

Predicting Credit Card Default Using Machine Learning

Jesse Mailhot | Mentor: Shmuel Naaman | Springboard Data Science Career Track

1/21/2019

Contents

- ❖ Introduction
- ❖ Dataset
- ❖ Data Wrangling
- ❖ Data Analysis
- ❖ Data Modeling
- ❖ Analysis Results

Introduction

In this project, I aim to solve the following problems:

- ❖ Which features are the best predictor of default?
- ❖ What is the relationship between default and other features?
 - ❖ Marital status, education level, gender, etc...

Dataset

Credit Card Default

- ❖ Demographic data: sex, marital status, education level, age
- ❖ Payment data: credit limit, bill amount, payment amount, payment delay
- ❖ April - September, 2005
- ❖ 30,000 Taiwanese credit card customers.

Source: UCI Machine Learning Repository

Data Wrangling

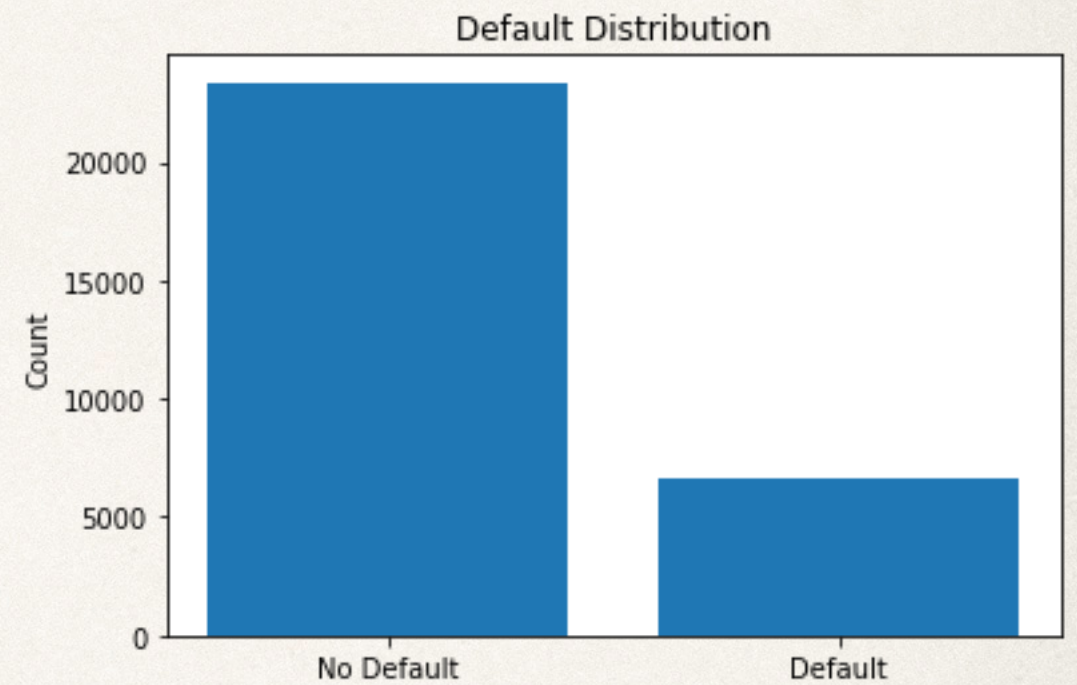
- ❖ Change SEX to 0 & 1
- ❖ Change unlabeled EDUCATION values to 'other'
- ❖ Change unlabeled MARRIAGE values to 'other'
- ❖ Binarize PAY_0 - PAY_6

Data Analysis

- ❖ Investigate each variable in relation to default
- ❖ Look for trends and outliers

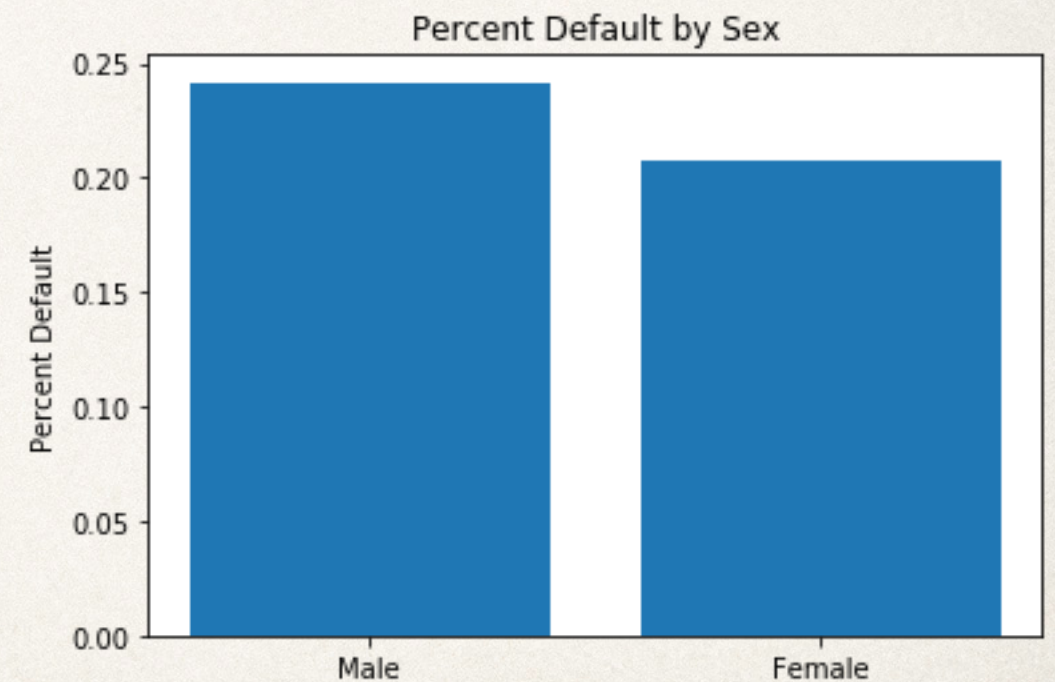
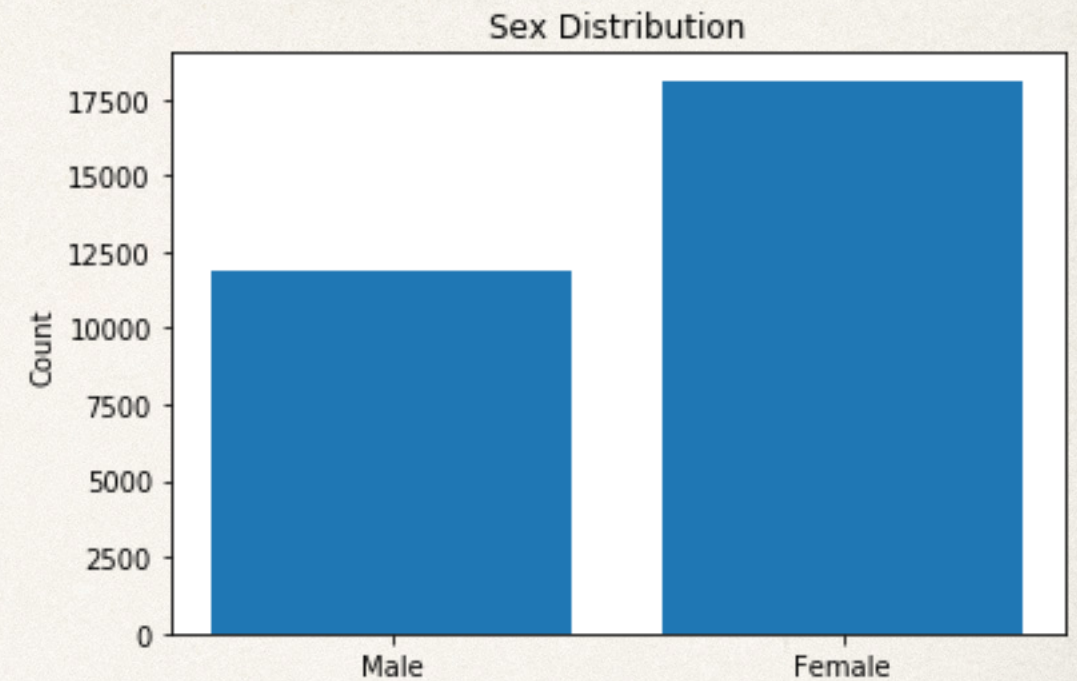
Default

