Project 3 (McNulty) Proposal

Jason Salazer-Adams

Overview

A general e-commerce site would like to answer a fundamental question, "Can they predict a customer's implicit filters, e.g. product type, price range, etc.?" I am planning to attempt to address this question by utilziing a Kaggle data set provided by RetailRocket. The data set was collected from a real e-commerce website and represents visitors buying behaviors in the form of view/addtocart/transaction events. Event definitions are,

- view - Visitor clicked on an item's page.
- addtocart - Visitor added the item to their shopping cart.
- transaction - Visitor bought the item.

There are three main data sets,

1. **Category tree** - hierarchical data set relating the lowest level category to the top level parent
2. **Events** - The visitor's events on the e-commerce site
3. **Item properties** - Weekly recording of properties for an item, e.g. category, price, availability, etc.

The data has been anonymized and all item properties have been hashed, except for category and available. The value of the properties are also hashed, except for numerical data. Based on a response from RetailRocket on Kaggle, it seems price and discount could be reveresed engineer. However, all other categories are not interpretable, which is a concern.

The purpose of the model is to be able to predict a property of the product which will be added to the visitor's cart. There will be bias in the model as I will only be looking at visitor's who eventually added an item to their cart, and there are visitors who never add an item to their cart.

MVP

An MVP would be predicting the top-level parent category (1 of 25) of the item added to the cart using the view event previous to the addtocart event.

Data Characteristics

I performed basic EDA of the data sets as I was concerned about the "raw" internet traffic data. Here are some basic characteristics of the data.

- The event split of view/addtocart/transaction is 2,664,312 / 69,332 / 22,457 observations.
- There are 1,407,580 unique visitors.
- There are 25 top-level item categories ranging from 1-290 sub-categories within each of thest top-level categories.
- There are 235,061 unique items.
- The events were recorded from 5/2/2015 to 9/17/2015
- There is a visitor which recorded 7756 events, which is 3400 more than the second highest. I will most likely need to exculde these heavy users or develop a sampling strategy to prevent bias.

Features

Here is a list of features I am currently considering.

- Category of the items viewed previous n steps from the addtocart event.
- Availabiity of the item viewed previous n steps from the addtocart event.
- Length of session defined as start being the first view, and end as the current view.
- Hour of day
- Day of week

Something else I am considering is a estimating a simple Markov chain with states being view/addtocart and transition being the item on the page. Unsure whether this will add predictive value to the model or not.

Challenges

Here are the challenges I foresee.

- Data cleaning tasks could be substantial and have an effect on the quality of the model.
- Size of data may require an AWS instance to run the model instead of my laptop.
- Item categories, properties, and values are anonymized which will make it hard, if not impossible, to interpert the model.