Predicting Credit Card Default Using Machine Learning

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

1/21/2019

Contents

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

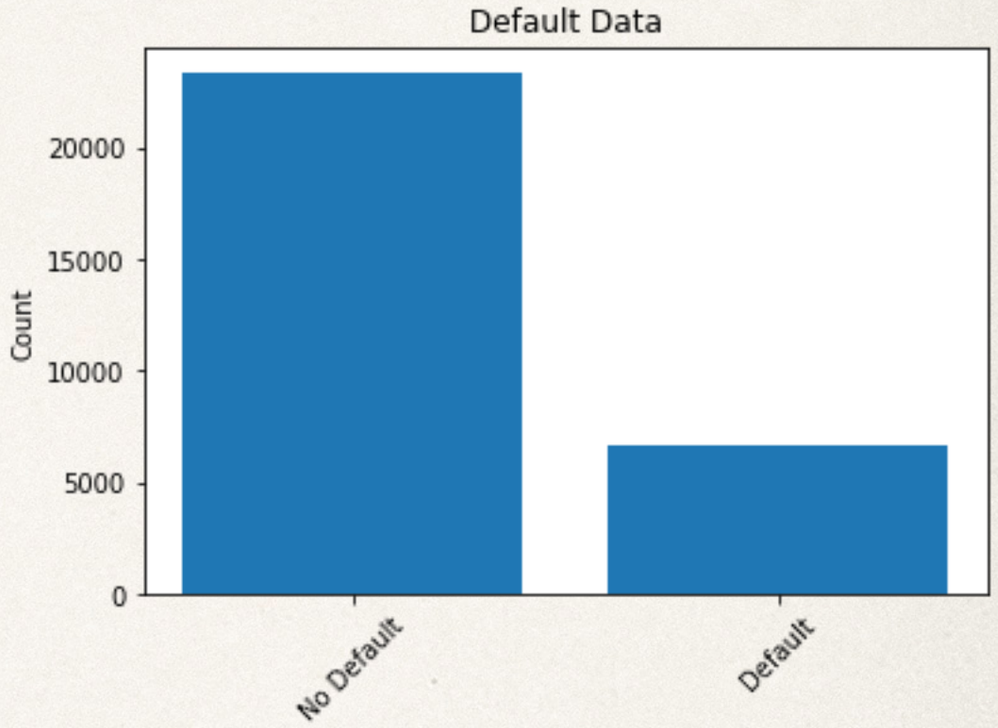
- Change SEX to 0 & 1
- Change unlabeled EDUCATION values to ‘other’
- Change unlabeled MARRIAGE values to ‘other’
- Binarize PAY_0 - PAY_6

Data Analysis

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

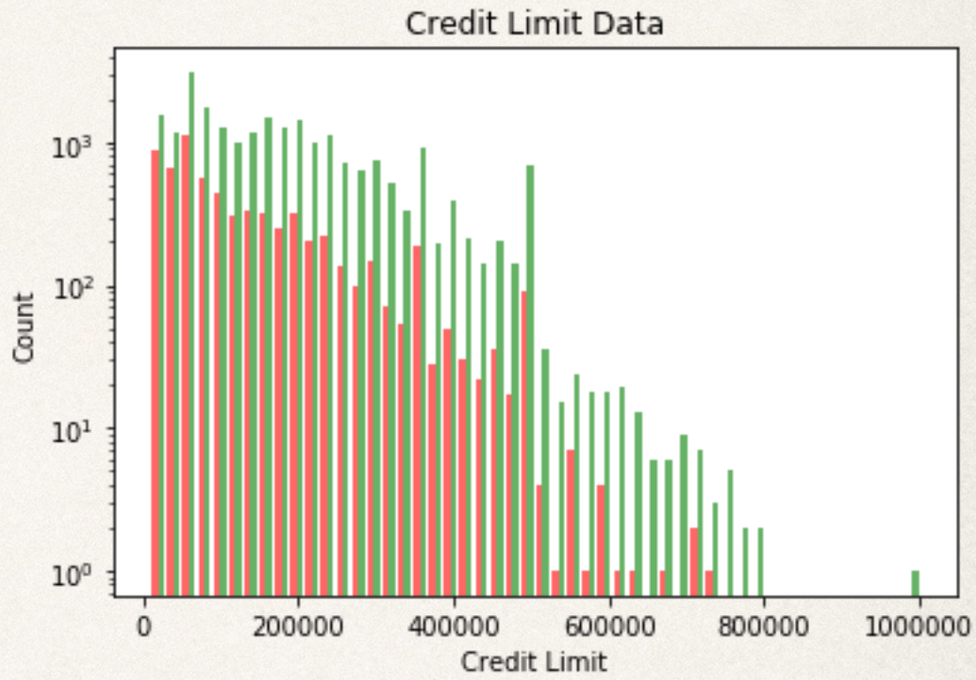
Default

- Unbalanced dataset
- 22.12% default percentage

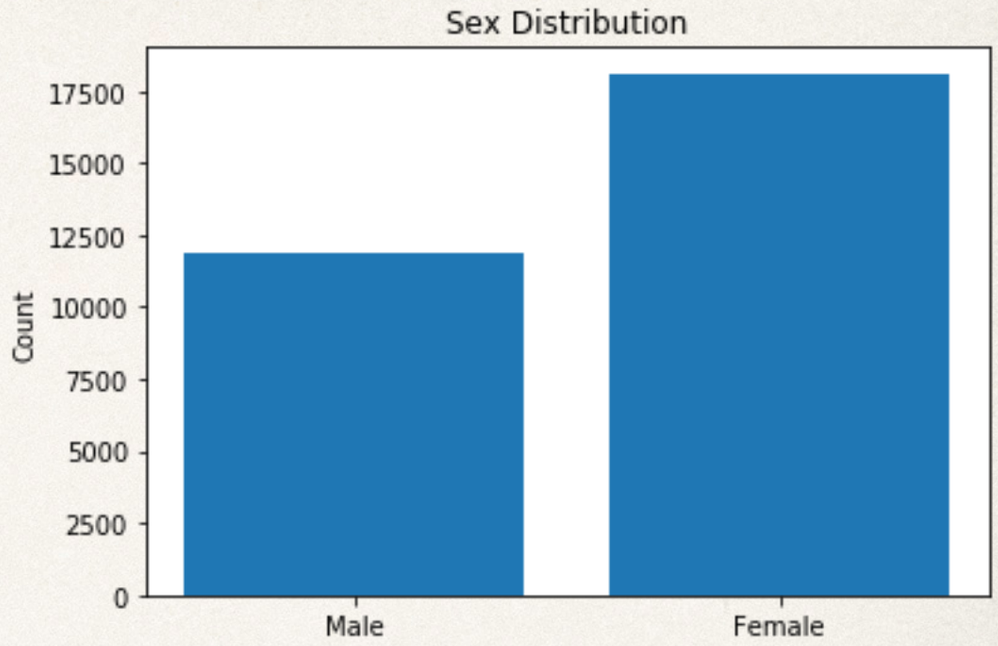


Credit Limit

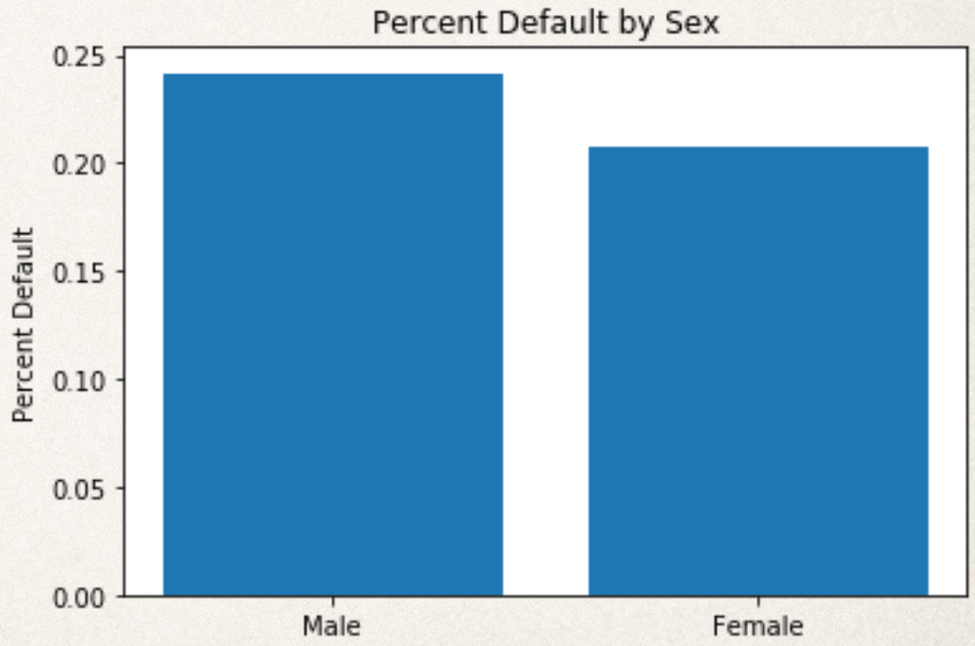
- Mean credit limit for defaulting clients is NT$47990 lower than non-defaulting clients



