Estimating the Probability of Credit Card Default

with Gradient Boosting

March 23, 2018

Springboard

Contents

Springboard

The Business Problem

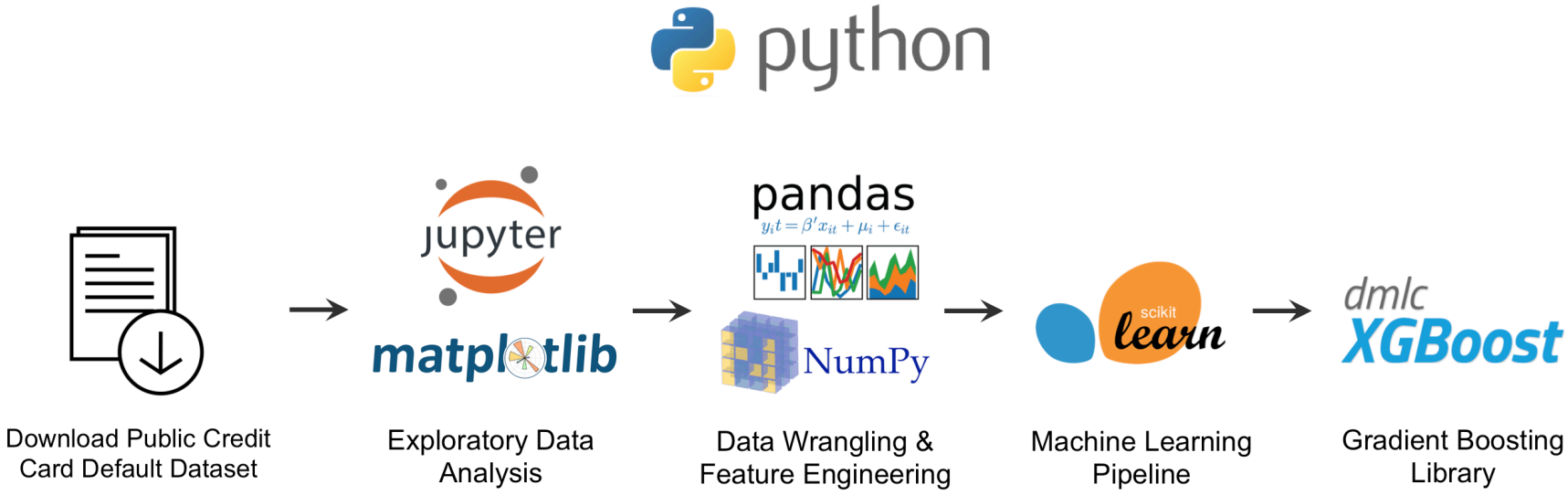
Springboard

- The Circumstances:
  - Assume that we are in the business of making credit card loans.
- The Problem
  - Estimate the probability that a given credit card client will default next month.
- Potential Business Impact
  - We could potentially use our estimates to:
    - 1) Identify low-risk clients for credit limit increases;
    - 2) Help create more reliable stochastic cash flows forecasts;
- Framing the Problem
  - We frame it as a binary classification problem.
  - Binary response variable for default next month.

The Tech Stack

Springboard

- The graphic below summarizes the main technologies and software libraries that we used at each step of the project.



The Dataset

General Description

- Source:
  - Data taken from a sample of Taiwanese credit card clients in 2005.
  - Originally used by Dr. I-Cheng Yeh in his 2009 paper:
    - "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients."

- Size:
  - Number of observations: 30,000
  - Number of explanatory features: 23

- Response variable **y**:
  - Binary response variable for whether or not a client defaulted in October, 2005.
    - **y = 1** if a client defaulted in October, 2005;
    - **y = 0** otherwise.
  - ~22 percent of the observations defaulted.

The Dataset

Explanatory Features

- 23 explanatory features:
  - 9 categorical variables;
  - 14 continuous variables.

