Predicting Airbnb Price Listings in New York City

Alexander Hurtado (hurtado@stanford.edu)

Introduction

In an era defined by a growing collection of digital information, we can leverage this data to better inform financial decisions, such as pricing. For this project, I focused on a public dataset enumerating Airbnb listings in New York City, along with their associated prices. These listings contain a plethora of information, including listing amenities, location context, host factors, and text descriptions. This project will attempt to tackle the task of price prediction using a shallow neural network trained on these features. Such a model could serve useful towards detecting price gouging, giving power back to the hands of consumers.

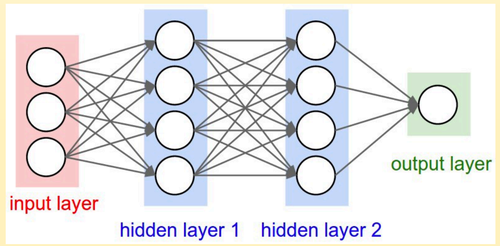
Dataset

The dataset provides a set of 106 attributes per Airbnb listing. Of these 106 attributes, we will focus only on 47 to extract features. Of the 47 attributes, 36 are either real valued (e.g. number of beds, host's response rate) or categorical (neighborhood, property type, host's verification status) and require little processing. From these 36 attributes, we can extract 90 features. The remaining 11 attributes contain text; these attributes are cleaned, tokenized, and their frequencies are counted.

We also took the liberty of casting the task of price prediction as a classification problem by bucketing the listings' prices into 5 buckets of approximately equal representation. These 5 buckets represent different ranges of prices, ranging from cheap listings (less than $79 per night) to expensive options (upwards of $200 per night).

Models

For this project, I trained three shallow 3-layer neural networks, each leveraging a different set of features. Each neural network was trained for 1000 epochs with learning rate decay via Adam optimizer. Each hidden layer had a size of 256. A general diagram of the shallow network is given below.



Baseline 1:
This baseline model took in two features: the latitude and longitude of the Airbnb listing. This model was meant to test the general capacity of the neural network to perform better than blind guessing.

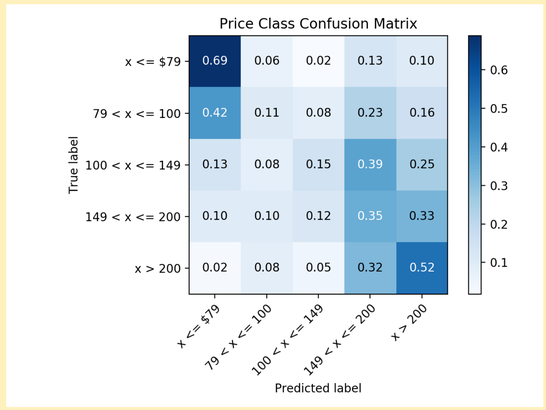
Baseline 2:
This baseline model took in 90 features extracted from 36 attributes that described various characteristics of the listing and its host. See "Dataset" for more information.

Advanced:
This model leveraged the full 90 features described in baseline 2 as well as ~10,000 features extracted from the textual data of the listing. In particular, the features represent a count of the 10,000 most frequent words that appeared in the listing's rules, summary, notes, description, etc.

Models	Accuracy	F1 Score
Latitude/Longitude Baseline	0.347	0.287
Listing Descriptor Baseline	0.443	0.397
Fully Features + Text Frequency Model	0.523	0.447

Discussion/Error Analysis

From our results, we see that the shallow 3-layer neural network incorporating listing, host, and text features performs the best on the task of price prediction. The performance of this model likely exceeds the ability of the baseline model because it better captures the characteristics particular to each listing. A confusion matrix of the model's class prediction is given below. From the diagram, we note that our model is rather good at predicting the extreme price classes, namely the less than $79 listings and greater than $200 listings. However, it does poor in correctly classifying listings in the $79 to $149 range.



Future Analysis & Work

In order to better understand how our neural network makes predictions on Airbnb listings, I plan on conducting an analysis on the relative importance of each feature towards price prediction. In particular, I will examine a gradient update on the input feature layer and note the sign and magnitude of the update for each feature.

I also plan on incorporating more advanced natural language processing techniques to extract features from the text data provided in the dataset.