Sex

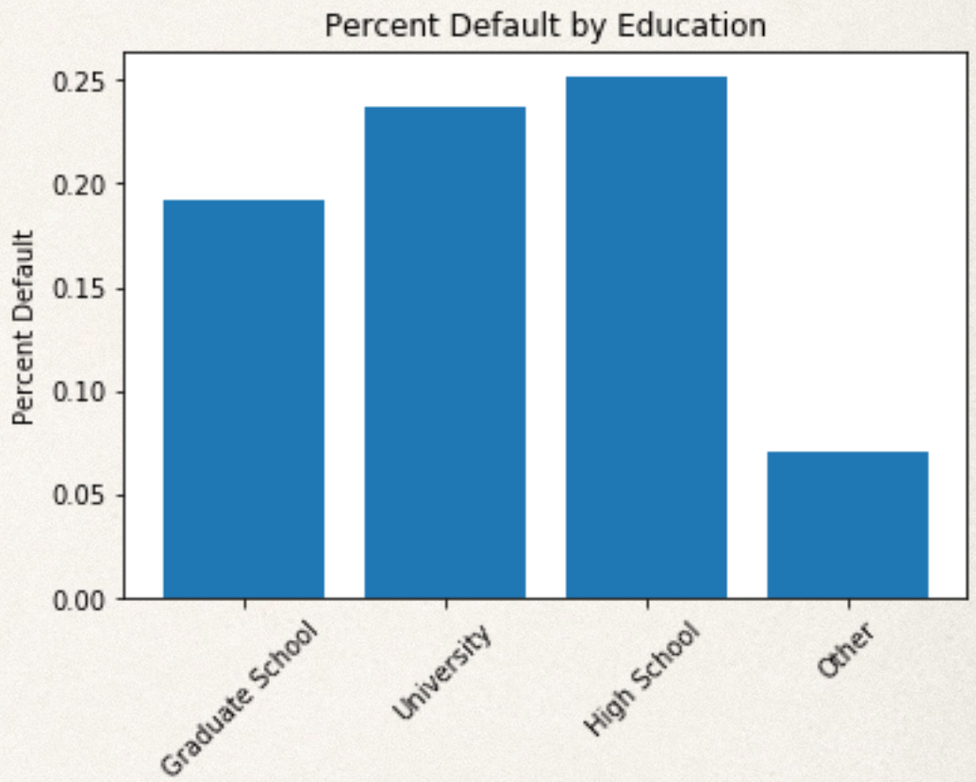


- More female clients than male clients
- Male clients have higher chance of default



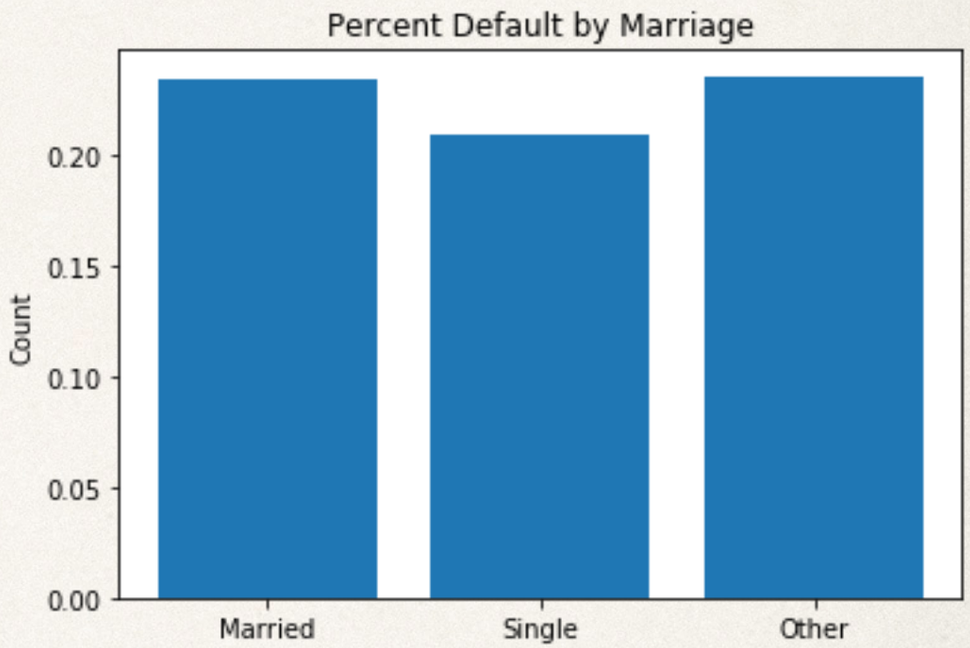
Education

- Negative correlation between education level and default
- Low sample size - Other

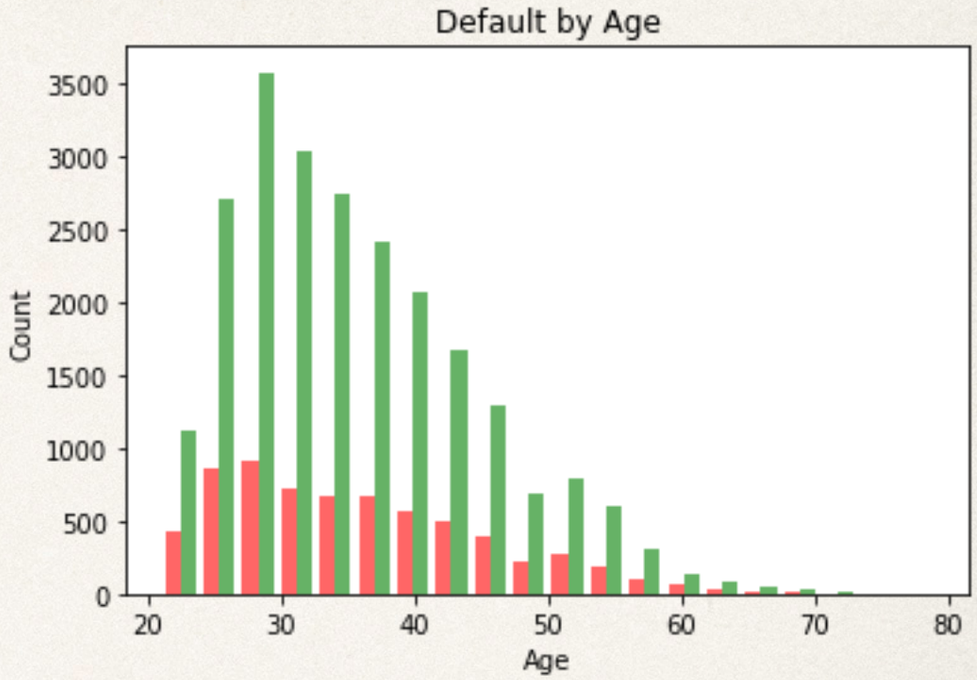


Marriage

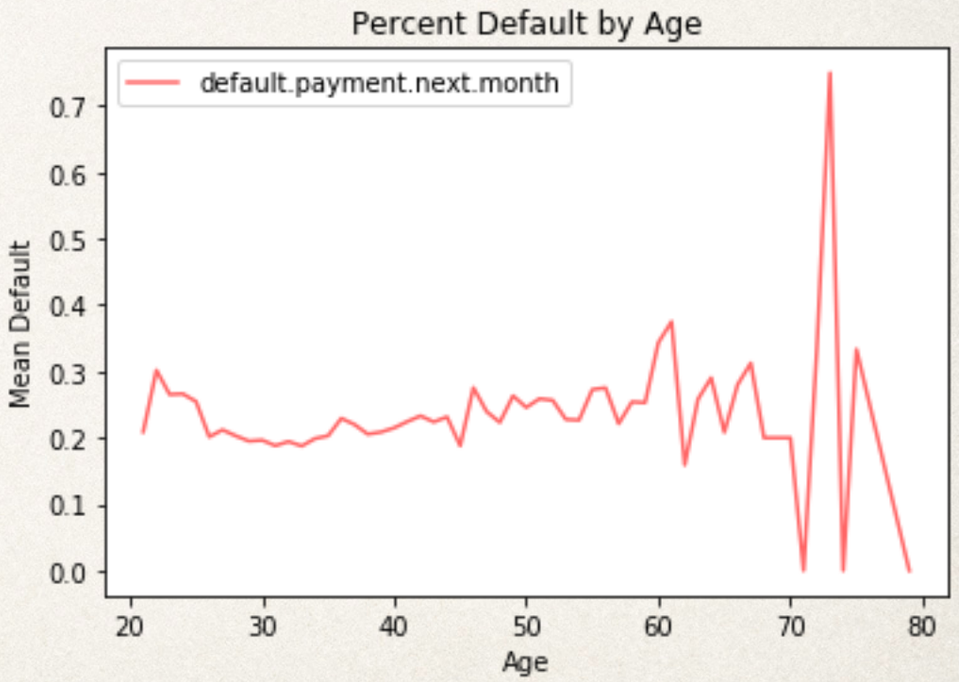
- Single has lower percent default than married
- Low sample size - Other
  - Other = divorced?



