

earnest

A data science internship

August 14, 2014

(censored for public view)

Predictive Algorithms

Build our models to predict applicant quality

- ▶ Primary motivations: 1) scalability, 2) improved decisions over hundreds of variables, 3) prioritizing applicants

Data Integration and Preprocessing

- ▶ Gather and merge important features from both Mongo collections and Postgres tables
- ▶ Make note of missing data, and how we should account for this while making predictions

Predictive Algorithms

Modeling

- ▶ Apply random forest on data, and regularize parameters
- ▶ Take advantage of random forest's in/out-of-bag

Inference and Priority Roadmap

- ▶ Analyze our results, and decide on what should be improved next before more optimization

Data Science Website



Infrastructure

- ▶ Build the framework from scratch
- ▶ Build webpages for clients, underwriters, queues, models, loans, and—soon to be deployed—weekly reports
- ▶ Automate the website updates with cron

Visuals

- ▶ Deliberate on useful, easily accessible visualizations
- ▶ (More suggestions on metrics are welcome!)

Mining Financial Transactions

Partial Dependence on

probability
Confidential!

- ▶ Find trends among our clients' spending habits
- ▶ Find most important keywords in distinguishing applicant quality
- ▶ Selectively refine the list of keywords and their underlying relationships

Collaborations

Marketing

- ▶ Data munging
- ▶ Reporting
- ▶ Consulting
- ▶ Potential blog post on all the intern's experiences/value?

Underwriting and Risk

- ▶ Important considerations for future modelling purposes
- ▶ Priority order for underwriting queue (pending implementation?)
- ▶ Automated decision-making
- ▶ Weekly reporting

Understanding our Competitors

Confidential!

- ▶ Web Scraping & Parsing
- ▶ Analyze common trends among their clients, and their volume of clients over time
- ▶ Examine their pricing model for different interest rates

For Future Data Scientists

- ▶ Coding Conventions
- ▶ Documentation
- ▶ File Management
- ▶ Finalize use cases for data science tools

There's more work to be done!

- ▶ More variables
- ▶ Data science website
- ▶ Predict decline reasons
- ▶ Predict variance in underwriter scores
- ▶ Fraud detection
- ▶ Unit testing (docker :)]

est