- ❖ Unbalanced dataset
- ❖ 22.12% default percentage



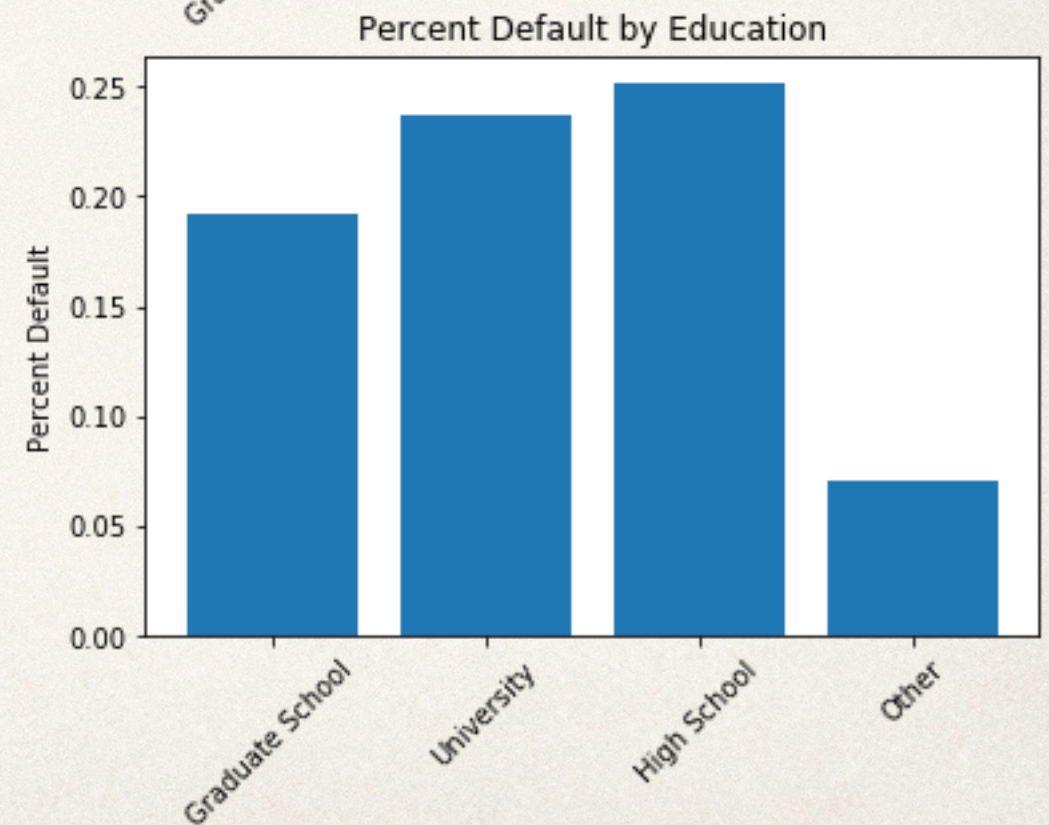
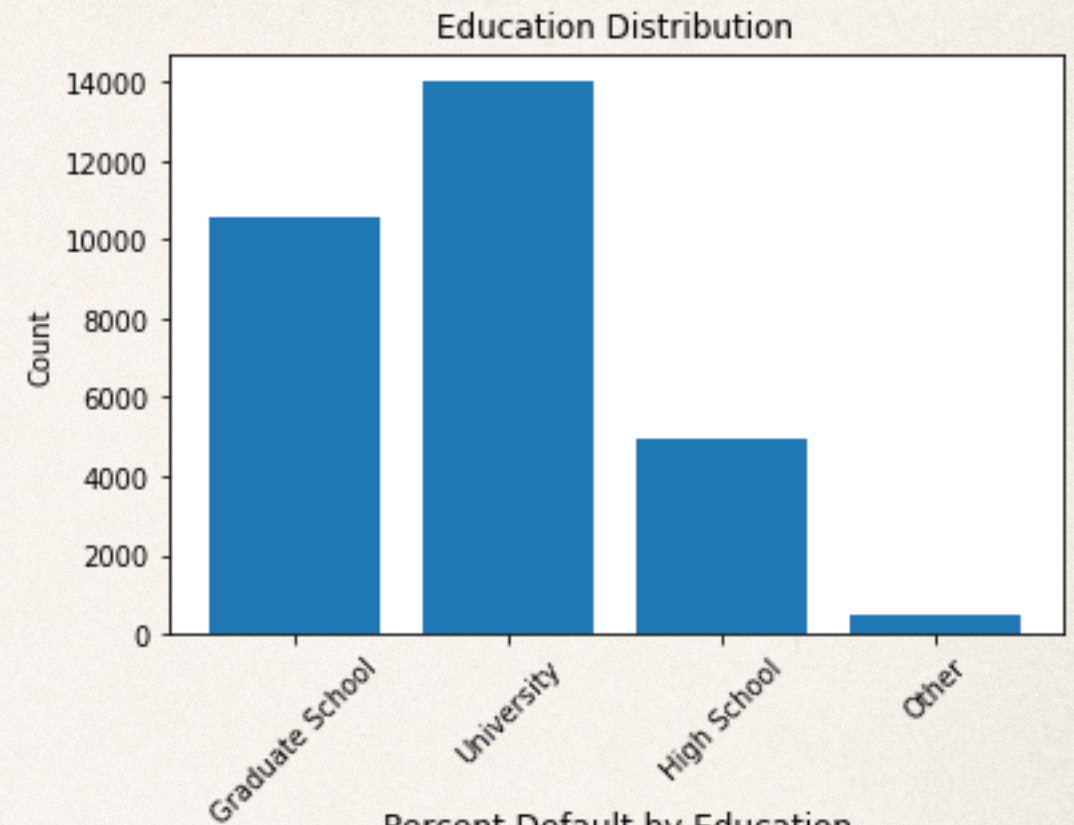
Sex

- ❖ More female clients than male clients
- ❖ Male clients have higher chance of default
- ❖ Chi-squared test results:
 - Test statistic 47.71
 - P-value $4.94e^{-12}$



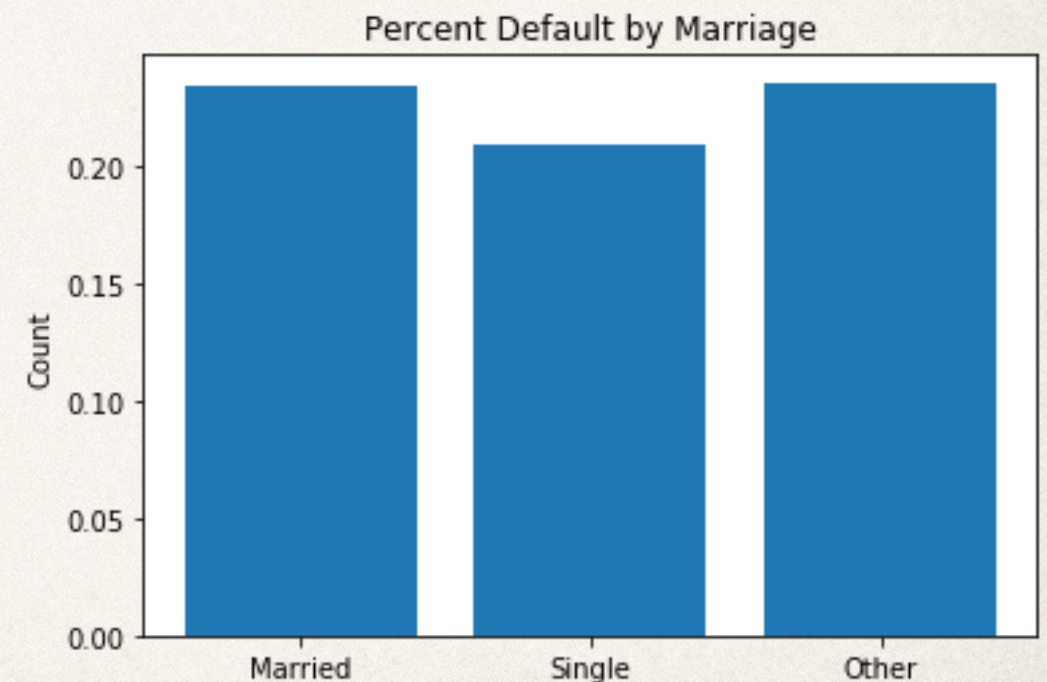
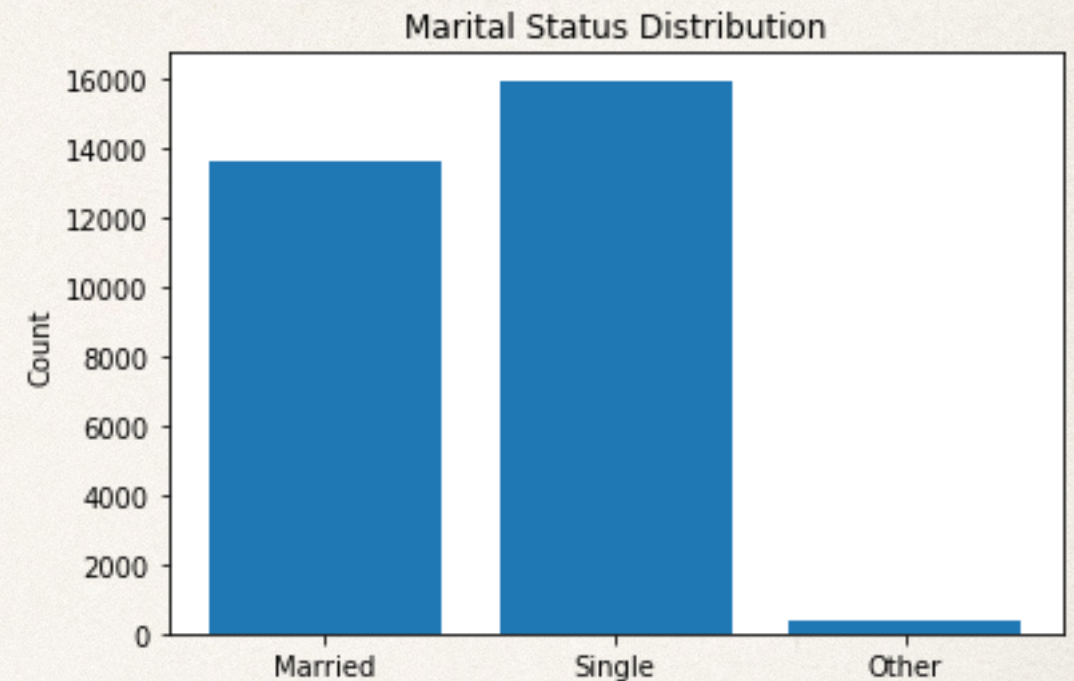
Education