Age

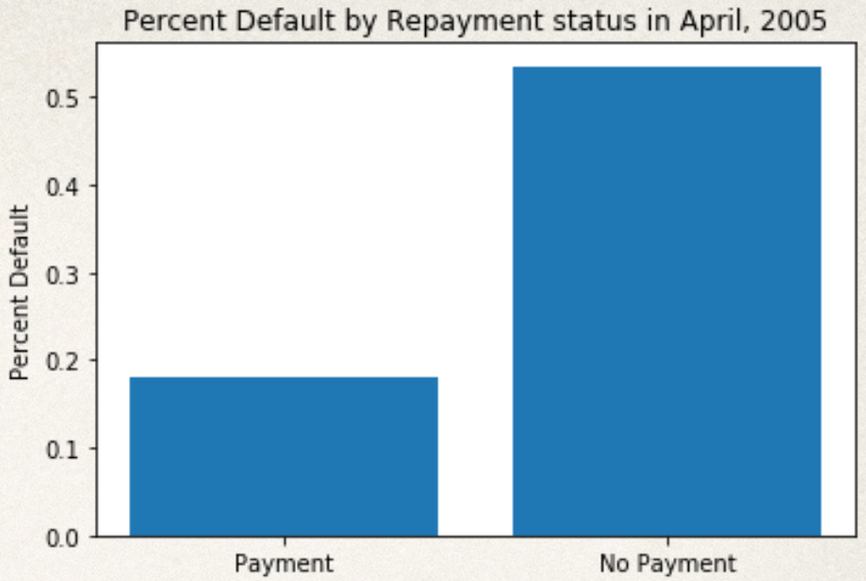


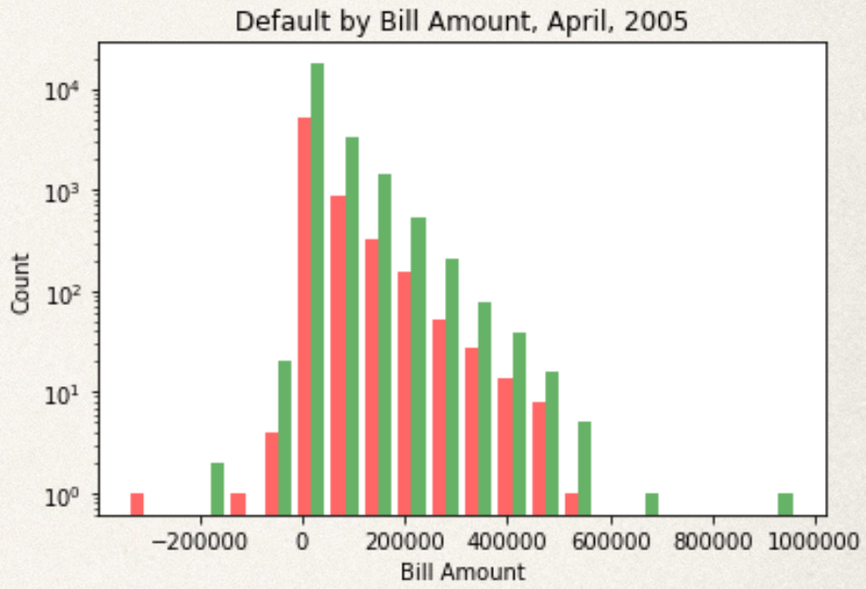
- Default highest in ages 20-25
  - Decreases after 25
  - Increases after 35
- Low sample size for ages 50+

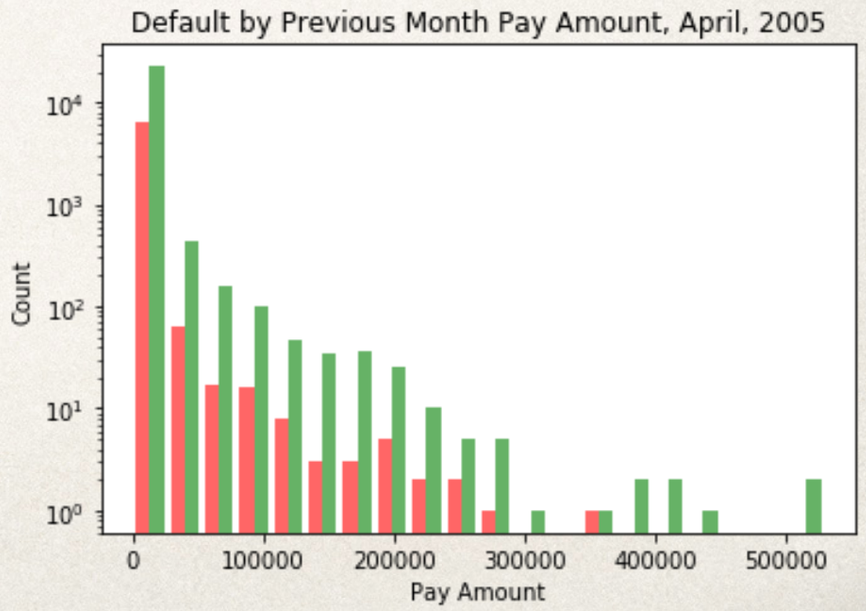


Pay Data
April, 2005

- Clients who made payments still defaulted (?)
- Clients with bill amount 0 or less defaulted (?)
- Negative correlation between pay amount and default

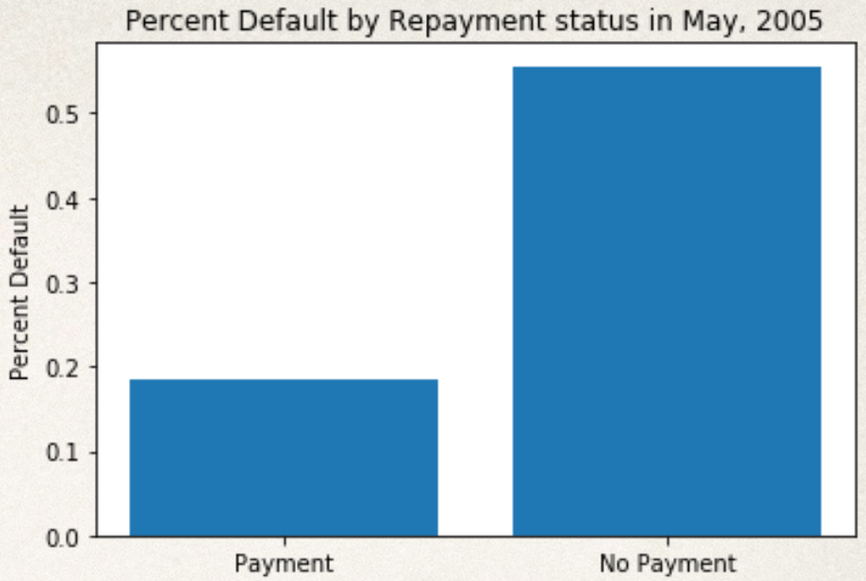


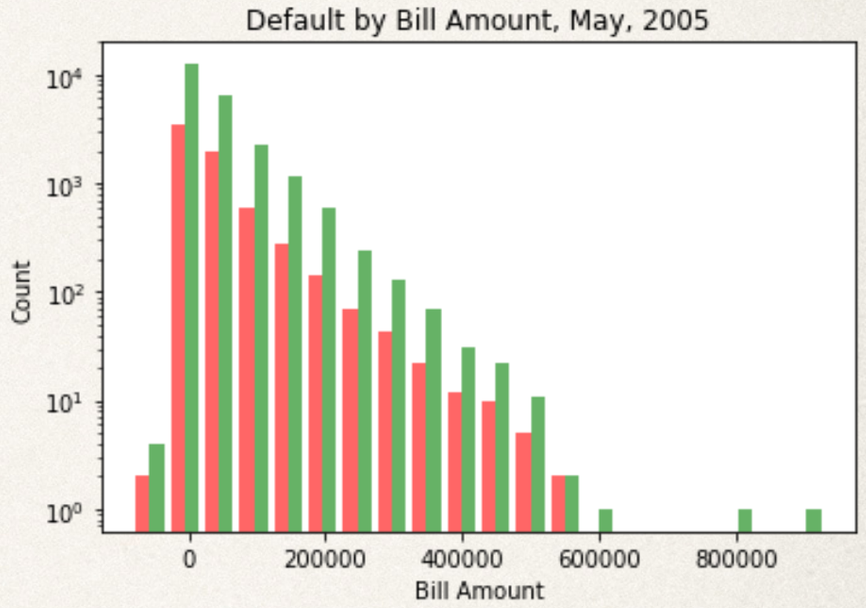


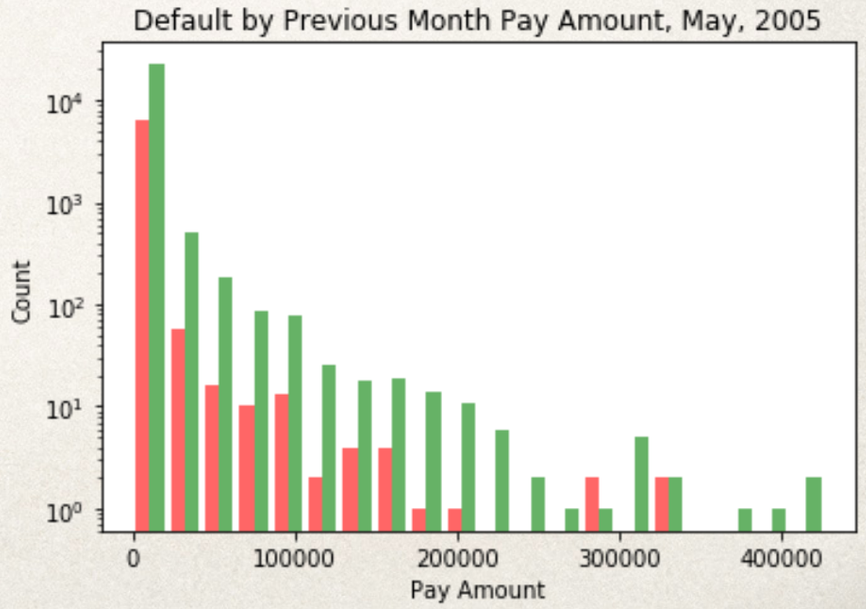


Pay Data May, 2005

- Clients who made payments still defaulted (?)
- Clients with bill amount 0 or less defaulted (?)
- Negative correlation between pay amount and default