- Categorical variables:
  - Level of education;
  - Marital status;
  - Gender;
  - Repayment status in 4/2005;
  - Repayment status in 5/2005;
  - …
  - Repayment status in 9/2005;

- Continuous variables:
  - Age;
  - Credit limit;
  - Amount billed in 4/2005;
  - Amount billed in 5/2005;
  - …
  - Amount billed in 9/2005;
  - Amount paid in 4/2005;
  - Amount paid in 5/2005;
  - …
  - Amount paid in 9/2005;

Feature Engineering

Two Ratios

- The dataset contains amounts billed and amounts paid for each month from April to September, 2005.
- Used these features and credit limit to calculate two sets of ratios:
  - 1) $\left(\frac{Amount\ Billed}{Credit\ Limit}\right)$
  - 2) $\left(\frac{(Amount\ Billed) - (Amount\ Paid)}{Credit\ Limit}\right)$
- Adjusts the amount billed (& amount paid) by the client's credit limit.
- Probability of default should increase as these ratios increase, *ceteris paribus*.

Feature Engineering

One-Hot Encoding

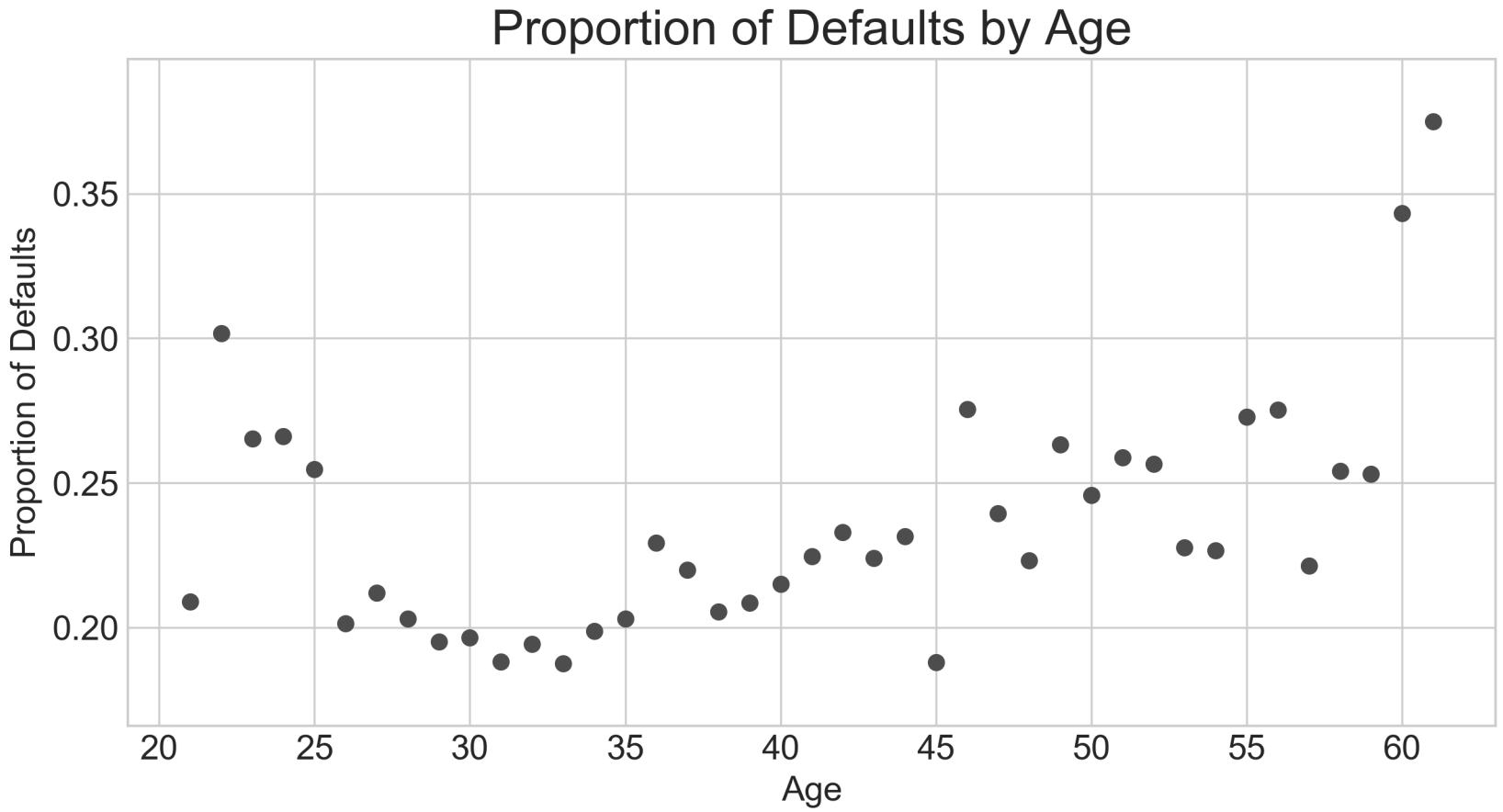
Springboard

- The nine categorical features in the original dataset had been encoded with integer values.