- ❖ Negative correlation between education level and default
- ❖ Low sample size - Other
- ❖ Chi-squared test results:
 - Test statistic 160.41
 - P-value $1.50e^{-34}$



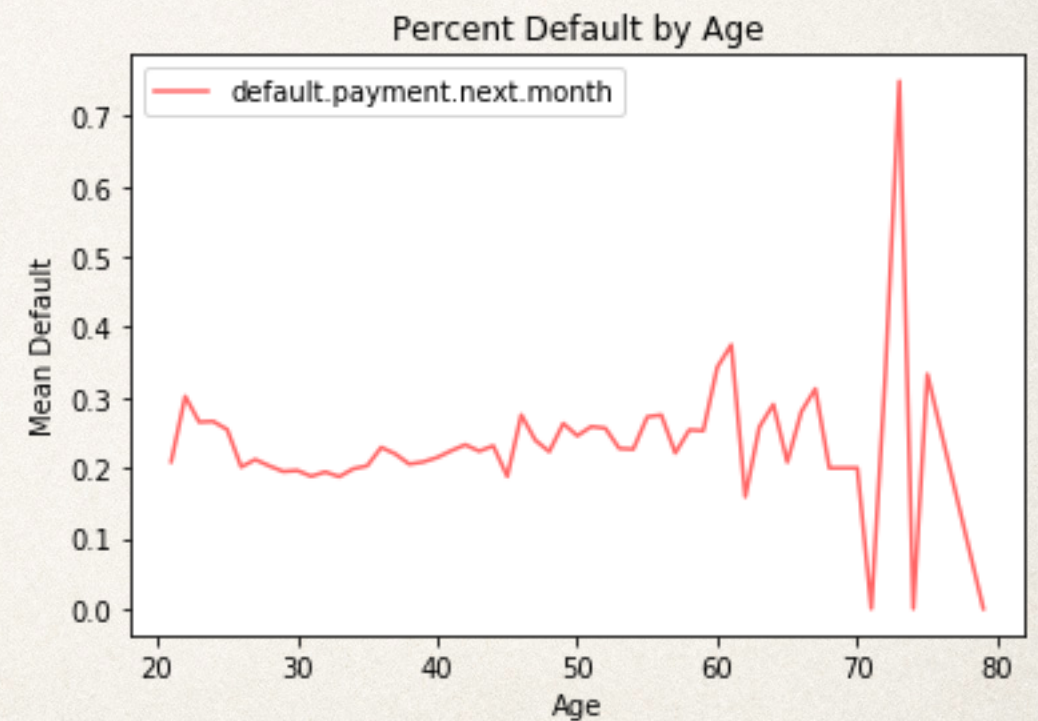
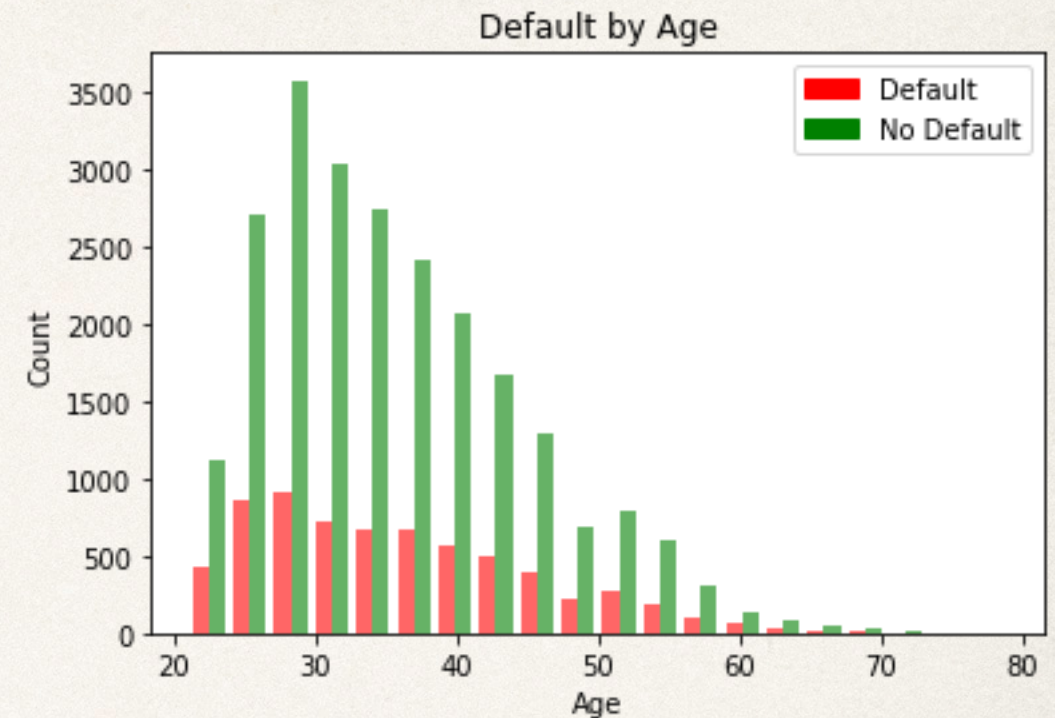
Marriage

- ❖ Single has lower percent default than married
- ❖ Low sample size - Other
 - ❖ Other = divorced?
- ❖ Chi-squared test results:
 - Test statistic 28.13
 - P-value $7.79e^{-7}$



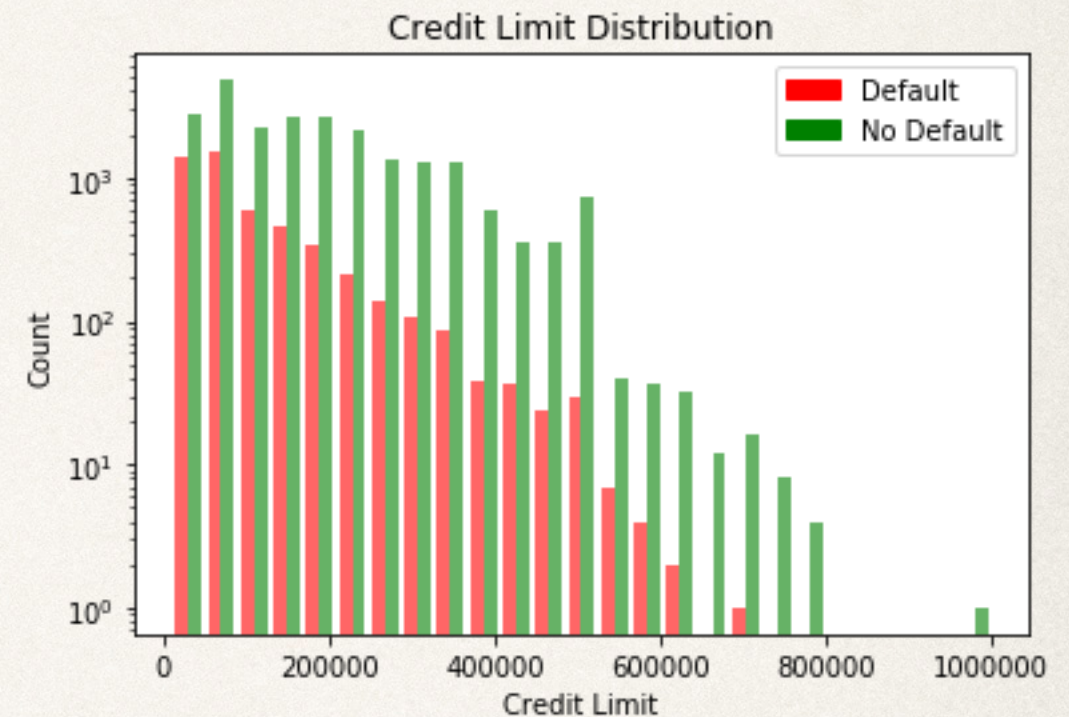
Age

- ❖ Default highest in ages 20-25
- ❖ Decreases after 25
- ❖ Increases after 35
- ❖ Low sample size for ages 50+
- ❖ T-test results:
 - Test statistic -2.41
 - P-value 0.016



