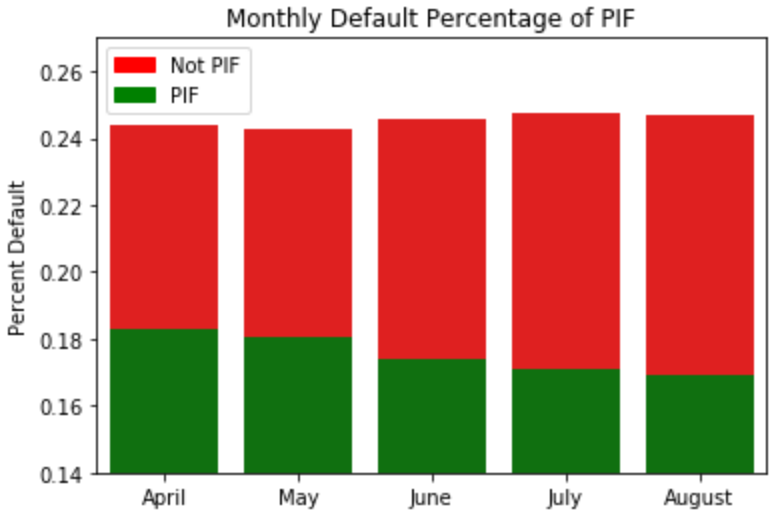
Feature Engineering

In order to help the model better classify default, I created the PIF feature. This new feature, labelled 'PIF_N', is a boolean feature that represents whether the payment amount in month N + 1 is greater than or equal to the bill amount for month N.

(ex. PIF_8 = PAY_AMT9 >= BILL_AMT8)

Since the dataset ends in September, I was only able to create PIF features from April through August.

After creating the PIF feature, I ran analysis to see what trends were present in the data. Below is a plot of the percent of clients who defaulted based on whether or not they paid their bills in full.



The clients who paid their bills in full still defaulted at a very high rate (16.92 for August). This doesn't make sense: how could someone who pays their bill in full each month default? Although we do not have data beyond September, 16.92% of clients who paid their August bill in full labeled as default seems unreasonably high, and definitely out of the ordinary. Further, some of these clients had a negative balance. This led me to believe that these clients did not actually default, but likely closed their accounts.

Feature Selection

Based on the findings from the analysis of the PIF feature, I decided to drop clients who closed their accounts from the dataset using the parameters *default = true & PIF_8 = True*. This improves the model's ability to correctly classify default and helps reduce the amount of false negatives produced. Clients who closed their accounts and were incorrectly labeled as default account for approximately 5.64% of the original population.

Since the PIF feature sufficiently captures the pay amount and bill amount information, the pay amount and bill amount features were not included in the model. The following features were included in the model: *credit limit, sex, education, marital status, age, pay status, and PIF.*

In order to properly train the model, the categorical features (education, marital status) must be split into dummy variables. Using Pandas *get_dummies* function, these columns were split into their corresponding dummy variables, and the 'other' categories dropped for data integrity. The dataset was then split into testing and training sets using *train_test_split* with 20% test size.

Model Selection

Given the objective of creating a model that predicts default, this is a classification problem. Using Scikit Learn, I tried out a few different models: Decision Tree, Naive Bayes, Random Forest, and AdaBoost. The decision tree performed reasonably well, especially since it includes the balanced class weight parameter for unbalanced datasets. The Naive Bayes classifier did not perform well with this dataset, nor did the AdaBoost classifier, likely due to the unbalanced nature of this dataset. The Random Forest model performed the best as it is more robust than a single decision tree and also includes the balanced class weight parameter to help mitigate against unbalanced datasets.

Metric Selection

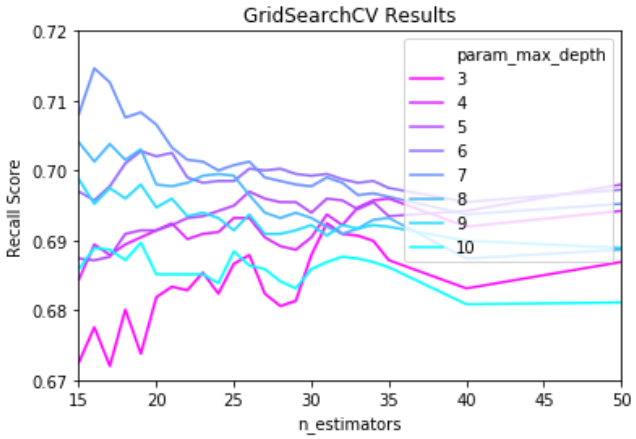
If this were a business scenario in the real world, I would be working for a bank whose primary goal would most likely be to save money. With that in mind, the most appropriate metric for this problem is recall score, defined below:

$$TP / (TP + FN)$$

Where *TP* represents the true positives (correctly classified as positive), and *FN* represents the false negatives (incorrectly classified as no default). In a business sense, it is more important to maximize the model's ability to identify clients who will default than to predict clients who will not. Since false negatives represent the clients who default but are predicted not to, it is important to focus on minimizing this group.

Model Tuning

Using Scikit Learn's *GridSearchCV* function, I maximized the model's recall score. The model performed best with a max depth of 7 and 16 estimators, which produced a recall score of 0.7147 on the training set.



Model Results

After tuning the model, I ran the model on the test set. The results are listed below.

Confusion Matrix	**Predicted no default**	**Predicted default**
No default	3973	709
Default	275	705

Recall score: 0.7194

Feature Importance

	feature	score
14	PIF_8	0.273711
8	PAY_9	0.225856
9	PAY_8	0.108314
10	PAY_7	0.076283
11	PAY_6	0.067449
0	LIMIT_BAL	0.055806
12	PAY_5	0.045125
15	PIF_7	0.044115
17	PIF_5	0.028379
13	PAY_4	0.025130
16	PIF_6	0.021164
18	PIF_4	0.009774
7	AGE	0.006657
2	GRADUATE_SCHOOL	0.003303
1	SEX	0.002296
3	UNIVERSITY	0.001859
6	SINGLE	0.001832
5	MARRIED	0.001518
4	HIGH_SCHOOL	0.001428

Analysis Results

Based on the feature importances derived from the random forest model, it is clear that the best predictor of default is whether or not clients paid their most recent bill in full. This seems somewhat obvious, but is now confirmed by the model. In regards to the demographic features, it is clear that they do not play as large of a role in predicting default. This makes sense, given that the default percentage between the categories of all of the demographic features were relatively close. Based on this model's feature importances, the demographic factors rank in the following order: age, education, sex, and marital status.

Further Research

This project mainly focused on predicting default based on the payment trends of clients. Below are some ideas for further research with this dataset:

1. Deeper analysis of false negatives generated by the model
2. Develop a model to predict default using only the demographic data
3. Further analysis on relationship of demographics and default