MIT Big Data Challenge: Transportation in the City of Boston

Model of Prediction Challenge

Melissa Gumrek

We found that the taxi pickup and dropoff data was by far the most informative dataset. SImple features such as the time of day and day of week captured most of the variation in the data. We also included features for the hour of day and the number of pickups and dropoffs nearby each test location. These features were included both for the time period of interest and for periods before and after the prediction interval. When available, the number of MBTA T rides at nearby locations improved performance. We experimented with adding weather and twitter data, but found these datasets did not improve our accuracy. We ultimately pooled all of our features into a random forest regression model, used a variant of greedy hill climbing to select the most informative features and built a separate model for each test location.

For the training set, most of our error, was driven by spikes in taxi demand at convention centers (e.g. Hynes and the Boston Convention Center). In general, we found that our model was able to predict the time of a spike but not necessarily its magnitude. This issue was likely pervasive in our prediction for the test dataset as well.

While many of our time-shifted features would require knowledge of the future, as mentioned above the most informative features were hour and day of the week. These features are likely to be highly informative in planning for taxi demand at specific locations. Robust data about major events would also be helpful to predict demand fluctuations. Unfortunately, we found the provided events data contained too little temporal information to be useful in predicting demand at high enough resolution.