**In which variables you should focus to increase your Airbnb occupancy**

A data based approach using Kaggle Boston Airbnb Open Data from 2016-09-06 to 2017-09-05

Airbnb business info as introduction

For this analysis I want to answer three questions:

1. Which location-related variable generate a more accurate model?
2. Can the model be improved by doing a sentiment analysis of the reviews?
3. Which variable dimension is more important to the model: host, property, or reviews?

**Data Preparation and Exploration**

Kaggle Boston Airbnb Open Data have three files: calendar, reviews, and listings.

**Calendar**

Calendar data has 1,308,890 rows and consist of 4 columns: the linsting\_id that identifies the Airbnb property, the date goes from 2016-09-06 to 2017-09-05, available that have an “f” if the property is not available and a “t” if the property is available, and the price if the listing is available.

**Tabla

Descripción generada automáticamente**

I used calendar data to calculate the occupancy percentage of each listing, applying the next formula per listing: sum(available “f”)/ sum(available “f”) + sum(available “t”). We can see the occupation percentage doesn’t follow a normal distribution, we have a high density in the low and high end of the percentage.

Gráfico, Histograma

Descripción generada automáticamente

**Reviews**

Reviews data has 68,275 rows and consist of 6 columns: the linsting\_id column is the property described by the review, the id identifies the review, the date goes from 2009-03-21 to 2016-09-06, the reviewer\_id identifies the Airbnb user who made the review, reviewer\_name is the name of the user who made the review and the comment the review itself .

Tabla

Descripción generada automáticamente

The review doesn’t have a quantifiable value, so I used the Amazon Comprehend service to process the comments. Amazon Comprehend is a natural-language processing (NLP) service that uses machine learning to uncover valuable insights and connections in text.

For this analysis, the detect\_sentiment function was used, this function receives text, in this case the review, and returns a score for neutral, positive, negative, and mixed sentiment. Finally, all the values were averaged by sentiment and listing\_id, and the information was saved in a csv file to be used in the analysis.

Tabla

Descripción generada automáticamente

**Listings**

Listings data is the main data source, it describes the Airbnb property and the host of that property, it contains 95 columns and 3585 rows. Not all columns are relevant for analysis, 49 columns were selected and divided in three dimensions: host, property, and reviews.

Here is a dictionary with all the select columns and the dimension to which they belong, more information about what each column means can be reviewed here.



**Data Imputation**

13 columns have more than 1% missing values, for these columns the following strategies were applied.

Tabla

Descripción generada automáticamente

-Drop columns: square\_feet, security\_deposit and cleaning\_fee because they have a very high percentage of missing values and are not critical variables for the analysis.

-Review columns, host\_response\_rate and host\_acceptace\_rate have two to nine percent of missing values, but they are important columns, so they will be filled with the mean.

-Host\_response\_time have 8.3 percent of missing values but is an important column so I will use a dummy nan column for the missing values.

-Columns: zipcode, bathrooms, bedrooms, host\_location, beds and property\_type have a very low percentage of missing values so for those columns the rows that contain nan values will be removed.

**Data cleaning**

Most variables are already continuous or categorical and can be processed using dummy variables, with the following exceptions:

For host\_response\_rate, host\_acceptance\_rate, price and extra people, characters like $ and % must be extracted to convert the values to continuous.

**Data Modeling**

The predict variable doesn’t follow a normal distribution and listing data have a lot of categoric variables, also, I want to understand which variables are more relevant to the model. For these reasons and because it tends to provide a high level of accuracy, I chose to use a random forest classifier algorithm.

(Maybe quickly explain the algorithm)

To use this algorithm first I convert the continuous output variable to categoric. Explain range an d categoric variable

**Which location-related variable generate a more accurate model?**

For this question there are three possible variables to use: zip code, coordinates (latitude-longitude) and the neighborhood. I create one dataframe per location variable with all the other select variables, then used the random forest classifier algorithm 20 times per dataframe, obtain the accuracy of the model and average the values.