Credit Limit

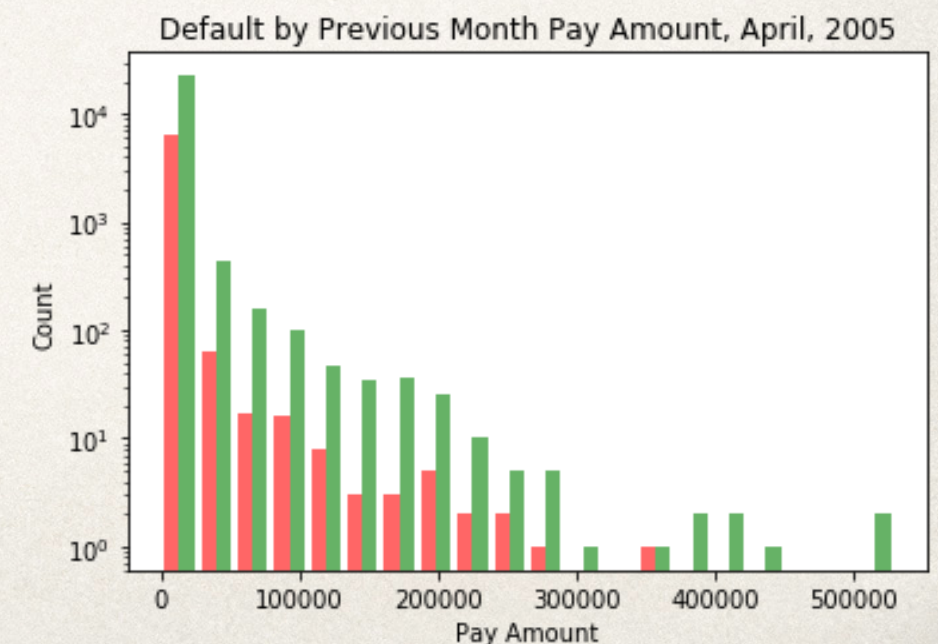
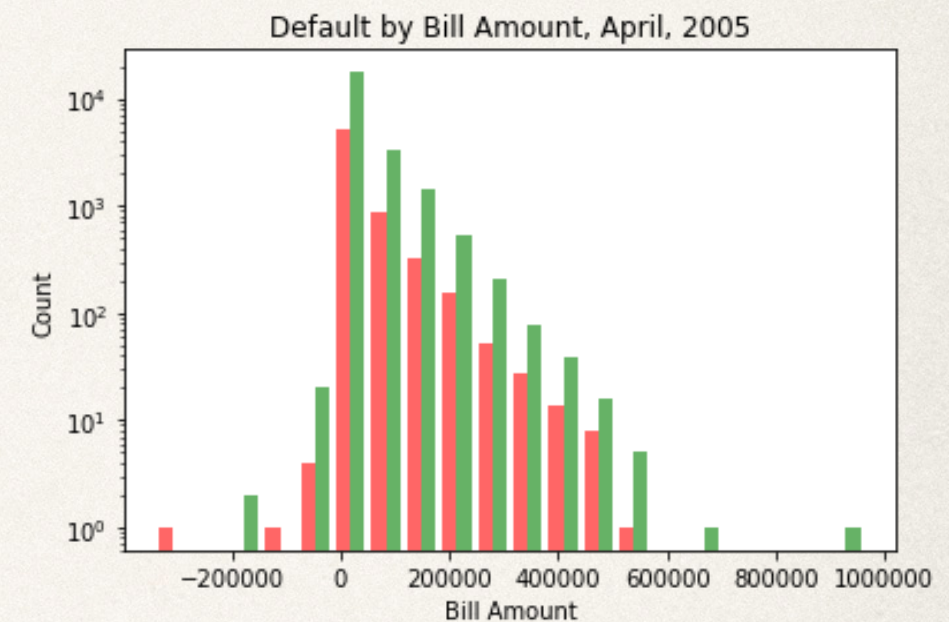
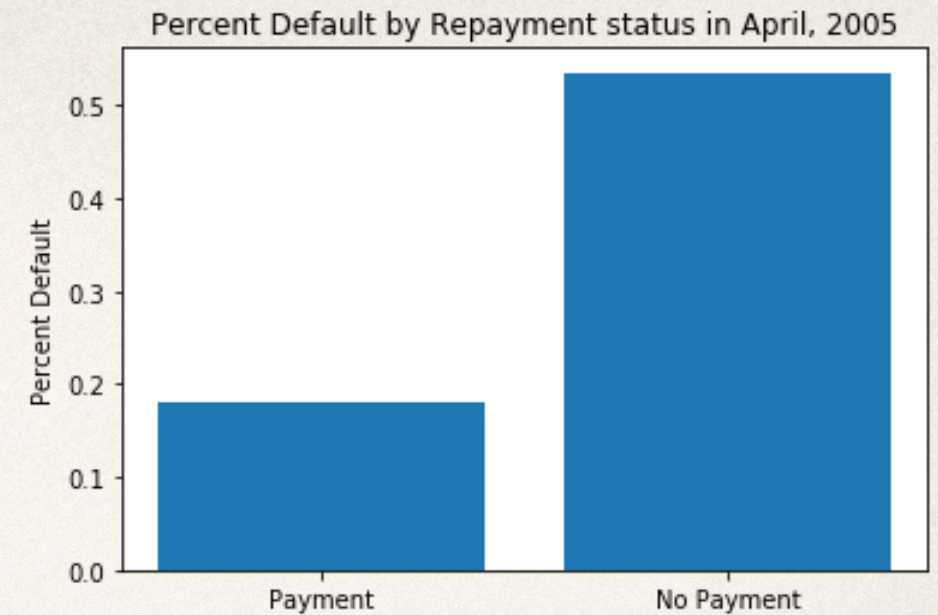
- ❖ Mean credit limit for defaulting clients is NT\$47990 lower than non-defaulting clients
- ❖ T-test results:
 - Test statistic 26.91
 - P-value $1.30e^{-157}$



Pay Data

April, 2005

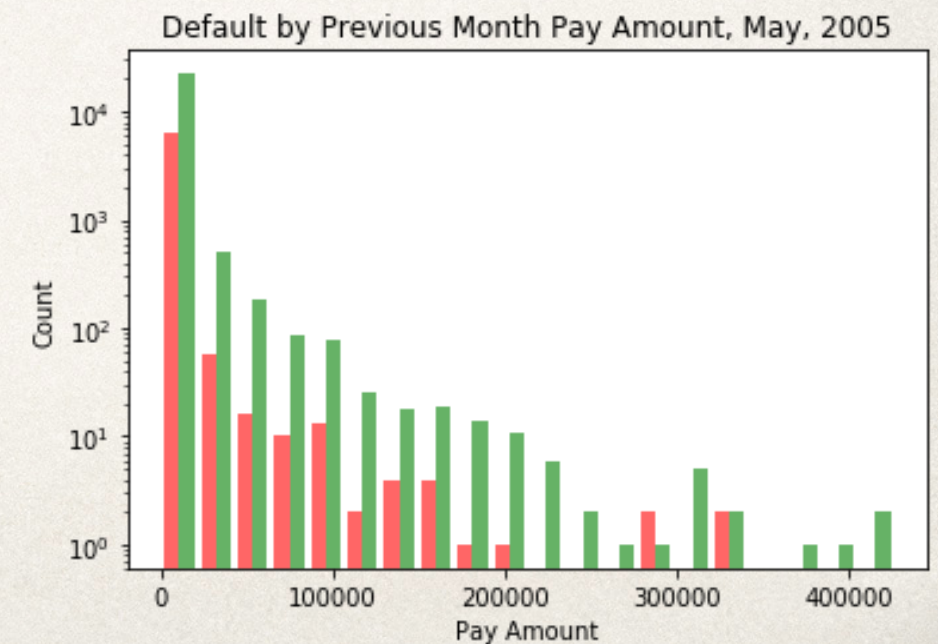
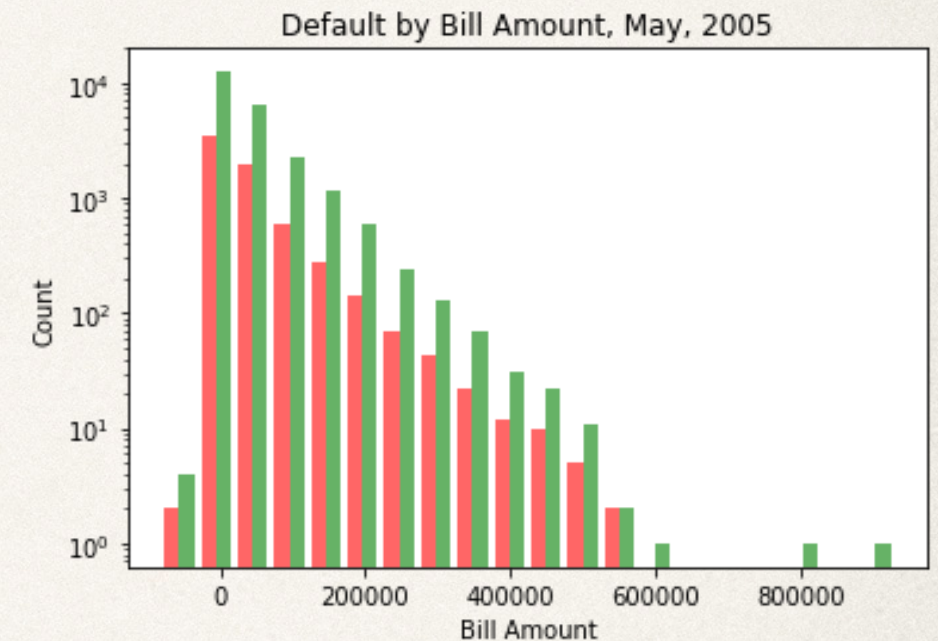
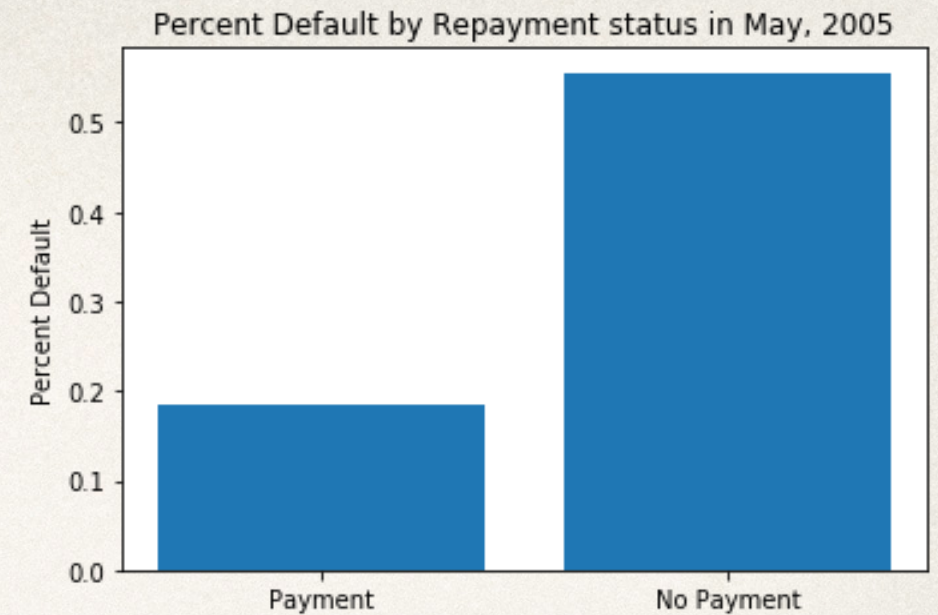
- ❖ 18.67% of paying clients defaulted
- ❖ Clients with bill amount 0 or less defaulted (?)
- ❖ Negative correlation between pay amount and default
- ❖ T-test results (bill amount):
 - Test statistic 0.930
 - P-value 0.352
- ❖ T-test results (pay amount):
 - Test statistic 9.22
 - P-value $3.03e^{-20}$