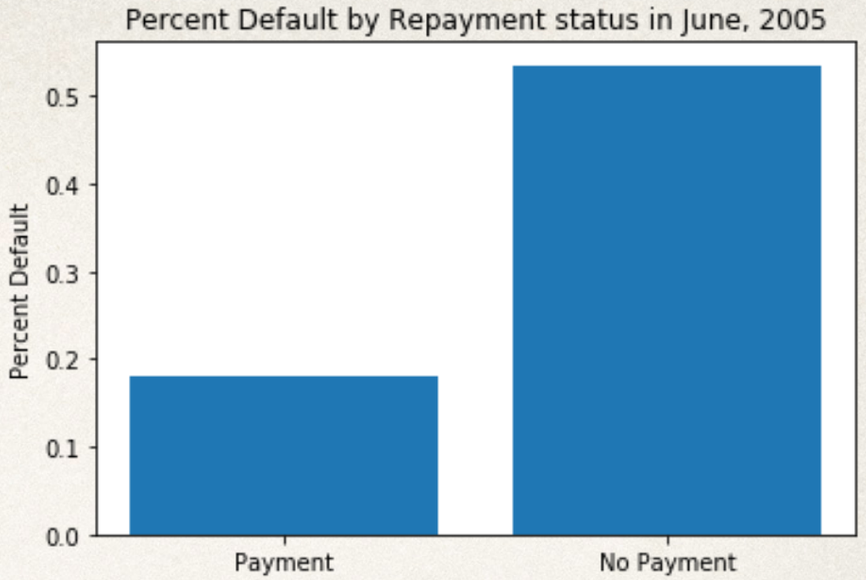


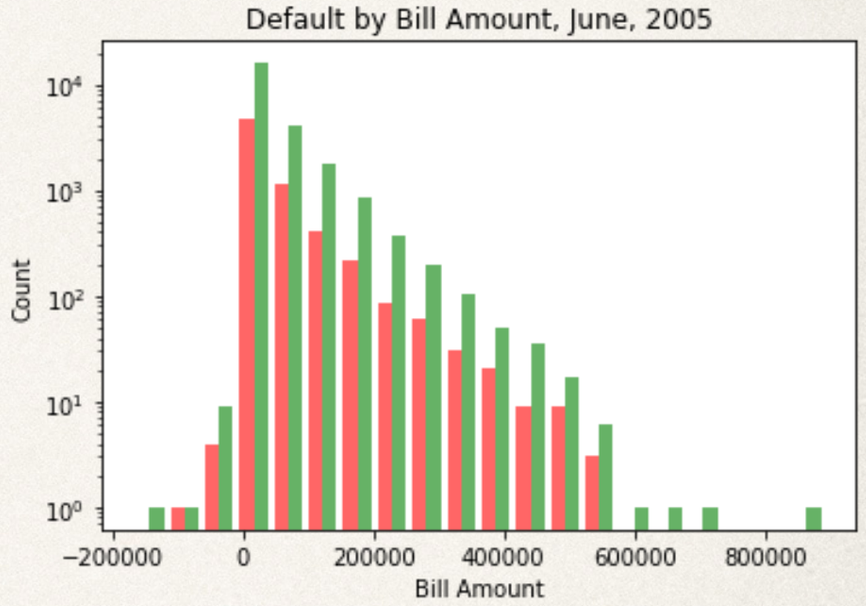


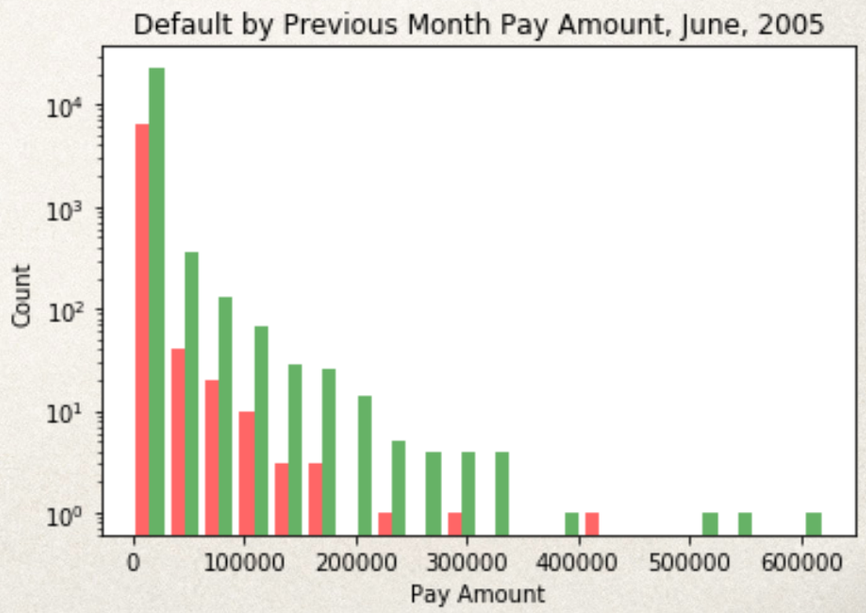


Pay Data June, 2005

- Clients who made payments still defaulted (?)
- Clients with bill amount 0 or less defaulted (?)
- Negative correlation between pay amount and default

