Predicting Netflix Use

jlevin1, dmutako, szitter

Goal

Everyone uses Netflix. The question here is can we predict Netflix use over the next few days/weeks. The use case for this project is that Netflix might need to predict demand in order to scale infrastructure. Thus, we are mostly interested in the following task setting: given historical usage numbers and data about new content, predict the total amount of content that will be streamed in the upcoming week (7 day period).

Data

Netflix was kind enough to sponsor our project since it's worth millions to them. They provided 2 years (2014-2015) of historical use per-day: number of servers, peak concurrent users, total unique users, and total sum of content streamed (in hours). We augmented the data by scraping the historical release dates of Netflix TV-shows and movies, including the movie's production budget. We joined the datasets on name of new movies added to Netflix. We threw away data on 1,000 movies (1%) for which names were not unique and could not be joined. Our final data table as rows with the following fields of interest: date, zip code, total usage, total new movies, total budget of new movies, mean budget of new movies, weather (1 if raining else 0)

Model+Evaluation Setup

We care about forecasting demand in the coming week based on the current week's usage data and the movies that will be added in the coming week. We do not expect to generalize to unseen locations (e.g. zip codes) but we do what to generalize to realistic future settings in which the movies being added are unseen. We therefore train on the data from 2014-July 2015, and test on data from August-December 2015. This ensures none of the movies added in our test data were seen in training. We have a gazillion training examples and 0.25 gazillion testing examples, which is large, so we don't need to use cross validation. We train a ridge regression and report accuracy using MSE.

Results and Analysis

Claim #1: The classifier trained using all features outperforms baseline models by a significant margin.

Support for Claim #1: Table 1 shows the accuracy of our full model compared to a baseline model that always predicts no change in usage (e.g. if the usage is x this week, the baseline will assume it is x next week).

Model	Test MSE
Baseline	1056
Our Full Model	54

Claim #2: Most of the predictive power comes from features related to previous week's usage; data about new movie content leads to only a small increase in performance.

Support for Claim #2: We train several versions of our model, using different subsets of features. We see most of the performance is coming from the "previous week's total usage" feature, and only a small gain comes from the movie content features. Specifically, movie features lead to a reduction of 3 points in MSE. We cannot say whether this change is significant.

Features	Test MSE
Just Previous Week Usage	58
Previous Week Usage+Total New Movie Budget+Mean New Movie Budget+Total New Movies	55
Full Model	54

Claim #3: The model is performing especially badly in the Northeast.
Support for Claim #3: Figure 1 shows MSE by region. We see that error is substantially higher in the northeast than elsewhere (MSE=68 in the northeast vs. 50 elsewhere). Investigating further, we see that, in our test data, the Northeast has significantly higher usage than in the training data (mean usage in NE in training data was 600+/-23 and in test data was 1000+/-52). We are not sure why this is. One theory is that Netflix expanded coverage in summer 2015. In order to investigate this, we would obtain a dataset of Netflix subscribers by zip code, and see whether subscriptions jumped in the NE more than in other regions during that time period.

[Insert fake chart here]