Pay Data

May, 2005

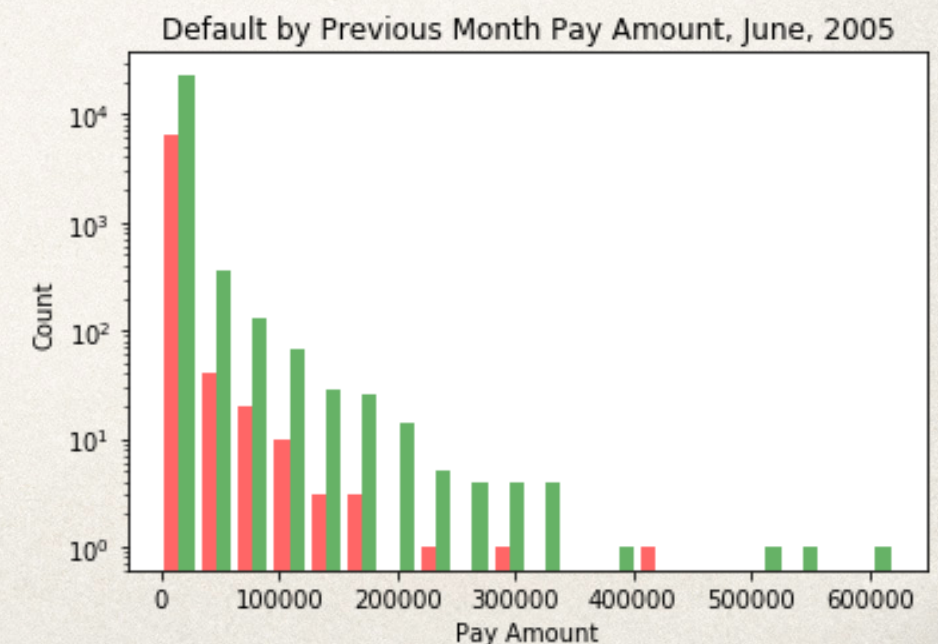
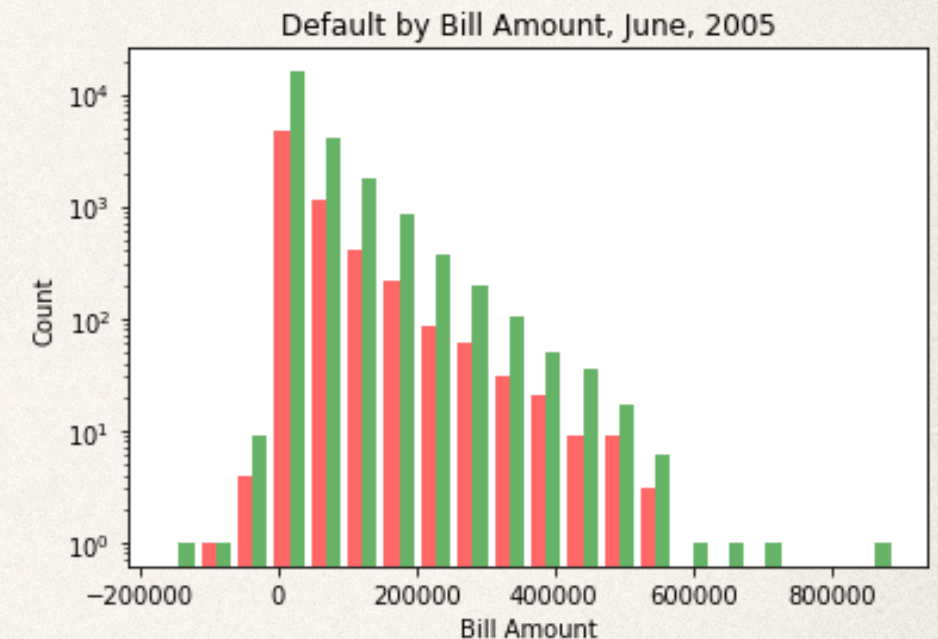
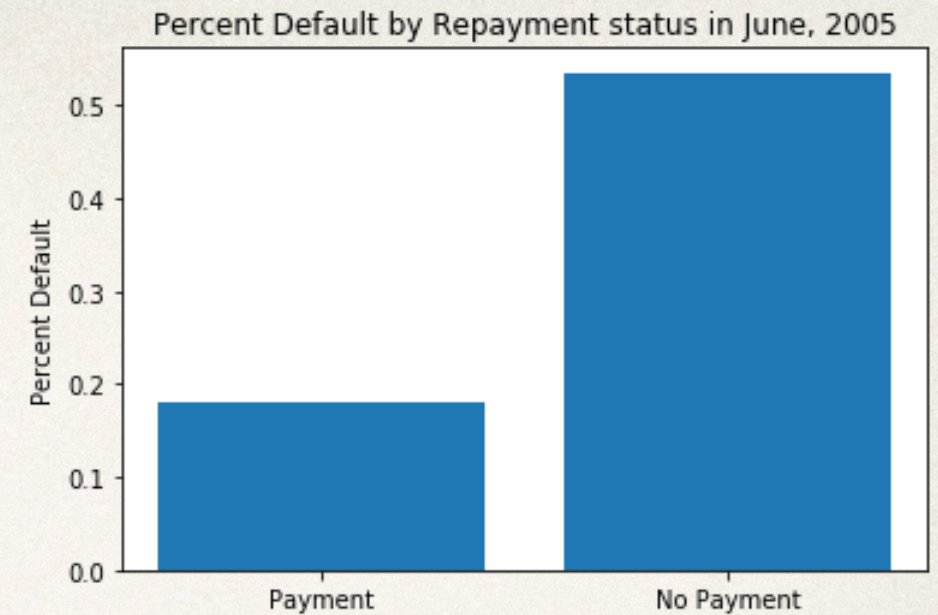
- ❖ 18.45% of paying clients defaulted
- ❖ Clients with bill amount 0 or less defaulted (?)
- ❖ Negative correlation between pay amount and default
- ❖ T-test results (bill amount):
 - Test statistic 1.17
 - P-value 0.241
- ❖ T-test results (pay amount):
 - Test statistic 9.56
 - P-value $1.24e^{-21}$