- We transformed these nine categorical features into 65 one-hot encoded vectors.

Exploratory Data Analysis

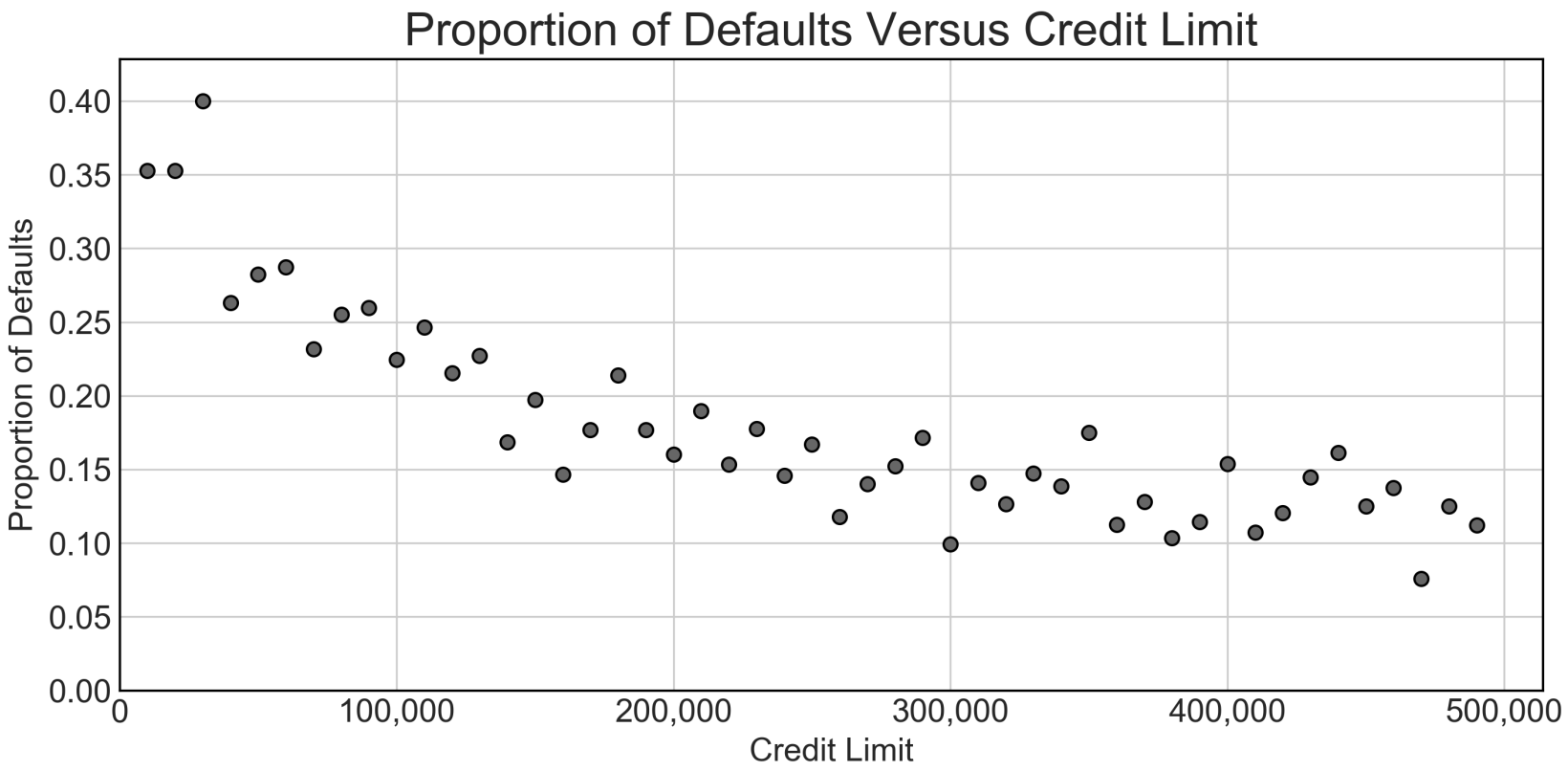
Proportion of Defaults by Age



- Note: Only plotted bins with at least 50 total observations.

