

# Capital One Data Science Internship

McLean, VA - Summer '16

Ryan Shiroma

## Presentation Outline

1. Getting the Internship
2. Company Background
3. The Work
4. Important things to know
5. What I Learned

# Getting the Internship

## Search Process

- Job board sites
- LinkedIn
- Company websites
- **Piazza**

The screenshot displays the Piazza Careers interface. On the left, a user profile for Ryan Shiroma is shown, including a photo, name, education (San Jose State University, Masters 2017, Major: Statistics), and social media icons. The main area shows an email inbox with a message from Capital One dated 2016/03/04. The message content is as follows:

Capital One  
Great! I'll apply tonight and get th...  
Learn about Piazza Careers

Capital One  
Hi Ryan,  
My name is [redacted] and I am in Campus Recruiting at Capital One. Currently, we are looking for really talented and qualified Data Scientists and after coming across your profile, I am very interested in chatting!  
I see that you are a Computer Science enthusiast with a background in engineering and coursework in statistics. That combination of coding and computer skills with the ability to model data is really important for this Data Science role!  
The Data Science Internship is an amazing opportunity that is a hybrid of skillsets from hard skills like statistics and computer science to strategic skillsets like individuality and creativity. Capital One is trying to harness the power of data to drive our mission to help population just below prime move to upmarket and change banking for good! On

# Getting the Internship

## Interview Process

1. Take-home dataset project
2. Three onsite one-hour interviews
  - a. Behavioral Interview
  - b. Role Playing Interview
  - c. Technical Case Interview



## Company Background:



- Top 10 bank
- Started as a credit card service in 1988
- Primary markets on East Coast and the South
- Heavily focused on technology and data science
- Large intern program



# The Work

## **Background:**

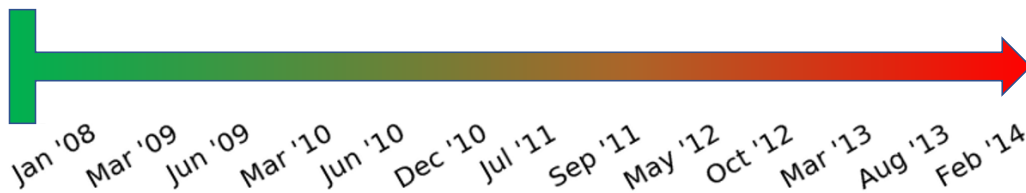
*Credit Risk Modeling:* Banks frequently model the likelihood that a given credit card holder is to default on their payments sometime in the future.

*Data:* transactions, payment history, external data...

*Model Types:* Logistic Regression, Tree based(Random Forest, GBM, etc)...

# The Work

**Primary Question: What are some important factors to consider in order to minimize overfitting on GBM models?**



Hyper Parameter	Settings
1. Depth	2, 4, 6
2. Trees	150, 300, 600, 900, 1200, 1600
3. Learning Rate	0.005, 0.01, 0.02, 0.03, 0.05
4. Minimum samples in a leaf	1 sample, 0.1%, 1%, 2%,3%, 5%
5. Maximum features considered for a split	Square Root, All

# The Work

## **My Analysis:**

- 1) fractional factorial design using models as “experiments”
- 2) Analyze the degradation effects of GBM hyperparameters over time

## **Results:**

Model degradation was sensitive to the values of two GBM model parameters and should be chosen carefully for future modeling.

*“Tree Depth” > 2 and “max-features” >  $\sqrt{N}$*



# Important things to know

- Relevant Statistics Courses
  - Regression(261A), Design of Experiments(261B), Classification(285)
- Programming, Programming, Programming...
  - R/Python/SAS
  - Unix terminal environments
  - Clustered computers

# What I learned

- **Technical Skills**

- Parallel Computing
- Python

- **Communication Skills**

- Presentation skills
- Non-technical communication of technical problems

## Main Take-Aways

- Apply to jobs through as many channels as possible!
- Don't neglect communication skills!
- Learn LOTS of programming!

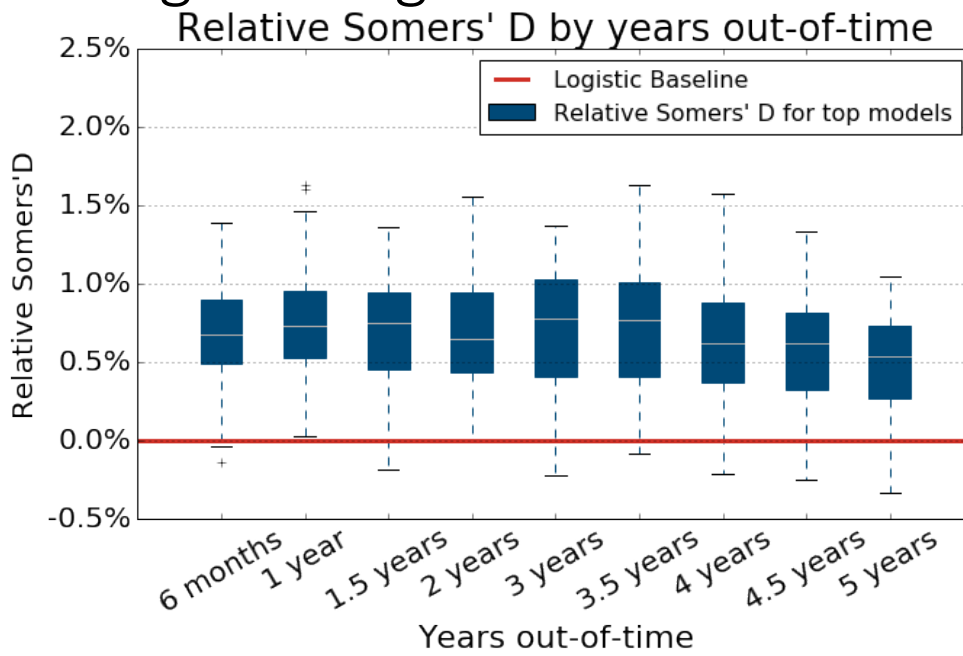
# Work Environment



Questions?

# Appendix

# GBM vs Logistic Regression over time



# Model Degradation over time