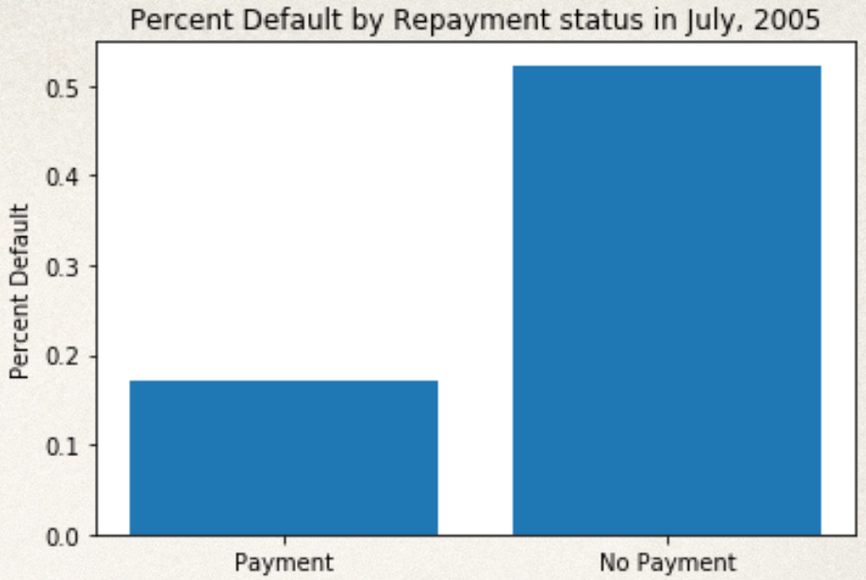


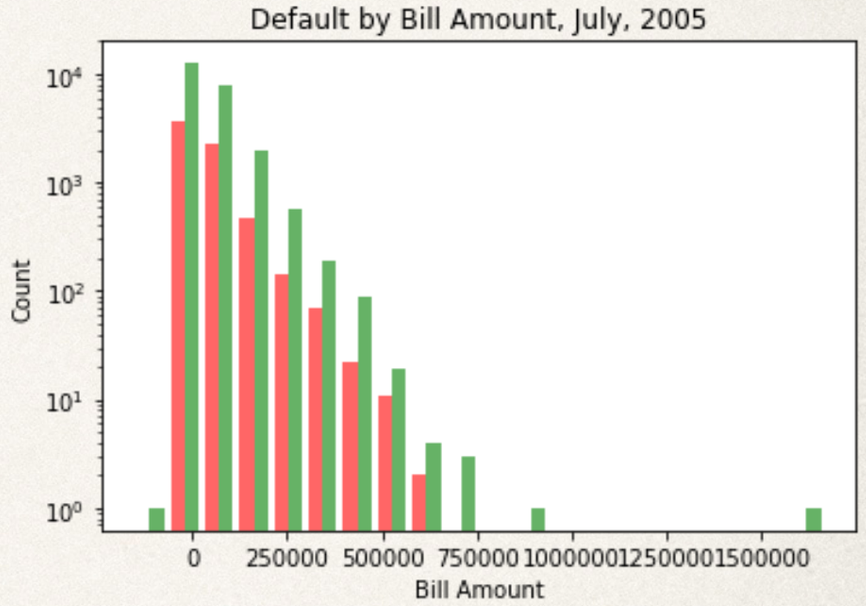


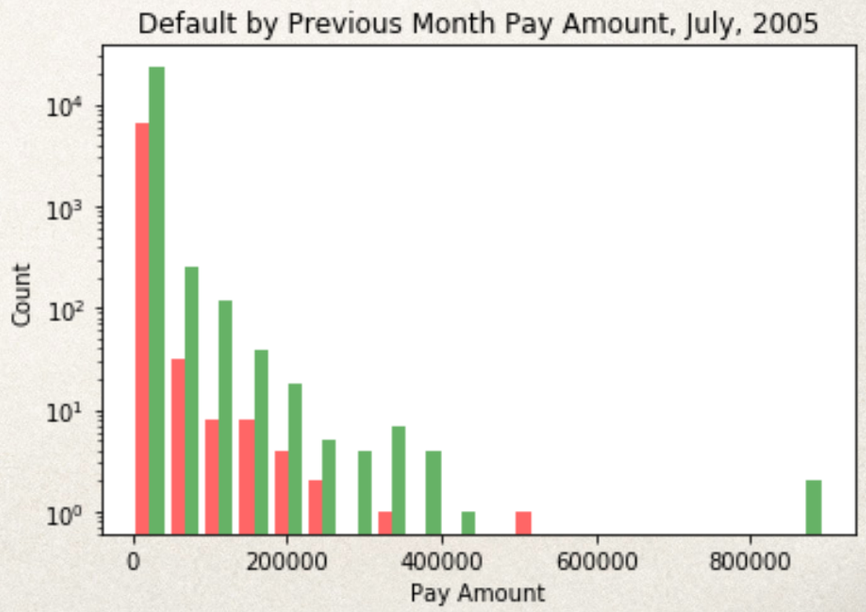


Pay Data July, 2005

- Clients who made payments still defaulted (?)
- Clients with bill amount 0 or less defaulted (?)
- Negative correlation between pay amount and default

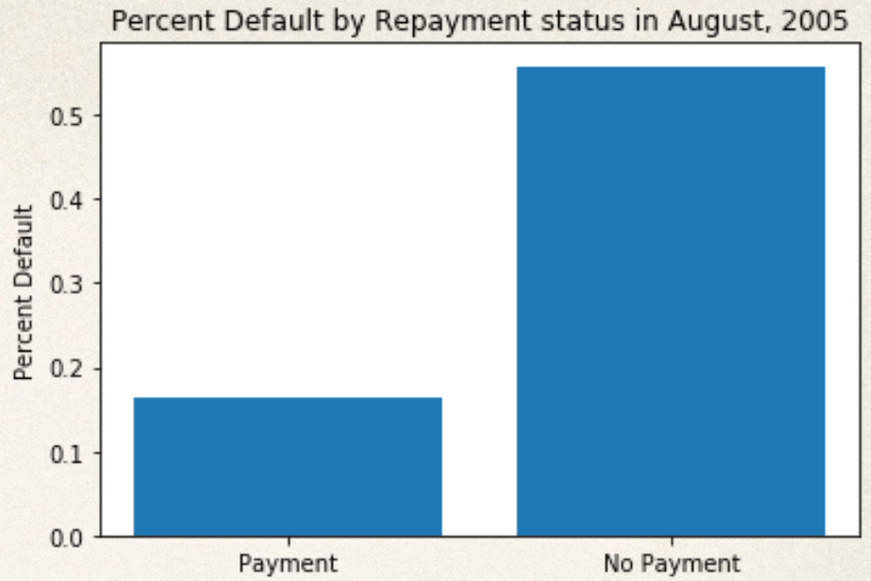


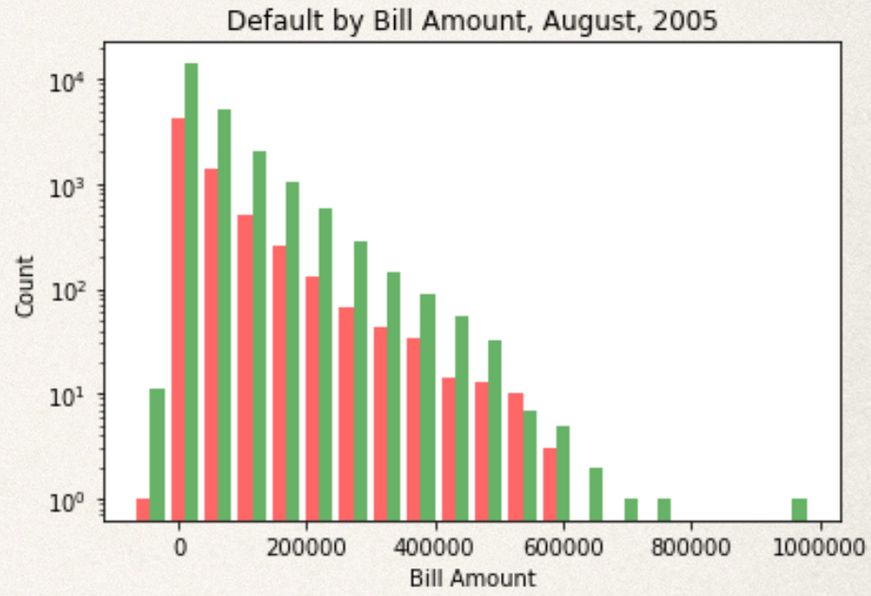


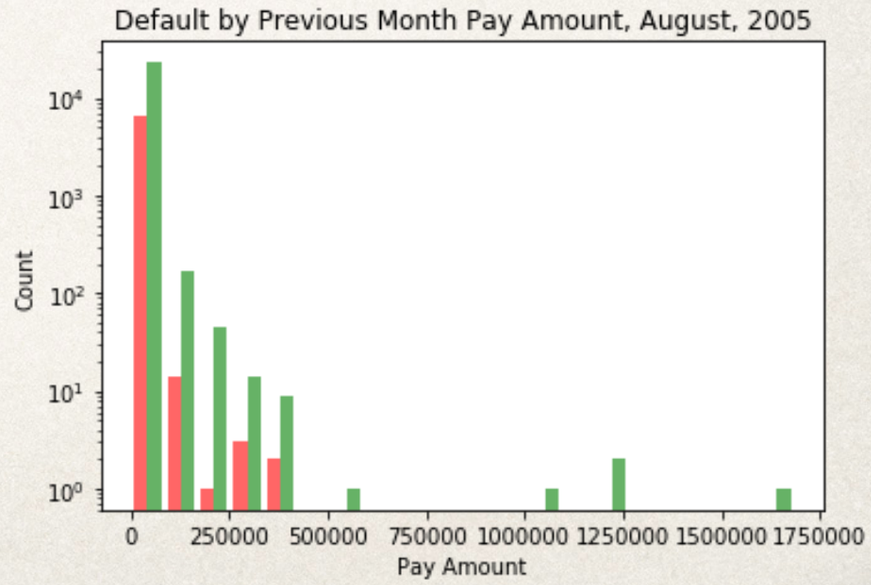


Pay Data August, 2005

- Clients who made payments still defaulted (?)
- Clients with bill amount 0 or less defaulted (?)
- Negative correlation between pay amount and default

