Ayo Opeyemi, Chase Davis, and Dino Rodriguez

TF: Christine Hwang

CS 109 / STAT 121A

Airbnb Suggested Pricing: Project Milestone #2

**Neighborhood and Price Prediction for San Francisco Airbnb Listings (Stanford):**

This research report describes efforts by Stanford researchers, Emily Tang and Kunal Sangani, to use text, images, and other forms of information provided in Airbnb listings to predict both the neighborhood of a listing within a certain city and to also predict the price of a listing. Tang and Sangani used data from the 7,029 Airbnb listings amongst 37 San Francisco neighborhoods. The data provided in this data set included text information (name, general description, bio of host, available local transit, and etc.), in addition an image thumbnail and other fields describing the amenities offered in a listing. The data set was refined to exclude neighborhoods in San Francisco where the total number of listings within a neighborhood was less than 70 listing, thus cutting our total number of listings to 6,764 amongst 27 neighborhoods.

For each listing, Tang and Sangani extracted 5 sets of features (listing information, multinomial bag of words, word class features, text sentiment features, and visual features). In addition, they download all images of listings and sampled 100 images to create a dictionary of words and then used k-means clustering to form 1000 visual words for the dictionary. They then used a Support Vector Machine and implemented sklearn’s SVM package. They noted that they had a slight overfitting problem and chose to address such by using Recursive Feature Elimination that recursively takes subsamples and eliminates features with small weights.

In the end, Tang and Sangani were able to predict housing prices with 81.2% accuracy (compared with the chance accuracy of 50%) and were able to predict neighbors with 42.2% accuracy. However, one shortcoming of their model included high variance, thus overfitting. This was revealed by the gap between the training and cross-validation scores.

**Predicting Airbnb Listing Prices with Scikit-Learn and Apache Spark (MPAR)**

Machine learning can be immensely useful for making assumptions about customer behavior which in turn is very useful for predictive recommendations for consumers. Nick Amato in this article dives into Python and scikit-learn package to examine an Airbnb dataset for San Francisco. He starts off by predicting list prices that a host might choose based on attributes for the listing – he also chooses to use Apache SPark for CV and tuning hyperparameters. Nick does a quick look at the dataset by examining features (of which there are more than 90) and then plotting the number of listings per neighborhood in San Francisco. He then cleans the NaN values from the set as necessary and other values that are out of the ordinary (0 for bedrooms, bes, or price). After cleaning, the dataset goes from 7029 points to 5246 entries. For categorical variables, he used the “get\_dummies()” method we’ve used in class to utilize one-hot encoding. He also handily uses the scatter\_matrix function to show each feature as a function of another which is a useful way to analyze collinearity quickly. After cleaning, Nick uses a bunch of scikit’s linear models to start fine-tuning: vanilla linear, ridge/lasso, ElasticNet, bayesian ridge, and Orthogonal Matching Pursuit. To compare between them, he chose to score using median absolute error because it's insensitive to outliers compared to alternatives. Bayesian ridge came out on top, but he chooses to use an ensemble method to get better results: though concedes dealing with the hyperparameters may be tricky. Scikit’s “GridSearchCV” exhaustively tries the parameter combinations and uses CV folding to find the best one. Obviously the brute force method isn’t the cheapest, but Spark comes in here to handle it across more machines. 500 iterations works well, and he decides to explore more hyperparameters while explaining the differences between MLlib in Spark versus Scikit’s version. He reduces the error down to $21.43 with this method and it takes much less time as well. When all is said and done, the most important features were accommodations, review\_scores\_rating, and shared room.