Airbnb - New York's Analysis

Jackson A. Prado Lima
Department of Computer Science, Federal University of Paraná
CP:19081, CEP: 81531-980, Curitiba, Paraná, Brazil
japlima@inf.ufpr.br

ABSTRACT

Airbnb is an outstanding hosting service used worldwide and provides information to its users in a simple and efficient manner. In this work, an analysis is made taking into account if incidents and cultural organizations influence the prices practiced in Airbnb. To help in this task, the present work performs an analysis using charts, maps and different types of machine learning. The data reveals that Queens district has a fair trade-off among price, cultural organization, and incidents, as well as, the accommodates diversity. Besides, the cultural organization has a high influence in to predict price.

1 INTRODUCTION

In some moments, the guests want to a local to be cheaper or more attractive considering local events to appreciate, besides a safe local is arguably necessary. Considering the Airbnb listing from New York City (NYC), how the cultural organizations and incidents impacts on listing price?

The paper is organized as follows. Section 2 describes how the experimental evaluation was conducted. Section 3 presents and analyses the results obtained. Section 4, describes the problems found during the work. Finally, Section 5 concludes the paper and discusses the future work.

2 EXPERIMENT DESCRIPTION

The hypothesis of this work[1] is that our approach is capable of found the influence of cultural organizations and incidents in to predict listing price as well as recommending a cheap and safe local (district/borough).

According to our goals the experiment was guided by the following research question: "How the incidents and cultural organizations impact on listing price?".

To answer the research question, we compared the result shows on maps, a collinearity matrix among the features, and how the accommodates are distributed. Additionally, we used a machine learning with seven regression models, and the results were compared using Median Absolute Error. Besides, two algorithms for classification were used to see the difficult to predict the most influent features.

2.1 Target Data

The investigation focused on four datasets collected in open data sites. These datasets are as follows: 1) Listings, a dataset that is an Airbnb snapshot of 02 April 2017 from NYC; 2) Incidents, contains all incidents from NYC for the year 2016; 3) Cultural Organizations, contains all cultural organizations from NYC in 2017; and 4) NYC Borough, contains information about boundaries of boroughs (districts). Further information about these datasets is available in Table 1.

Table 1: Data sets used in the experiment

Dataset	Description	Size	Format
Listings [1]	Detailed listings data, including various attributes (features) of each listing such as number of bedrooms, bathrooms, location, etc.	153,7MB	csv
Incidents [3]	All valid felony, misdemeanor, and violation crimes reported to the New York City Police Department for all complete quarters in 2016.	124,2MB	csv
Cultural Org. [4]	Listing of all Cultural Organizations in the Department of Cultural Affairs directory.	333,3kB	csv
NYC Borough [2]	Polygon boundaries of boroughs (water areas excluded).	400,5kB	GeoJson

3 RESULTS AND ANALYSIS

This section shows and discusses the results to answer the research question.

Figure 1 shows, per district, a matrix of each feature as a function of another which is useful to check for any collinearity among the features. The cells running through the diagonal of the matrix contain a histogram with its values on the X axis. The features chose to be analyzed were: price, cultural organizations, and incidents.

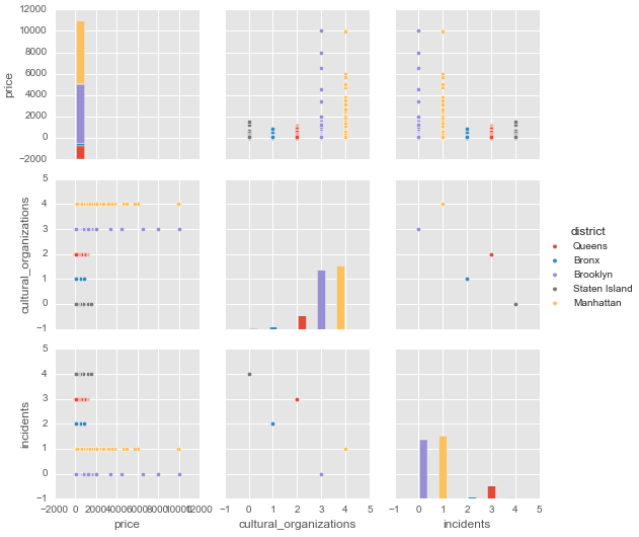

Figure 1: Collinearity Among the Features

[1]The project is available at https://github.com/jacksonpradolima/airbnb-ny-analysis

Based on the figure, we do see that Queens and Bronx are in the trade-off among the features. To help in to check this observation, Figure 2 shows that Manhattan district has the highest listings mean price, and Queens and Bronx are the lowest. In Figure 3, we can observe that Manhattan stands out followed by Brooklyn. On the other hand, Brooklyn stands out followed by Manhattan when we consider incidents (Figure 4). In this sense, we can affirm that Queens and Bronx have the best trade-off among mean price, cultural organizations and incidents.