Exploratory Data Analysis

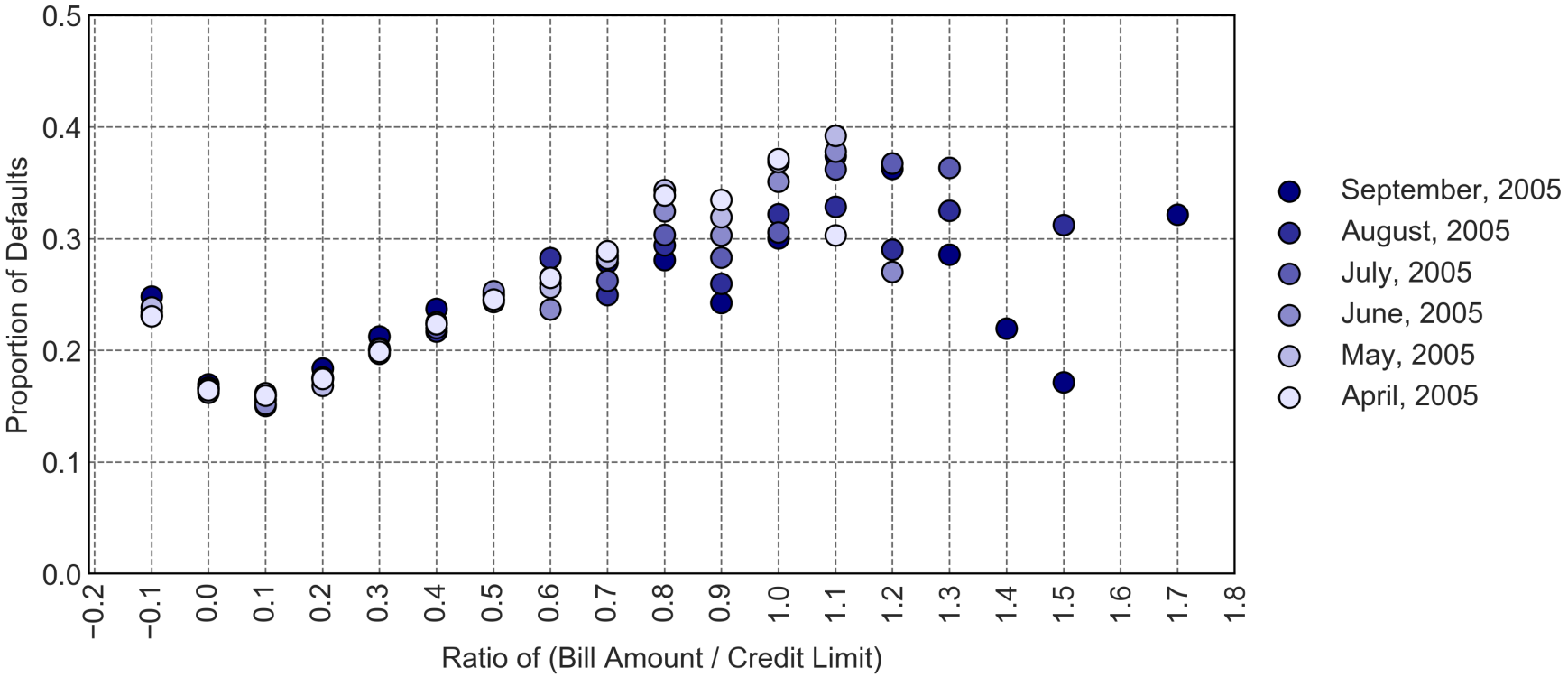
Proportion of Defaults Versus Credit Limit



- Notes: - Credit limit observations were binned using a bin size of 10,000 Taiwan New Dollars.
  - Only plotted bins with at least 50 total observations.

