Mobile App Rank Prediction

Data and prizes courtesy of

Algorithmic Venture Capital

- Rocketship VC makes Venture Capital investments based on data and algorithms
  - Team from Stanford, Amazon, NASA

App Rank Prediction Problem

- Consider a cohort of 10K mobile apps, each with less than 1M downloads today
  - “New kids on the block”
- Some of these apps will soar to over 100M downloads by the end of the year. Most will languish.
  - The next Snapchat!
- Problem: Observe the cohort for 8 weeks, then predict Top 100 at the end of 1 year
  - Top 100 = Top 1%

Data

- ~40K iOS apps
  - Basic data including app name, category, company, etc
- Daily app downloads for 8 weeks, and downloads at end of 1 year
- App Ratings (count of 1-5 star ratings)
- App Reviews
  - Review count, review text
- Usage
  - Open rate, active users, average session time, avg sessions/user,…

Evaluation and baseline

- We have a baseline model that uses only the time series of downloads
  - Ignores everything else (ratings, reviews, usage)
  - We will share the baseline predictions

- Evaluation
  - You will be asked to predict the top 100 for a “withheld cohort” of 10K apps
  - Your score = Overlap between predicted top 100 and true top 100

Prizes!!

- Each team that beats the baseline by at least 10% will get a prize

- Grand prize for best predictions:

  Oculus Rift!

- Be a VC: if you help us identify an interesting app, you can be part of the evaluation process

For more info on the dataset

Email me!

Anand Rajaraman
anand@cs.stanford.edu

Real Estate Listing Prediction

Data courtesy of Smartzip

Problem Statement

- 95 million homes in the US

- What is the probability that a home will be “on the market” within the next 12 months?

Data

- 95 million homes
  - Geo (census tract, county, ..)
  - Estimated price
  - Property attributes (bedrooms, building area, lot size, year built, property type, etc)
  - Historical data (last sold price and date, turnovers in last 10 years, defaults, listings)
  - Outstanding loan, LTV/equity
  - Neighborhood demographic data
  - Data about owners (employment, children, etc)

Evaluation

- For each property, output a probability of listing
- Sort by probability and keep the Top 20%
- Lift = (Top 20% Listing Rate)/(Overall Listing Rate)
- Optimize for Lift at various levels (nation, state, county)
- Company has a predictor you will try to beat

For more info on the dataset

Email me!

Anand Rajaraman
anand@cs.stanford.edu