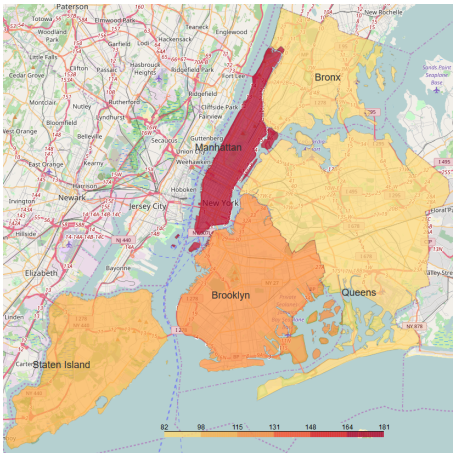

Figure 2: Collinearity Among the Features

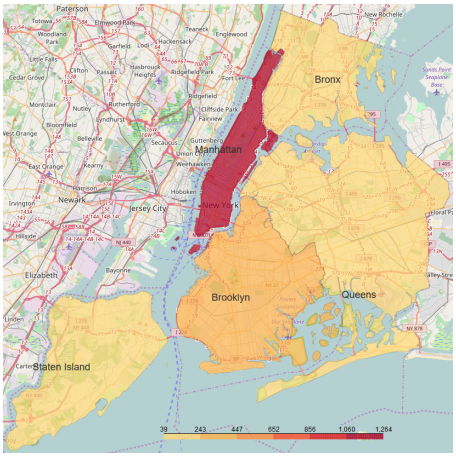

Figure 3: Collinearity Among the Features

Figure 5 shows mean price grouped by the number of people that can be accommodated (number of beds) with a limited of four accommodates, when we can begin see differences, per district: Queens, Manhattan, Bronx, Brooklyn, and Staten Island. We can observe that Manhattan's district has more options than other ones. It's interesting to highlight Queens district outperformed the Bronx in the diversity of accommodates.

Next, we used machine learning and for this we split our dataset into an 80% train and 20% test. In this step, we run seven regression

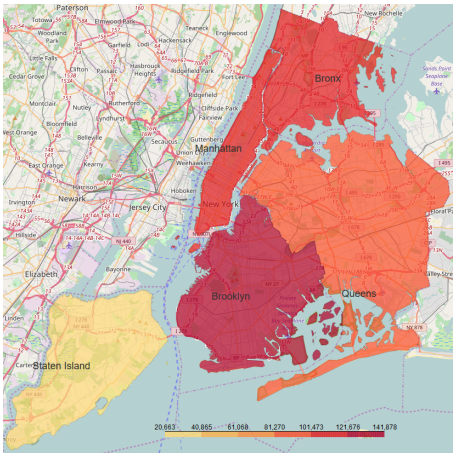

Figure 4: Collinearity Among the Features

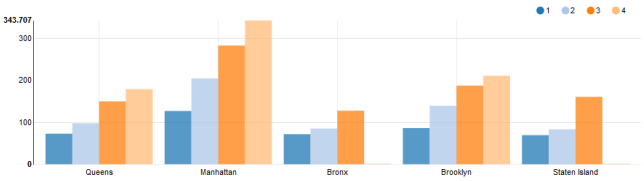

Figure 5: Mean price grouped by the number of people that can accommodate

models: Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Bayes Ridge Regression, Orthogonal Matching Pursuit (OMP), and Gradient Boosting Regressor (GBR); and two algorithms for classification: Decision Tree and Random Forest. First, we compared the regression models, where GBR was tunned with 300 iterations and used an exhaustive "grid search" which only tries all the supplied parameter combinations and uses cross-validation folding to find the best one.