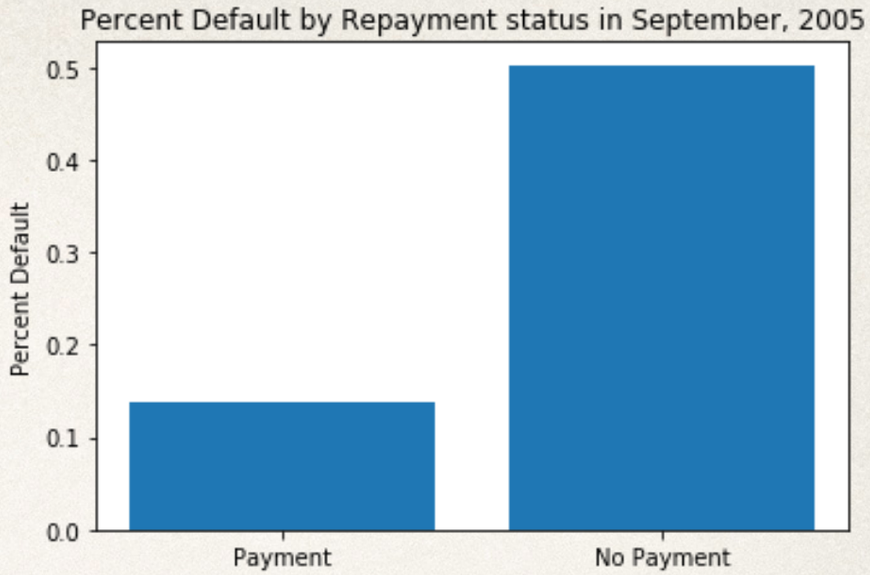


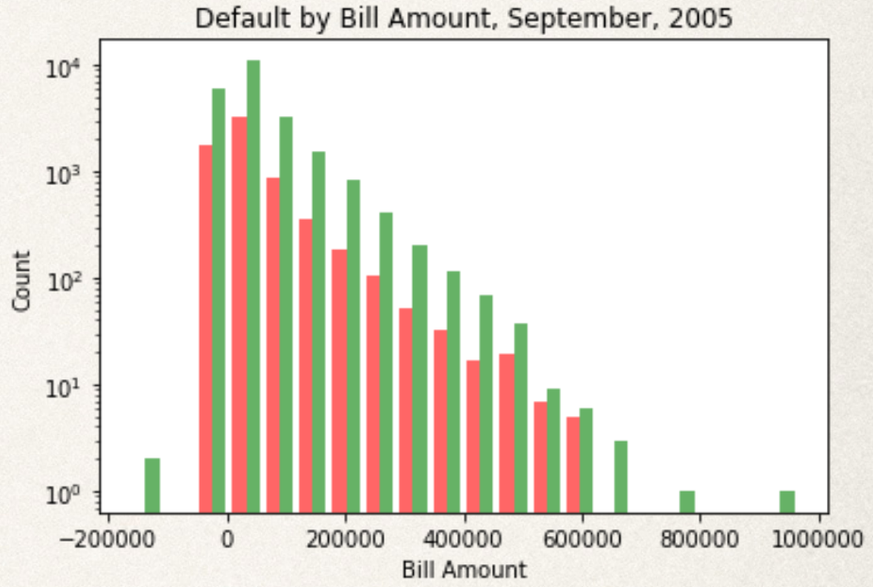


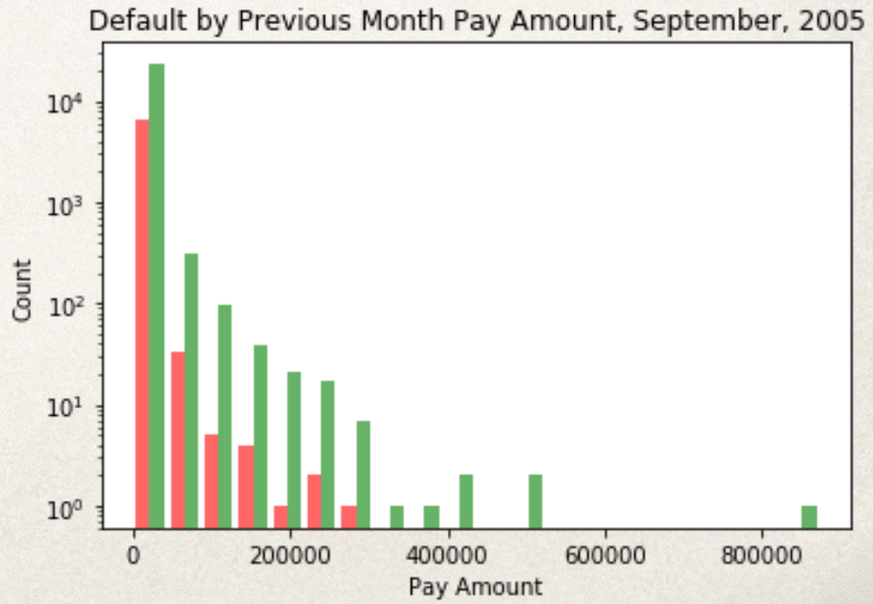


Pay Data September, 2005

- Clients who made payments still defaulted (?)
- Clients with bill amount 0 or less defaulted
- Negative correlation between pay amount and default







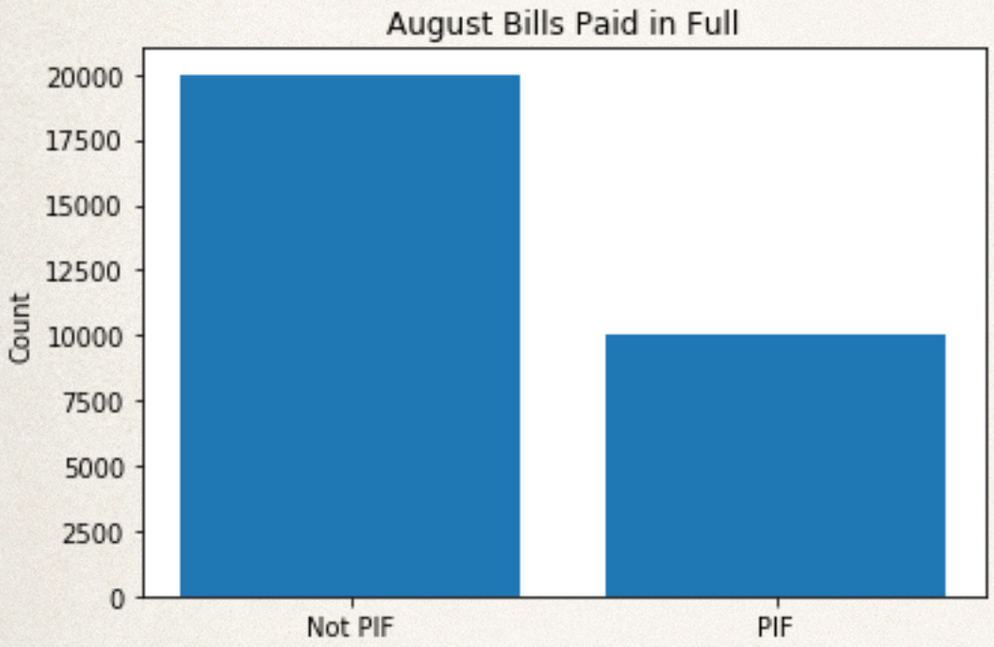
Data Modeling

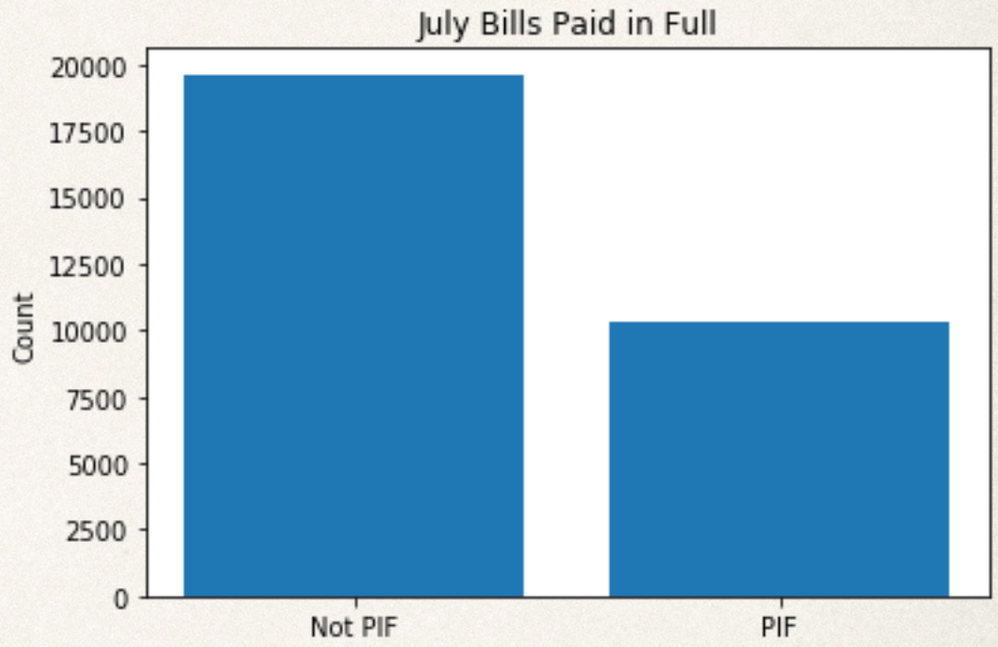
- Engineer ‘PIF’ feature
- Split categorical columns into dummy variables
- Feature selection
- Model selection
- Metric selection
- Model tuning

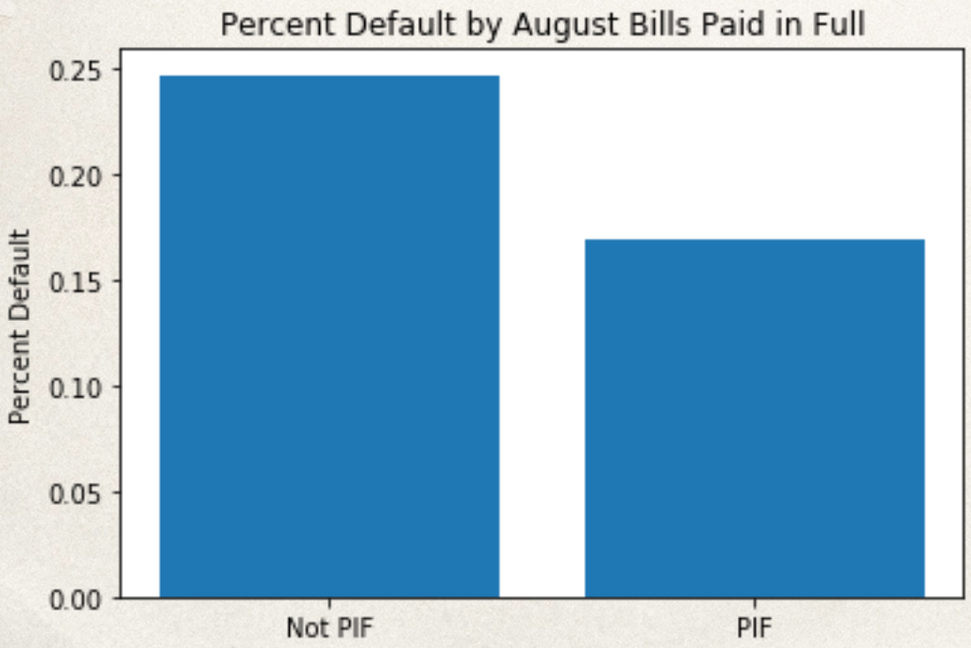
Engineering ‘PIF’