Pay Data

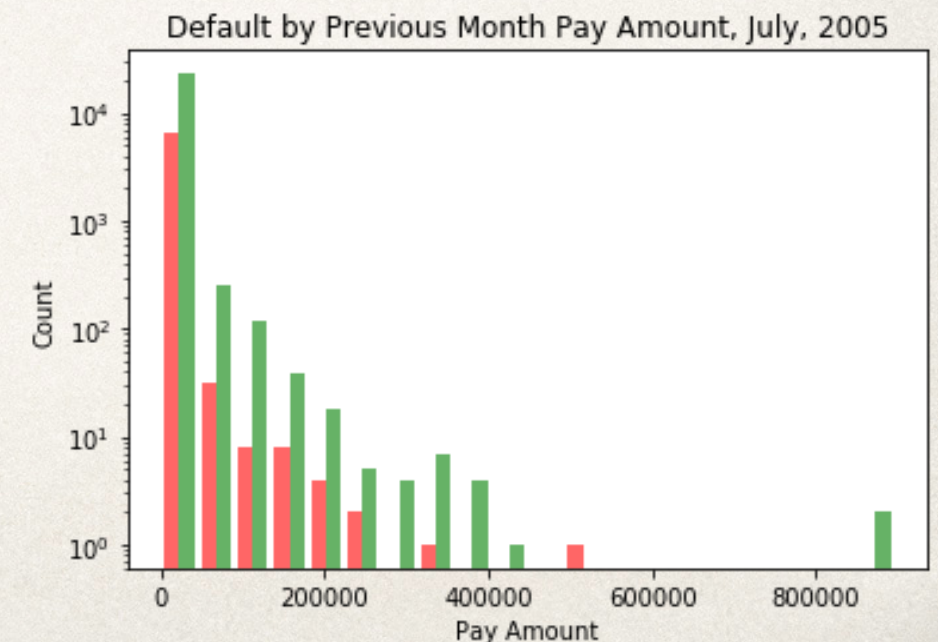
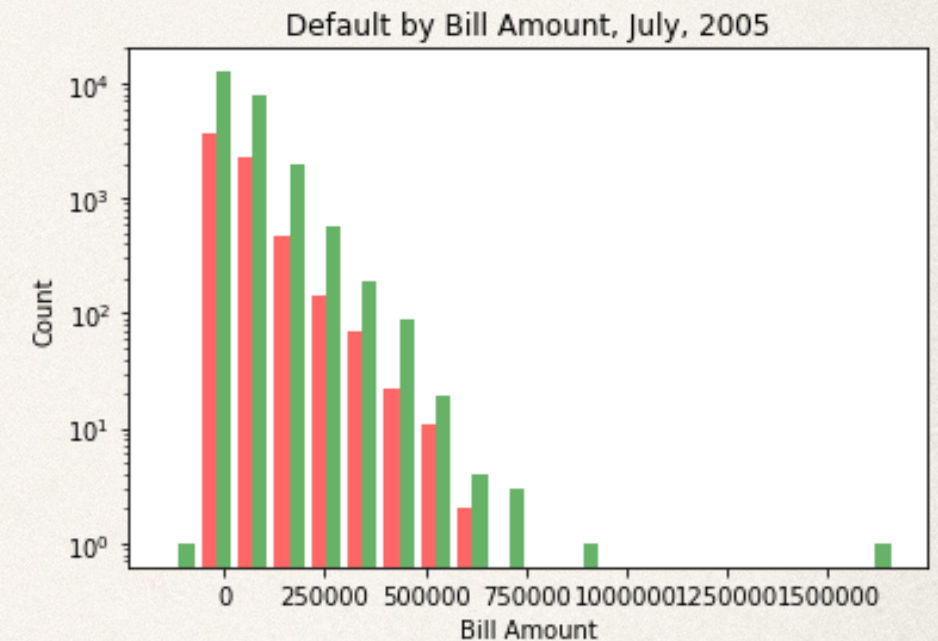
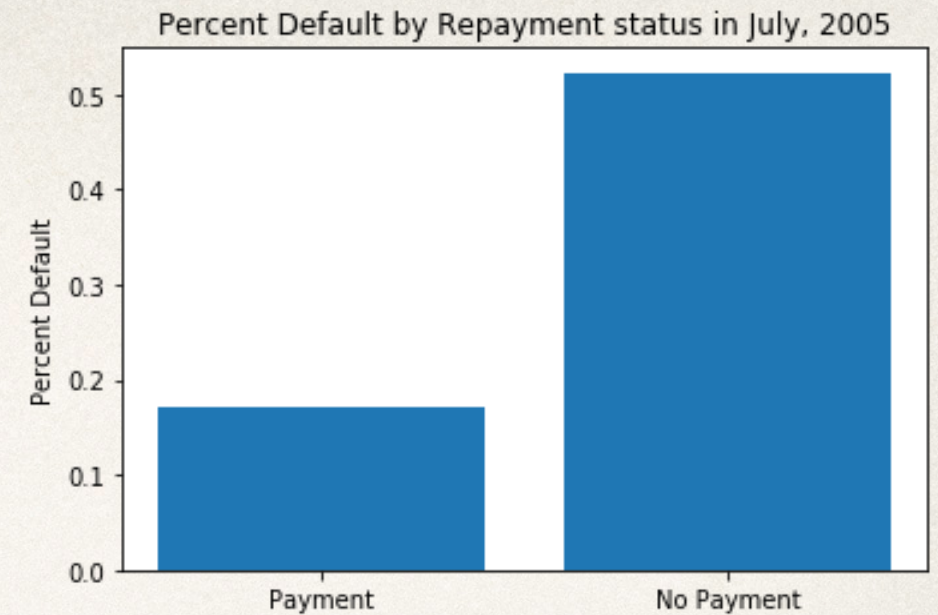
June, 2005

- ❖ 17.95% of paying clients defaulted
- ❖ Clients with bill amount 0 or less defaulted (?)
- ❖ Negative correlation between pay amount and default
- ❖ T-test results (bill amount):
 - Test statistic 1.75
 - P-value 0.079
- ❖ T-test results (pay amount):
 - Test statistic 9.85
 - P-value $6.83e^{-23}$