Exploratory Data Analysis

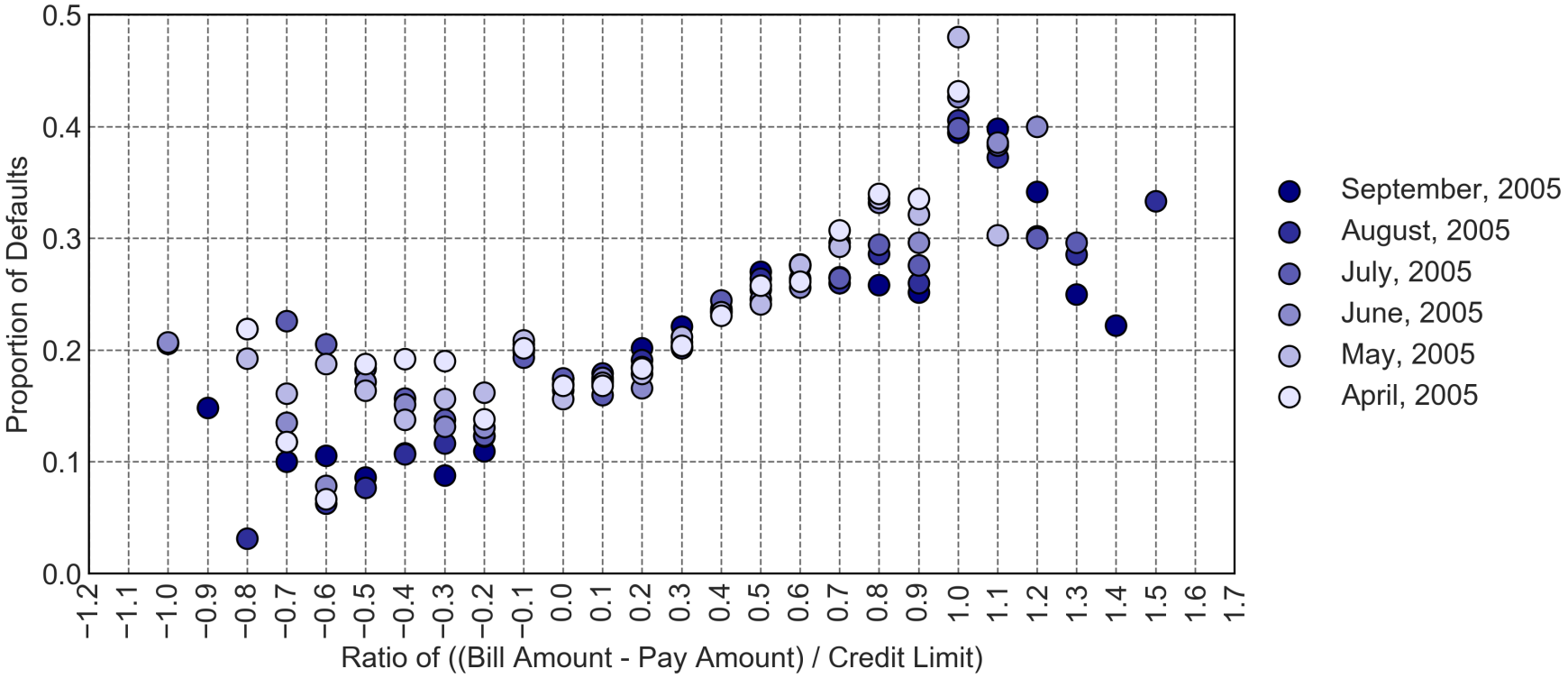
Proportion of Defaults Versus the Ratio of (Bill Amount / Credit Limit)



- Notes: - Ratio observations were binned using a bin size of 0.1.
  - Only plotted bins with at least 50 total observations.

Exploratory Data Analysis

Proportion of Defaults Versus the Ratio of ((Bill Amount – Pay Amount) / Credit Limit)



- Notes: - Ratio observations were binned using a bin size of 0.1.
  - Only plotted bins with at least 50 total observations.

