H2O World

Lead scoring for real time bidding

Introduction

Rushcard : Prepaid debit card
Activehours : On-demand payroll

- Problems in Marketing, Digital Acquisitions, Lead scoring, Lifetime value predictions, Operations, Customer Service, Risk, Retention

Scoring leads

- Traditionally done after someone has signed up. Use post signup information, append additional data-points (3rd party, demographic, etc.)
- Use the assigned score to send promotions, incentives, targeting offers etc.
- Useful for increasing usage/value of a cohort of acquisitions
- Practical considerations - insufficient data during signup, no viable use case other than display banners or offers, cost of scoring real time.

Scoring leads - Now

- Programmatic buying, real-time bidding, customized ad serve.
- Optimized for conversion, mostly topline.
- Can we optimize for conversions based on projected lifetime value?

What is available at signup?

- Web browsing data: Pages visited, time spent on site or app, source of the lead, any additional touchpoints, time of day, day of week

- IP address/location matched with zipcode statistics/ census data: median income, demographic breakup - race, gender etc., time zone, unemployment rates, growth in income, population etc.

More .. in this example

- Survey questionnaire answers
- Age, gender
- Type of product selected
- Price range browsed
- Id verification results
- Additional scoring information from data vendors based on email address

Sample features

- 70 actual features - individual categories broken out to create about 219 features
- Predict likelihood of profitability. We'll use a simple categorical variable for that today
  - Try comparing with a cost function based variable

Pre-processing

- All the data comes from a data-warehouse.
- Web tracking data has a lot of missing values.
  - Categorize all missing values into new group?
- Categorize continuous variables like Age?
  - Increasingly, not a good idea anymore
- Dependent variable - Activated, Direct Deposit or a combination of the two. Can you also include a cost function here?
  - For e.g. Activated = $25, DD = $100, Non- activated = $5
- Trust the tool more as compared to intensive cleaning processed earlier. Trial and error is easier now. Just run.

Models

- Random Forests
- GLM (binomial) - try playing around with the cutoff rate, can lead to interesting scenarios
- GBM
- Naive Bayesian?