Project Report

Prepared for: Ilkay Altintas

Prepared by: Paul de Fusco, Sankarshan Acharya, Parinaz Azdavari

Advisor: Julian McAuley

January 15, 2018

MAS DSE Capstone Project Group 9

## Challenge

**Airbnb** is an American company which hosts an online marketplace and hospitality service, for people to lease or rent short-term lodging including vacation rentals, apartment rentals, home-stays, hostel beds, or hotel rooms. The platform provides the property host a set of features for the management and adverting of the rental, including a calendar associated with the unit to set pricing and availability.

The calendar user interface provides a recommended unit price for each available day based on its pricing algorithm. If a host sets a price that is deemed to be too high, the UI shows a red dot next to the calendar day price but no further details or statistics are shown. Furthermore, the details related to how the price recommendation is made are hidden to the host. While a host can browse similar listings in the area to see what prices are being advertised by competing hosts, the AirBnB platform doesn’t show any market information. Finally, it is well known that the price recommendation is often well below what could be charged, at times even multiple times lower when particular events causing unusual spikes in demand take place (e.g. the Comic Con convention).

AIRBNB unit Pricing model - mas des capstone project

The datasets will be described in more detail in the data sources section, but it is relevant to mention that AirBnB datasets are obtained by scraping the site and are not directly published by the company. As a result, while a price feature will be provided, it will include blanks for days when a unit is rented or taken off the market (without a distinction between the two). Furthermore, AirBnB purposely hides the exact address for rental properties, so the narrowest localization can be made by subsections of zip code areas.

## Opportunities as a set of questions

The shortcomings described above represent a set of opportunities which can be addressed with data science tools. To mention a few:

* What is the optimal price for a given rental unit at a particular point in time. The notion of optimality should be defined with a specified objective function and clearly determined set of constraints. Market factors e.g. competing unit prices, count, etc. should also be considered by the model. Reviews data can also contribute to the optimal price for a unit.
* Can additional datasets be integrated to:
  + Learn more about the units advertised i.e. associate listings with more features.
  + Learn more about expected demand for a particular time period.
  + Learn more about what the competition will do
  + Learn more about service fee that should be charged
* Additional opportunities exist in leveraging the reviews data to make predictions about each postings occupancy, which in turn might help determine optimal price for the unit.

## Data sources

The main data source for the project is the datasets found at [www.insideairbnb.com/getthedata](http://www.insideairbnb.com/getthedata)

At this link, datasets are provided for each major city/region in the world. For each, the following datasets are available:

AIRBNB unit Pricing model - mas des capstone project

* ‘neighbourhoods.geo.json’
* 'reviews.csv'
* ‘.DS\_Store'
* ‘neighbourhoods.geojson'
* ‘listings\_summ.csv'
* ‘neighbourhoods.csv'
* ‘stations.geojson'
* ‘listings.csv'
* ‘calendar.csv'
* ‘reviews\_summ.csv'
* ‘neighbourhoods.json'

The most salient of them are the listings and the reviews datasets. The listings contain information for each listing as it was posted by the host, including posting image data. The reviews contain a high amount of text comments regarding past stays. Calendar provides a fuller view of a unit’s availability, while neighborhoods can be used to better geolocalize postings.

**Additional Datasources**

Zillow Real Estate API: <https://www.zillow.com/howto/api/APIOverview.htm>

Gov foreclosure data: <https://catalog.data.gov/dataset/test-api>

Local government data: <https://github.com/sunlightpolicy/opendata/blob/master/USlocalopendataportals.csv>

More Real Estate API’s: <https://www.programmableweb.com/category/real%20estate/apis?category=20211>

More are WIP

## Approach

The main difficulty in modeling the price is caused by the fact that the listings field ‘price’ is at times blank. As mentioned above, the blanks relate to occurrences (dates) when the listing was either occupied or made unavailable by the host - but we are not provided information about which of the two it is. This problem makes the training data less than ideally reliable.

As a solution, the team has started dissecting the data for San Diego county and identifying listings that offer a fair mix of blank and non blank price field instances, preferably with patterns of frequent alternation of blank and non blank price instances within short time periods. By identifying listings that show a more informative pattern of activity the team will fill blanks with hypothetical prices under different scenarios/assumptions (market ‘temperature’ and additional environmental variables can be used in the model). In specific, the team will match vacant listings in the same market that exhibit similar blanks patterns, and/or compare similar listings in similar markets, and/or consider additional factors from other datasets. A large enough dataset can be constructed by applying the same methodology to similar markets and appending the results of this exercise into a final, more reliable version of training data.

AIRBNB unit Pricing model - mas des capstone project

We have initiated work on the above and will continue doing so over the next two weeks. We will then seek Julian’s advice to start creating models and considering the different mathematical approaches to the project objective.

## Team Roles and Responsibilities

Julian: Project Advisor

Responsibilities:

* Provides thought leadership, rates outcomes of development process and gives constructive feedback to help the team move forward.

Paul: Team Lead & Scrum Master

Responsibilities:

* Provides thought leadership and helps meet project objectives with development efforts and participation to discussion
* SCRUM Master: does anything possible to help the team perform at their highest level. This involves removing any impediments to progress, facilitating meetings, and doing things like working with the product owner to make sure the product backlog is in good shape and ready for the next sprint.

Sankarshan: Data Scientist & Scrum Product Owner

* Provides thought leadership and helps meet project objectives with development efforts and participation to discussion.
* Scrum Product Owner: has a vision of what he wishes to build, and conveys that vision to the scrum team. Manages subset of product backlog.

Parinaz: Data Scientist & Scrum Product Owner

* Provides thought leadership and helps meet project objectives with development efforts and participation to discussion.
* Scrum Product Owner: has a vision of what he wishes to build, and conveys that vision to the scrum team. Manages subset of product backlog.

## Project Coordination and Communication Plan

We have created a github repository, a slack channel, and will use email. We will probably adopt a softer version of agile methodology where we have frequent standup meetings where we cover the following:

* What I have done since the last time we met
* What I will do before the next time we meet
* Do I have any blockers

We will aim to have at least two standup per week, and meet on Saturdays to code together. Standups can be conducted via email if necessary.

## Bullets for each team member’s individual contributions in Step 1

Sankarshan and Paul:

* + Had calls and meetings to discuss and agree to the plan.
  + Created initial ipython notebook where they conducted data exploration with San Diego county datasets
* Parinaz joined a bit later but also participated to calls to discuss project scope and feasibility.