Gradient Boosting

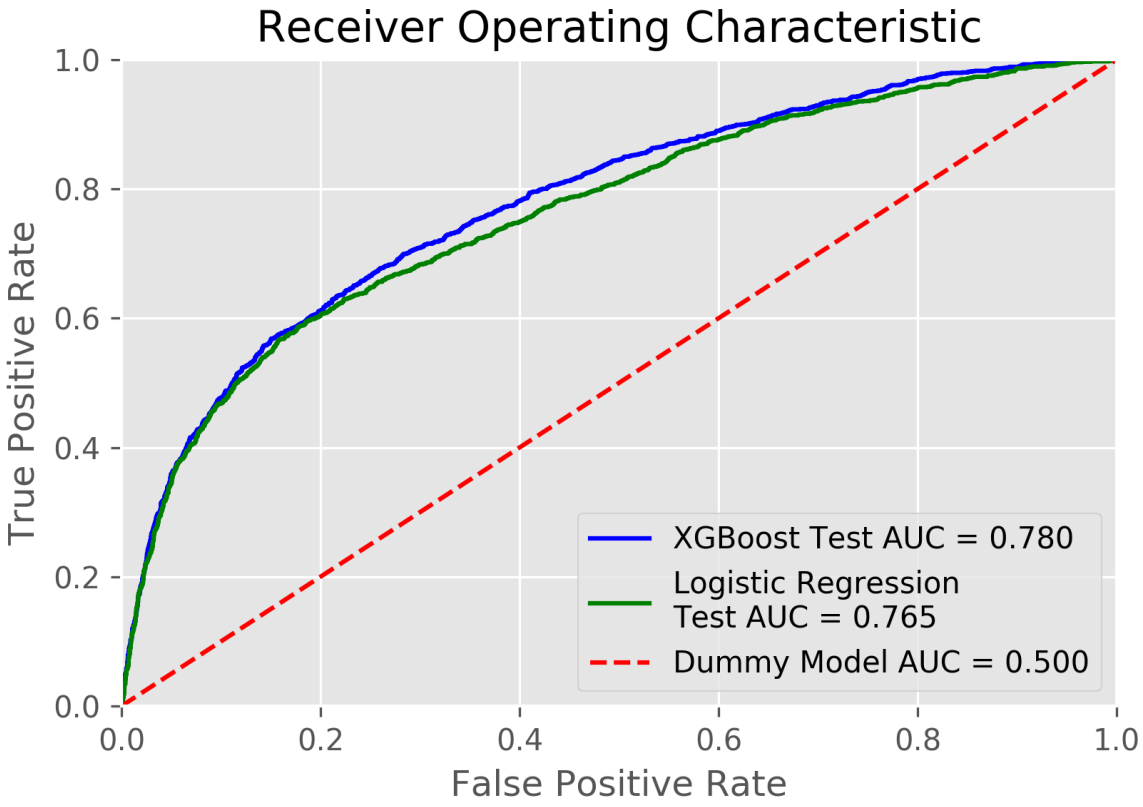
What is it?

- What is Gradient Boosting?
  - An ensemble of decision trees (weak learners)
  - Fit in a stage-wise manner:
    - In each consecutive stage, a new decision tree is added to compensate the shortcomings of existing decision trees

- Gradient Boosting Pipeline
  - Preprocessing steps:
    - 1) Variance Threshold
  - Perform cross-validated grid search to identify optimal hyperparameters, e.g.:
    - Learning rate
    - Number of estimators
    - Maximum depth
    - …

