Capstone Project Proposal - Predicting AirBnB Listing Price and Utilization in Toronto

Introduction:
AirBnB manages a short term rental platform, where people can list their apartment or home and allow others to rent a room or even the entire property for days at a time. In mid-2017, according to one article, they were on pace to serve over 100 million customers for the year, representing an estimated 25% of leisure travellers and 23% of business travellers. AirBnB, like other members of the relatively new and controversial sharing economy, has managed to tap into an inefficient market - rooms or entire homes that are unused and could be earning the property-owner (and sometimes, probably illegally, the renter) some extra income.

For those listing a property on AirBnB (referred to as "hosts"), one of the most difficult decisions they need to make is how much money to charge for their listing. If the price is set too high, they may be skipped over by customers looking for a better deal. If they charge too little, they could be leaving a great deal of income unearned. AirBnB provides a predictive pricing tool to hosts, but the tool has received complaints for not being accurate enough. Some hosts even believe AirBnB may be artificially decreasing prices to keep new customers coming into the platform. A small number of third party tools exist that attempt to solve the pricing issue, but these do not take desired utilization (the rate a location is booked) into account. My tool will seek give potential hosts the ability to see what different pricing outcomes will have on their utilization rate, and see the optimal price for the number of days they want to be booked.

Problem Statement:
I plan to develop two models. A regression model to predict the price of a listing, and a second ensemble model to predict overall utilization, which will use the predicted price as an input.. The first goal will be to provide a more accurate prediction of what price hosts should list their property at, based on a variety of factors gleaned from multiple datasets. Potentially relevant variables from AirBnBs own data can include property amenities, the number of reviews, the average rating, the description of the property, the number of guests allowed, the time of year, and more. There are already a handful of third-party pricing tools available, so I will seek to differentiate my model by joining AirBnB data to other sources to provide a complete picture of neighborhood popularity, or how "up and coming" it is, under the assumption that more popular areas will be able to set higher prices. This will be done by joining data from Yelp's academic dataset.

The second goal for my model will be to predict utilization rate - defined as the number of days the property is booked out of the days it is made available by the host. This model will be based on the recommended price, or a price the host provides, so they can perform a "what-if" analysis. This could be a harder problem to solve because the AirBnB dataset includes the number of days a location was booked, but not the number of days it was made available by the host, so the true availability may need to be estimated or based on certain assumptions.

Client Description:
The client for this project is any host placing a listing on AirBnB that wants a second opinion on the pricing recommendation AirBnB has provided. Hosts will be able to use this tool to determine the optimal price to set their listing at, depending on how many days they would like their listing to be booked and how much money they would like to earn.

Data Description:

The primary data to be used in this project comes from Inside AirBnB, a website that scrapes and maintains data from all listings on AirBnB and makes them available by city. This data comes in two files:

- Listings - detailed data on each listing in the city. Consists of 96 variables. These include details on the listing itself, such as neighborhood name, latitude and longitude (masked to within 500ft of the actual location), type of property, number of rooms and beds, and many more.
- Calendar - this is a narrow (only four columns) but long dataset which contains, for each day of the year, the price the listing was charging and whether it was available. This could be used for tracking price change seasonality as well as estimating overall availability for a given listing.

I also plan to make use of Yelp's Academic dataset to determine which neighborhoods have the most popular restaurants and businesses. This dataset contains information on businesses and restaurants in approximately 10 cities and is provided as a series of JSON files. I will make use of the Business file, because it contains number of business near each AirBnB listing and their average rating. These data points will be fed into my predictive models as features.

Approach:
I plan to follow the CRISP-DM methodology for this project, which stands for Cross-Industry Standard Process for Data Mining. This is made up of six phases, which are iterative.

- Business Understanding - this involves framing the question and learning as much as possible about the business and how the data to be used is collected. I have already framed the question, but I plan to do some research and learn how others have approached similar pricing problems.
- Data Understanding (EDA) - this phase involves performing exploratory data analysis, generating summary statistics and data visualizations, to look for interesting trends or patterns in the data.
- Data Preparation - data transformations are made in this phase. Here I will remove unnecessary fields and create derived variables, for example taking the listing amenities and separating them into multiple binary flag fields. I will also join the AirBnB dataset to the Yelp dataset, using either Neighborhood Name or Latitude/Longitude to join on.
- Modeling - the actual model-building is done in this phase. To predict pricing, which is a continuous numeric target, I plan to start with a simple linear regression, then move on to more complex models such as random forests and compare the results. To predict utilization I will calculate the difference between the predicted price and the actual price, to see whether a listing being over- or undervalued impacts how often it is booked.
- Evaluation - after building models against a training dataset I will run them against a test subset of data, which was removed prior to the modeling phase. I will need to do this twice, first to test the price predictions, then again to test the utilization prediction.
- Deployment - deploying the model will involve building an interactive tool where potential AirBnB hosts can enter an address and their desired utilization and view the optimal price to set their listing at.

Deliverables:
The deliverables for this project will include:

- The project report writeup
- A presentation and associated slide-deck
- Jupyter Notebooks containing all code used for data exploration, transformation, model-building, and evaluation
- If time allows - an interactive dashboard where users can enter relevant data points and view a prediction for their listing.