Vince Laird

CAPSTONE 2 PROJECT MILESTONE REPORT
DATA SCIENCE - CAREER TRACK

E-commerce: Predicting User Actions

The second capstone project will be based on the RecSys Challenge 2015. The website and details about this challenge are here: http://2015.recsyschallenge.com/

This is a problem of coding a recommender system. The data used provides a sequence of click events / click sessions. For some sessions, there are also buying events. The goal is to predict whether the user is going to buy something or not, and if the user is buying, what would be the items the user would most likely buy.

This information is valuable to e-commerce businesses as it can indicate what items to suggest and how to encourage a user to buy (such as via promotions, discounts, etc.)

The data represents six months of activities of an e-commerce website selling random items such as garden tools, toys, clothing, electronics and more.

Data set

The data set provided on the RecSys website contains click information, buy information and test data. The click information and buy information will be utilized; since the challenge is over, determining a final score using the test data will not be possible, as the correct “answers” for the test data are not available.

Exploratory data analysis was performed on the data; the click data set contains 33M rows of information, and the buy data set contains 1.15M rows. Further information regarding the data set can be found in the notebook here:

https://github.com/vincelaird/springboard/blob/master/capstone2/Capstone%202%20milestone%20notebook%20EDA.ipynb

Initial findings

Given the characteristics of the data, we will use tree based learning algorithms such as decision trees, random forest and gradient boosting. We will be able to extract features from the timestamp data provided, categorical information related to the item and the item's price.