Model Performance

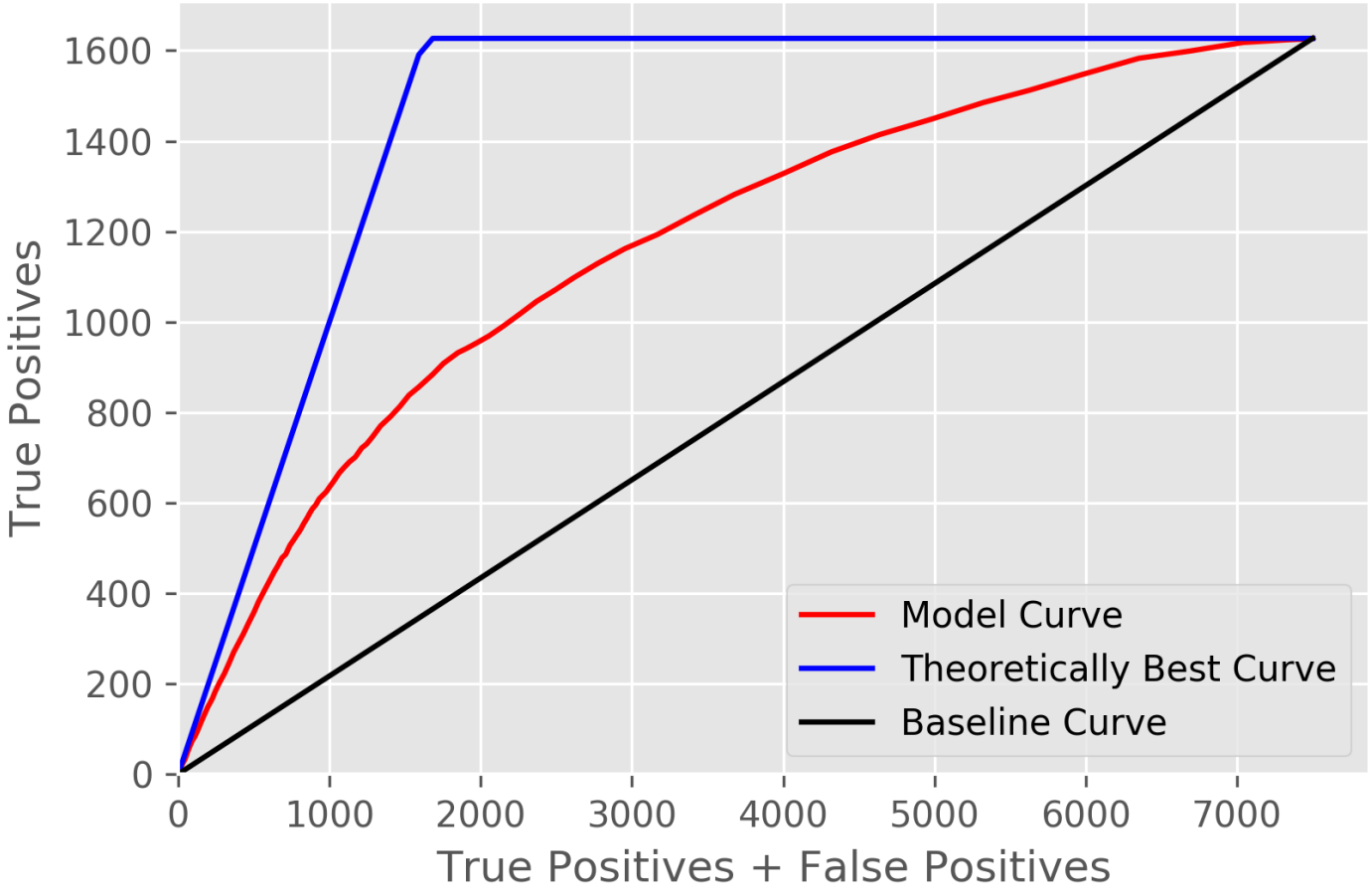
ROC AUC Score



- Our XGBoost model achieved a ROC AUC score of 0.78 on the test set and a cross-validated ROC AUC score of 0.78 on the train set.

Model Performance

Lift Chart – Test Set



Model Performance

Lift Curve Area Ratio

Method	Error Rate		Lift Curve Area Ratio	
	Training	Validation	Training	Validation
Results of Yeh & Lien (2009)				
K-nearest neighbor	0.18	0.16	0.68	0.45
Logistic regression	0.20	0.18	0.41	0.44
Discriminant analysis	0.29	0.26	0.40	0.43
Naive Bayesian	0.21	0.21	0.47	0.53
Neural networks	0.19	0.17	0.55	0.54
Classification trees	0.18	0.17	0.48	0.536
My Results				
XGBoost	0.17	0.18	0.659	0.560

- Our XGBoost model achieved a Lift Curve Area Ratio of 0.56
- This outperformed Yeh & Lien's scores on the validation set.

Thank You

Springboard