When to Book: Predicting Flight Pricing

Qiqi Ren – qiqiren@stanford.edu | CS229

Predicting

As someone who purchases flights frequently, I would like to be able to predict when the best time is to buy in order to get the best deal. I chose to look at flights in a short-term range, which I defined as up to three weeks in advance.

Given some information about the flight and the time of request, we should be able to predict whether we should book the flight or wait.

Data and Features

I collected data from a major travel website, looking at five different destinations: Boston (BOS), Chicago (CHI), Portland (PDX), Los Angeles (LAX), and New York (LGA). For each of these five destinations, hourly requests were made to get all available flights in the three week range starting from that date. The date range of data collected is from 11/8 to 12/10.

I chose to use the current day of the week, the day of the week of the flight, the current time of day (categorized into four 6-hour time periods), the flight's time of day, the number of hours until the flight, number of stops, the duration in minutes, and the price of the flight as features. In some of the experiments, I used one hot labels to encode the day of week and time of day features as 7 and 4 different binary features, respectively.

The labels were 1 for "should wait" and 0 for "should buy." They were determined by whether the minimum price for a flight with the selected date and destination would decrease. . I chose to consider only the minimum priced flight for each request for a date and destination.

Models

I used the scikit-learn implementations of the following models:

Logistic Regression:

- hypothesis function h

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

SVM:

- Optimization problem
- Used RBF kernel

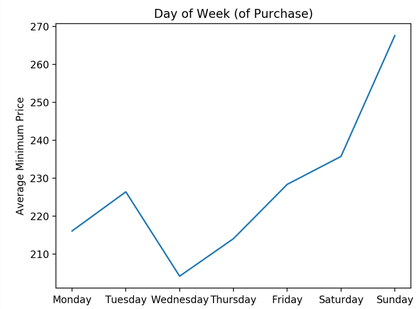
$$\max_{\gamma,w,b} \ \gamma$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \ i = 1,\ldots,m$$
$$\|w\| = 1$$

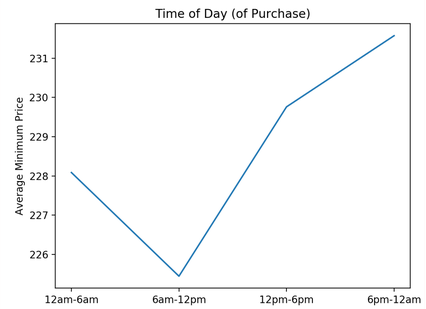
K-Nearest-Neighbors, Random Forest:

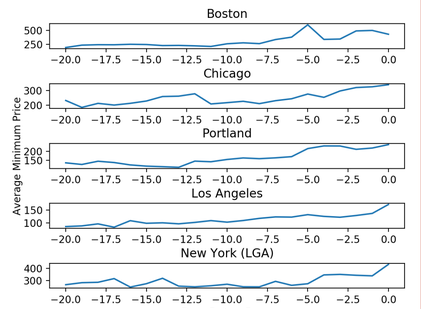
- k=3, L2 distance for kNN, n=18 for Random Forest

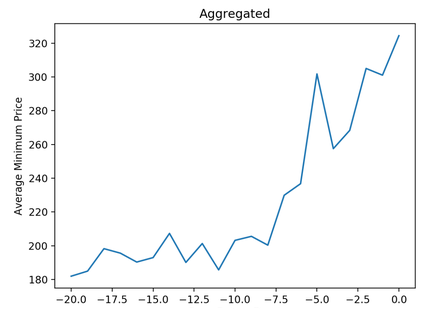
Results

Model	Training Set Accuracy	Test Set Accuracy
Logistic Regression (one hot labels)	69.1 (69.9)	69.0 (69.4)
SVM	98.2	94.1
kNN, k = 3	95.9	91.3
Random Forest, n = 18	99.3	95.2









Discussion

The initial model I tried as a baseline was logistic regression, which quickly showed that the data was likely not possible to model linearly, since the percentage of positives in the data was about 39%, so logistic regression had only slightly better performance than naively labeling all data points as negative (61% accuracy).

The classification models I tried (SVM, k-Nearest-Neighbors, Random Forest) performed well. However, the lower test set accuracies show that these models are overfitting the data. I also believe overfitting may have inflated the accuracies, since the test set data and training set data have overlapping points since data was randomly split but flight pricing does not change hourly and the data points I collected are hourly.

Future

The next steps are to segment the data further in order to reach more interesting results. For example, I may need to depart after a certain time, or prefer to take a flight with no more than one layover. I also want to continue to look at regularization and feature selection to avoid overfitting and get more robust results.