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
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?

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

Relative Somers’ D by years out-of-time

- 6 months
- 1 year
- 1.5 years
- 2 years
- 3 years
- 3.5 years
- 4 years
- 4.5 years
- 5 years

Years out-of-time

Relative Somers’ D

- Logistic Baseline
- Relative Somers’ D for top models
Model Degradation over time