Pay Data July, 2005

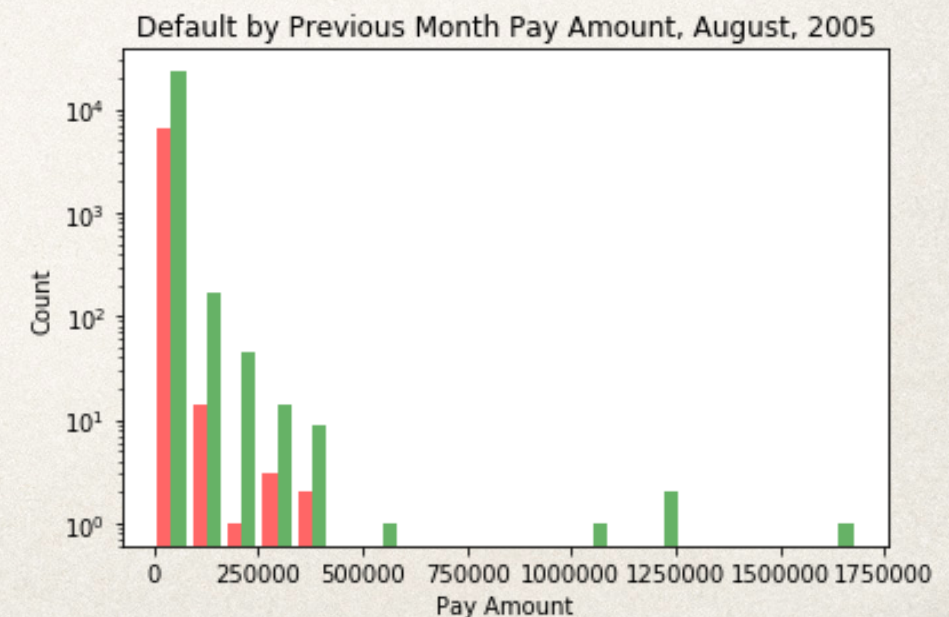
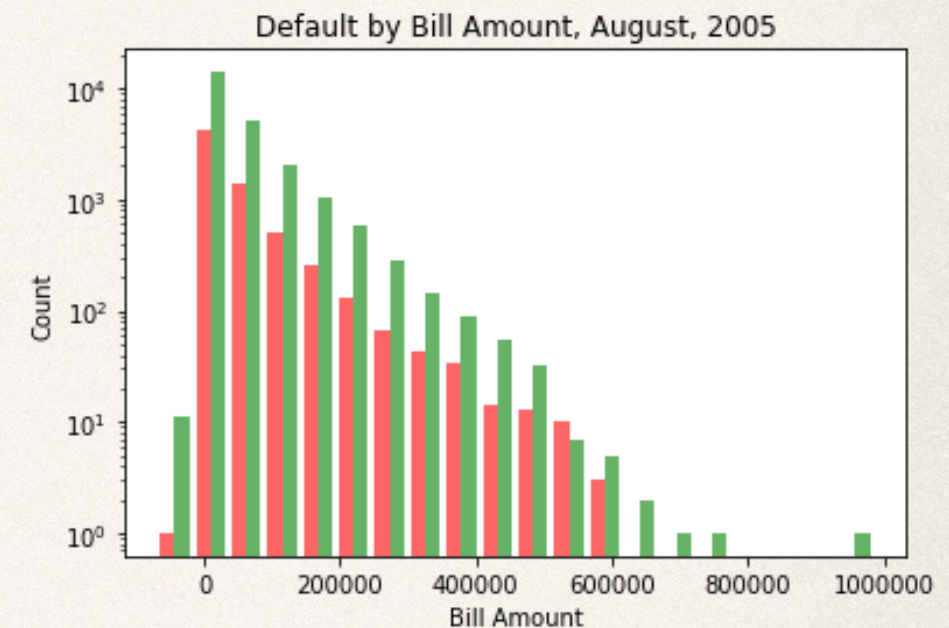
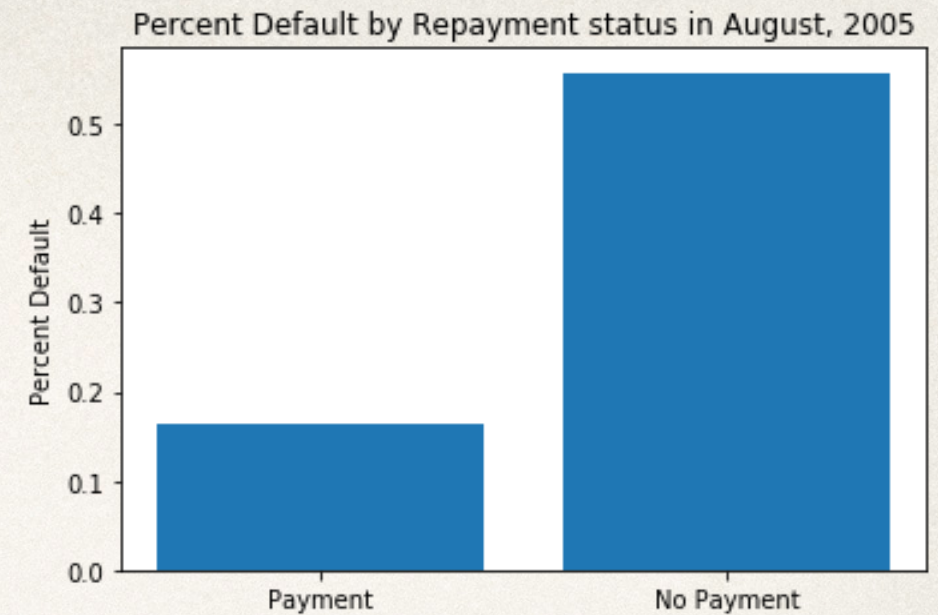
- ❖ 17.19% of paying clients defaulted
- ❖ Clients with bill amount 0 or less defaulted (?)
- ❖ Negative correlation between pay amount and default
- ❖ T-test results (bill amount):
 - Test statistic 2.43
 - P-value 0.015
- ❖ T-test results (pay amount):
 - Test statistic 9.75
 - P-value $1.84e^{-22}$



Pay Data

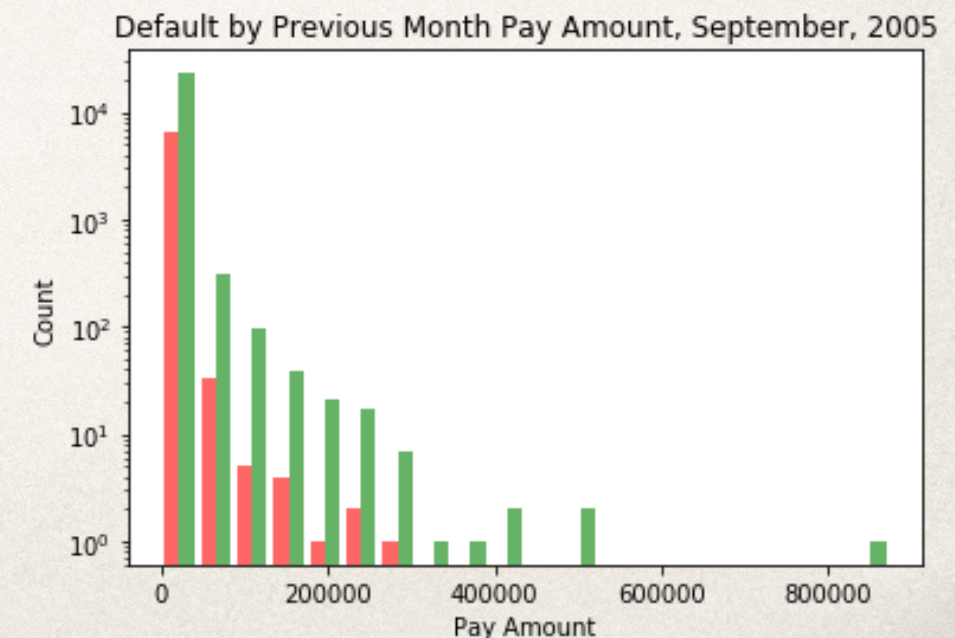
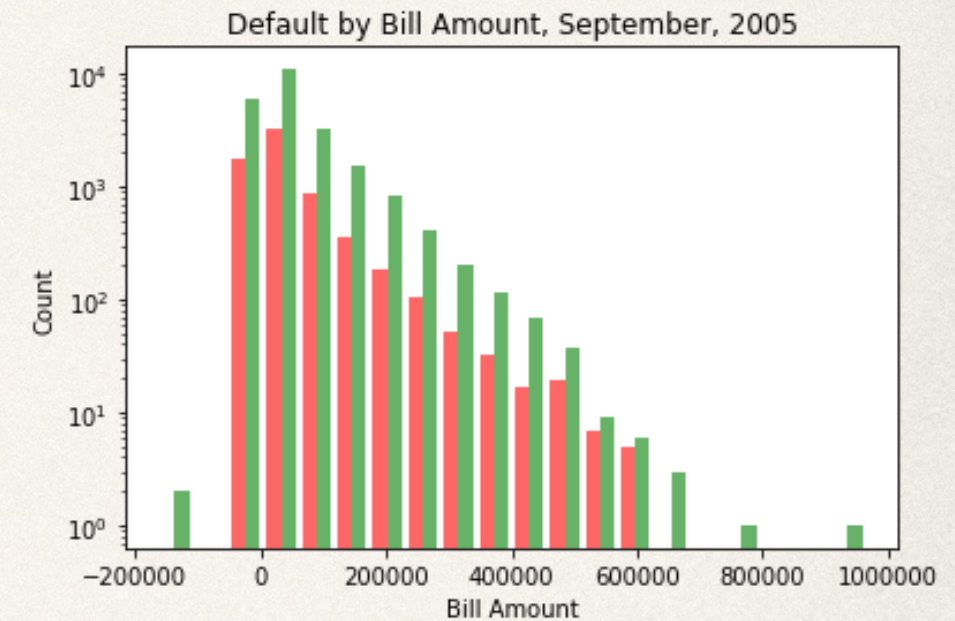
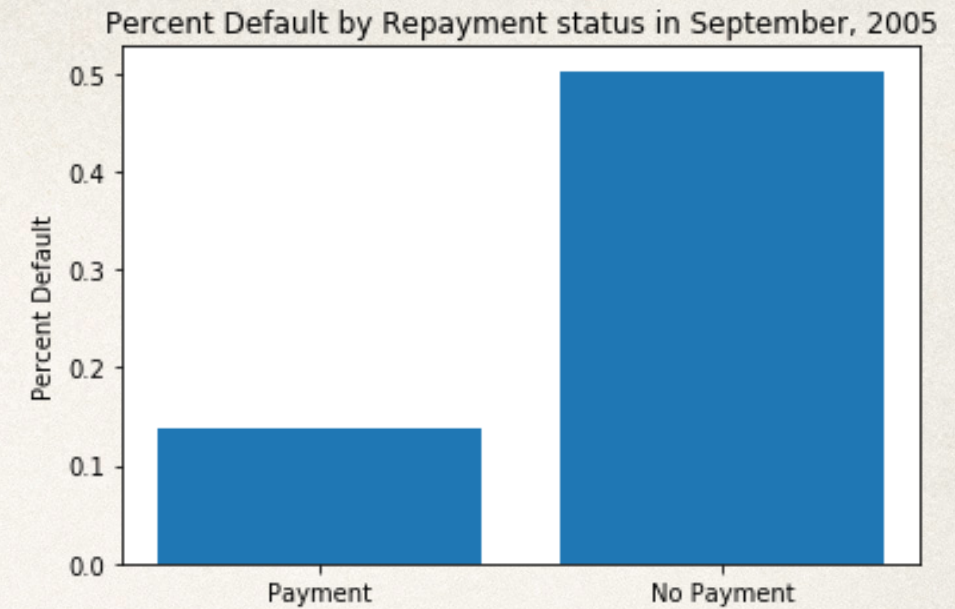
August, 2005