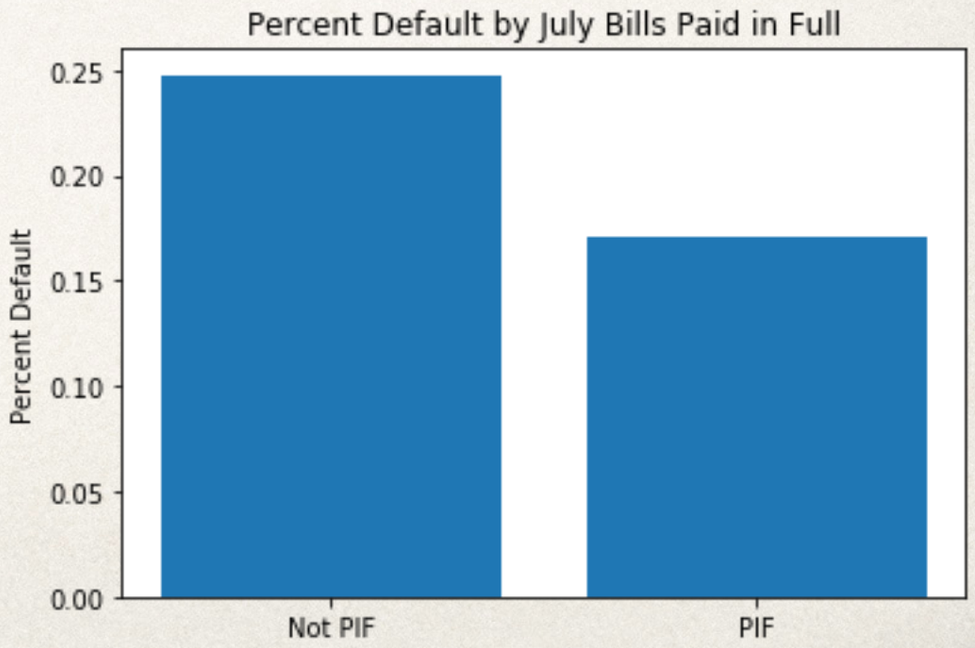
- PAY_AMT(N) >= BILL_AMT(N - 1)
  - August to April