- ❖ 16.27% of paying clients defaulted
- ❖ Clients with bill amount 0 or less defaulted (?)
- ❖ Negative correlation between pay amount and default
- ❖ T-test results (bill amount):
 - Test statistic 2.46
 - P-value 0.014
- ❖ T-test results (pay amount):
 - Test statistic 10.16
 - P-value $3.17e^{-24}$



Pay Data September, 2005

- ❖ 13.83% of paying clients defaulted
- ❖ Clients with bill amount 0 or less defaulted (?)
- ❖ Negative correlation between pay amount and default
- ❖ T-test results (bill amount):
 - Test statistic 3.40
 - P-value 0.001
- ❖ T-test results (pay amount):
 - Test statistic 12.66
 - P-value 1.15^{-36}



Data Modeling

- ❖ Engineer 'PIF' feature
- ❖ Split categorical columns into dummy variables
- ❖ Feature selection
- ❖ Model selection
- ❖ Metric selection
- ❖ Model tuning

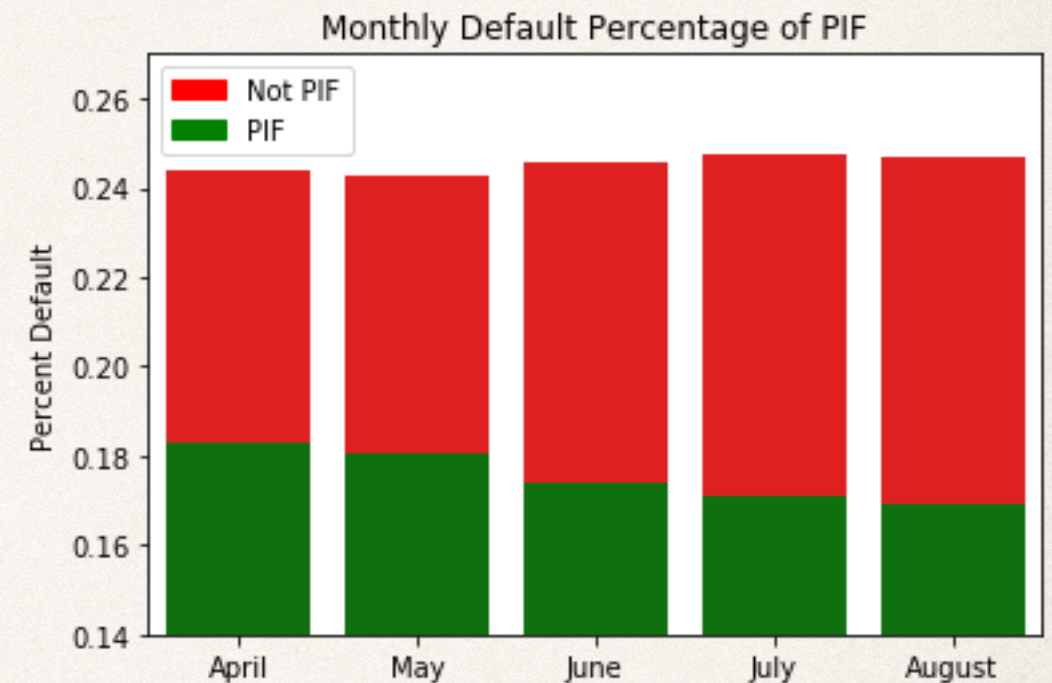
Engineering 'PIF'

- ❖ $\text{PAY_AMT}(N) \geq \text{BILL_AMT}(N - 1)$
- ❖ August to April

EDA for PIF

Columns

- ❖ 16.92% of clients who paid their August bill in full defaulted
- ❖ Negative balances
- ❖ Closed accounts labeled as default?



Feature Selection

- ❖ Drop clients who closed their accounts
 - ❖ *PIF8 = True & default = True*
- ❖ Exclude bill amount, pay amount columns
 - ❖ Described by 'PIF'

Split categorical columns

- ❖ Marriage, Education
- ❖ Pandas *get_dummies*

Model Selection

- ❖ SciKit Learn
 - ❖ Decision Tree
 - ❖ Support Vector Machine
 - ❖ Random Forest
 - ❖ AdaBoost

Metric Selection

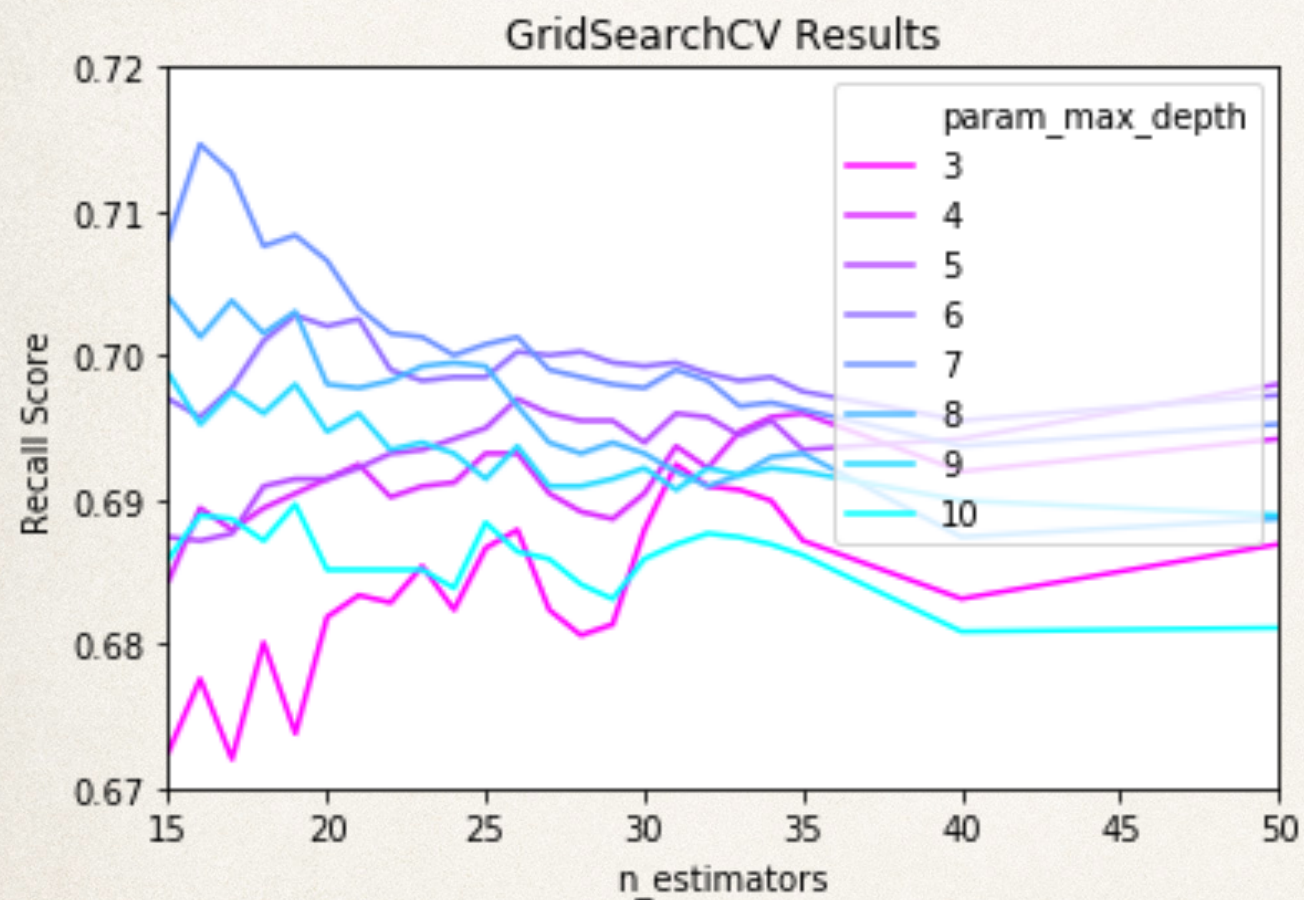
Goal: save the bank money

- ❖ Accuracy: $(TP + TN) / (FP + FN)$
- ❖ Precision: $TP / (TP + FP)$
- ❖ Recall: $TP / (TP + FN)$
- ❖ F1: $2 * (precision * recall) / (precision + recall)$

Model Tuning

- ❖ SciKit Learn *GridSearchCV*
- ❖ Maximize recall score
- ❖ `max_depth` & `n_estimators`

Model Tuning



	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

Model Results

Confusion Matrix		
	Predicted No Default	Predicted Default
No Default	3973	709
Default	275	705

Recall Score: 0.7194

Analysis Results

- ❖ Payment features best indicator of default
- ❖ Demographic features not as clear
 - ❖ Married clients are more likely to default than single
 - ❖ Males more likely to default than females
 - ❖ Default is highest for customers age 20-25

Further Research

- ❖ Dive deeper into false negatives data
- ❖ Try using only demographic data
- ❖ Further analysis on relationship between demographics and default