EDA for PIF Columns

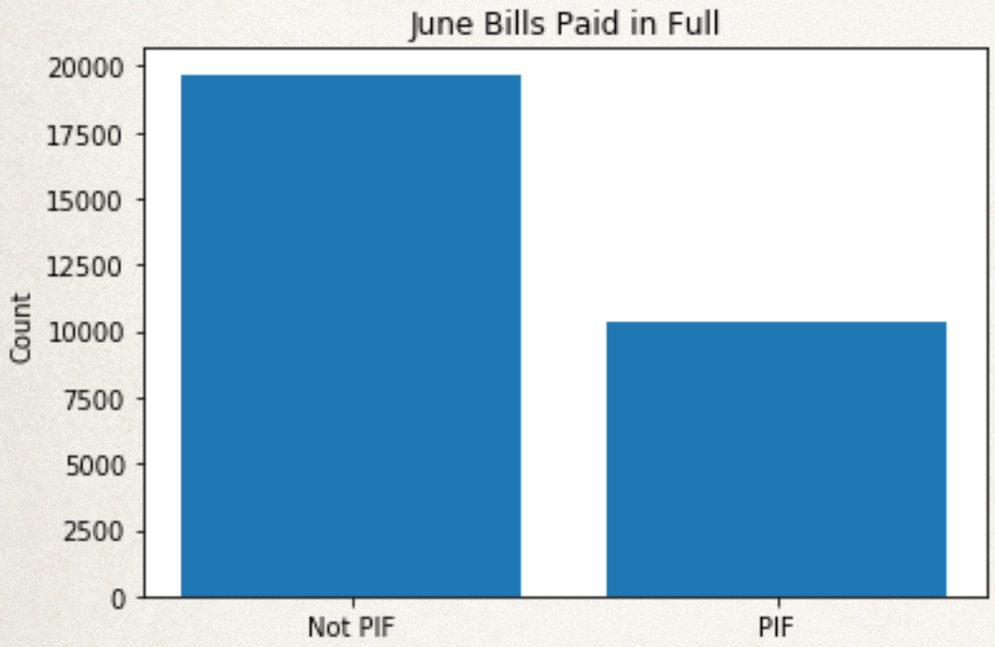


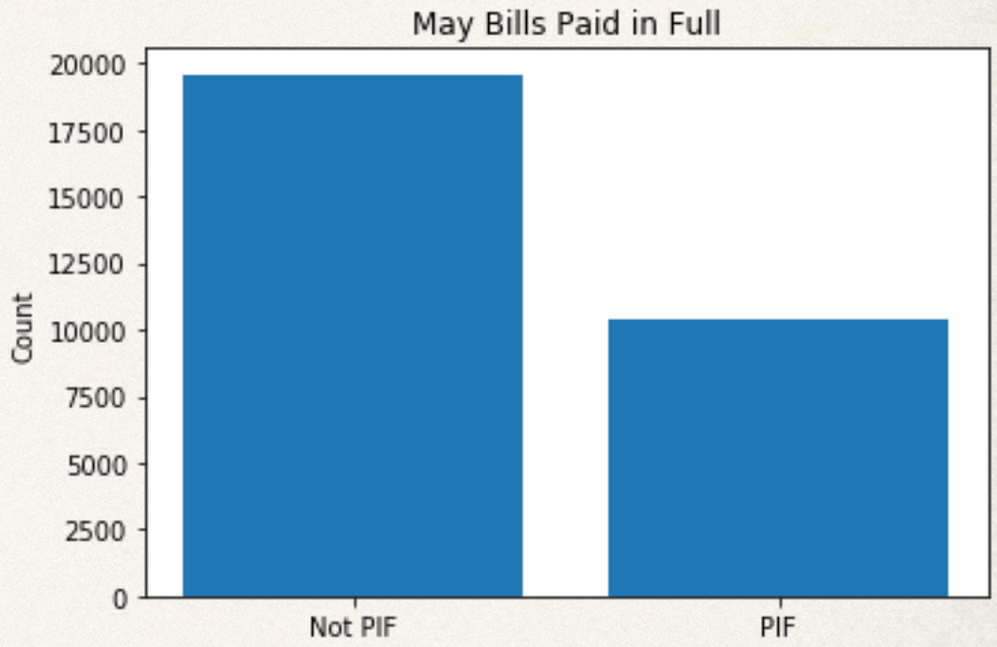


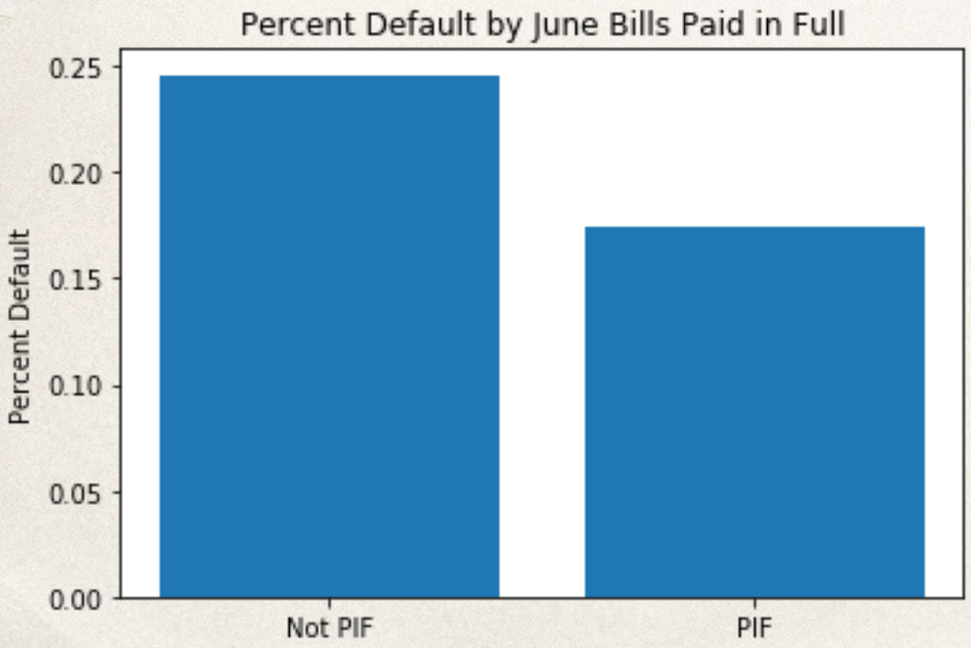


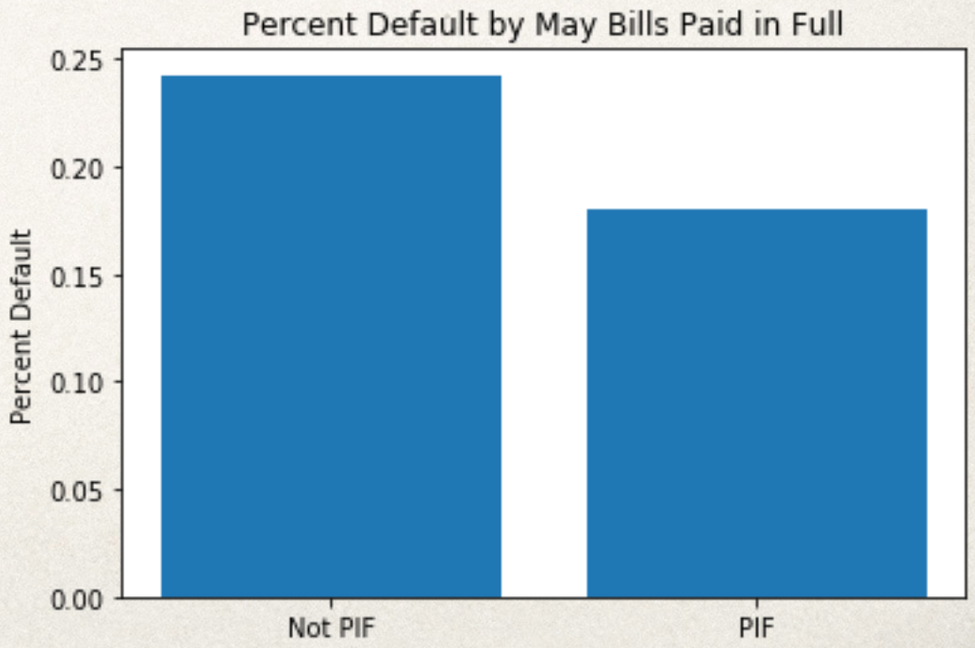


EDA for PIF Columns

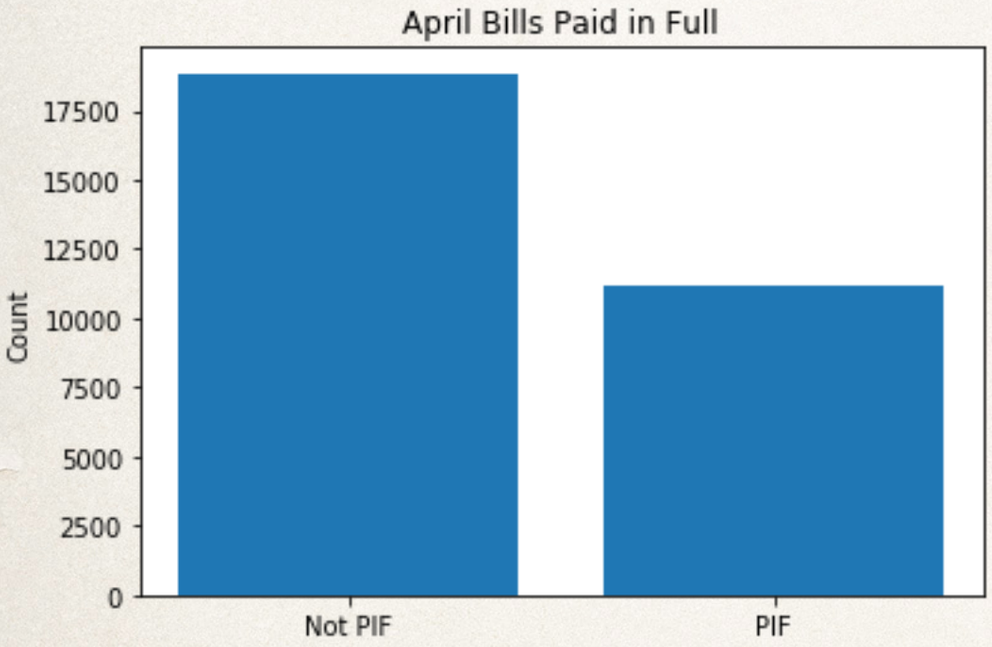


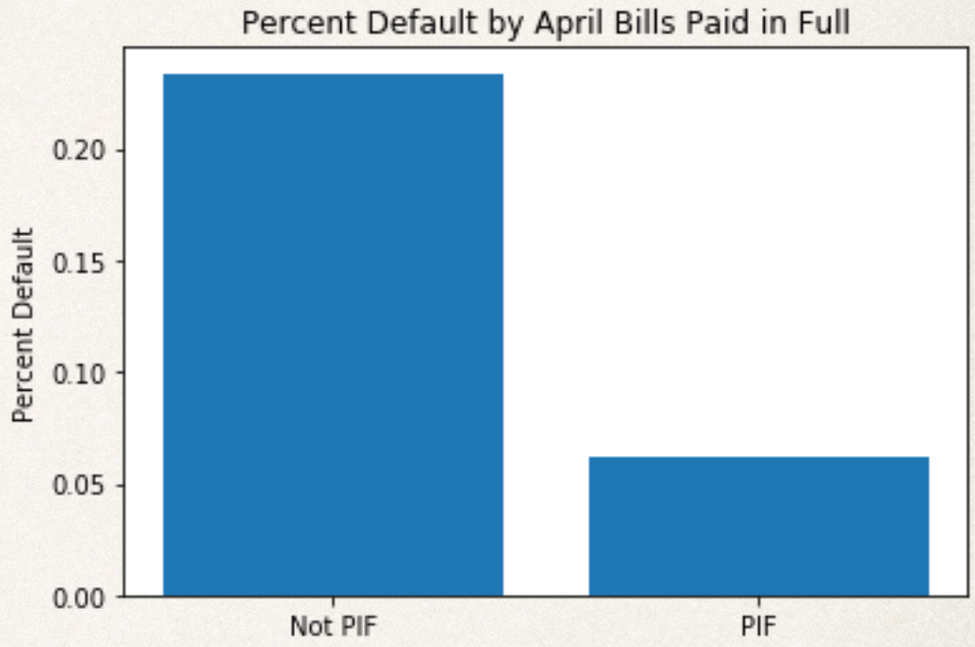






EDA for PIF Columns





Feature Selection

- Drop clients who closed their accounts
  - PIF2 = 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

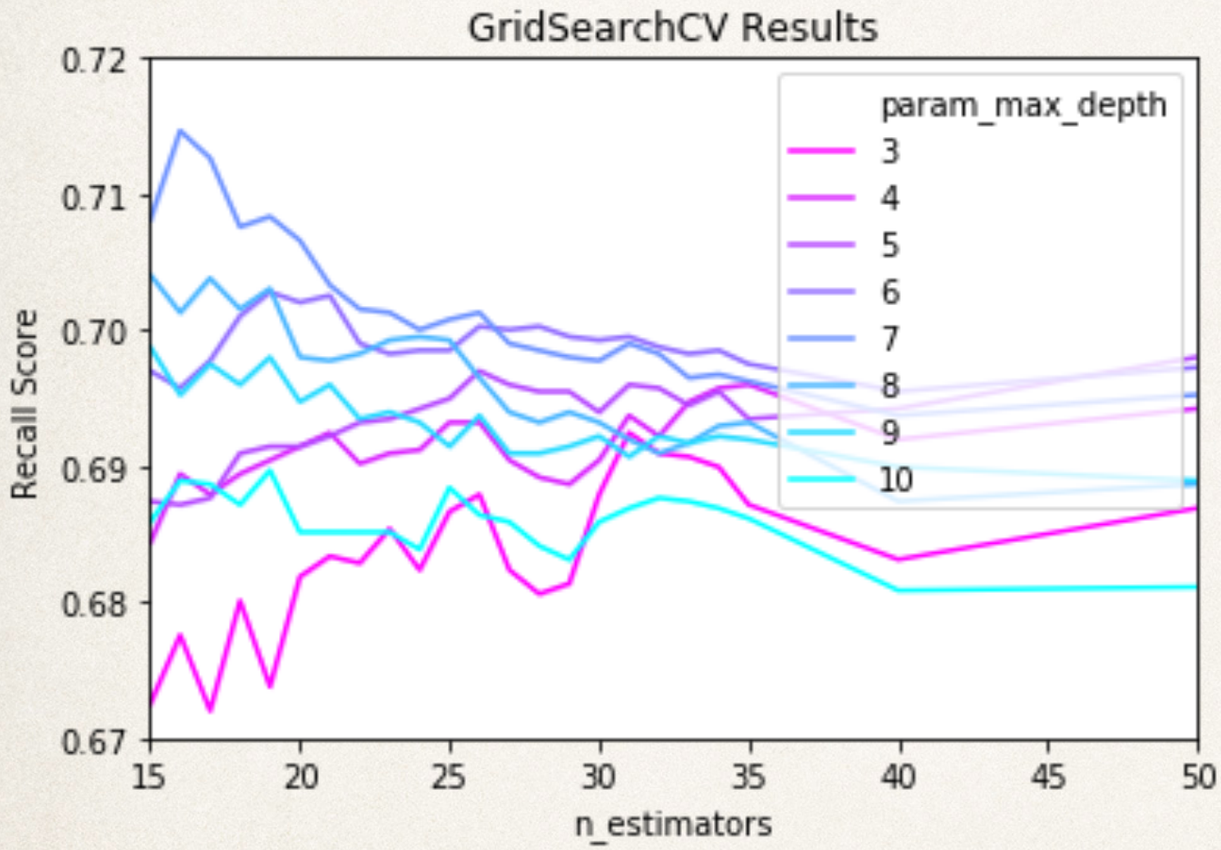
Goal: save the bank money

- 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
- Test different models