Figure 6 shows the regression models their Median Absolute Error. Intuitively, Median Absolute Error is less sensitive to outliers than Mean Squared Error and translates nicely to a dollar amount that is about price.

When analyzing these estimators by their Median Absolute Error, we see that GBR outperforms the other ones with a Median Absolute Error of $23.86, approximately 31% less than Lasso Regression that appear in second with a Median Absolute Error of $34.59.

After, we calculate the feature importances to see which features were most influential in predicting the listing price, Figure 7. This show a relative scoring of how important each feature is about the feature with the most importance.

Clearly, some of the variables have more influence than others, and the most important feature is the "Entire home/apt" attribute; this indicates whether or not the unit is shared with other people, and has the most effect in setting the price. This feature is followed by accommodates, cultural organizations and Manhattan district, and the feature incidents were not so influent.

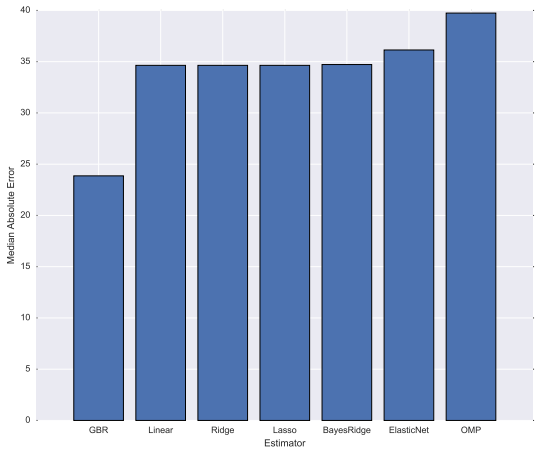

Figure 6: Price - Median Absolute Error

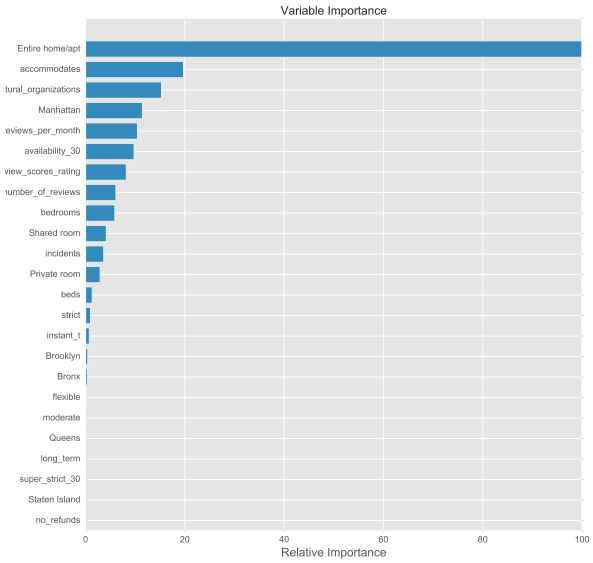

Figure 7: Gradient Boosting Regressor - Feature Importance

In the last, we tried to predict the three most influent features using algorithms for classification. Table 2 shows in bold the best test error values.

Table 2: Classification by Feature Importance

Feature	Decision Tree	Random Forest
Entire home/apt	**0.0000**	0.0009
Accommodates	**0.4115**	0.4151
Cultural Organizations	**0.0000**	**0.0000**

Analyzing the results, we found two districts in the trade-off among price, cultural organizations and incidents: Queens and Bronx; where comparing these districts Queens outperform the Bronx in the diversity of accommodates. In relation the regression models, we can conclude that the local is not so influent in to predict the price, but the room type. In relation the algorithms for classification, the Decision Tree outperforms Random Forest and the feature with more difficult to predict was "Accommodates". Besides, the cultural organizations have a strong relation with the data where there was not test error. Thus, we can conclude that cultural organization has a strong influence on listing price and where there are more cultural organizations is more expensive and consequently attracts more incidents.

4 PROBLEMS FOUND

We identified some points that can be problems found during the work: (i) The listings data set had problem to read it, and a workaround was done using R language; (ii) The Apache Spark regression evaluator has not the median absolute error metric, for this reason it was used sci-kit learn library.

5 CONCLUDING REMARKS

This work introduces an analysis trying to discover if the cultural organizations and incidents influence on listing price. The idea is to check the collinearity the features, try to predict the price using regression models, verify the level importance of the features chose and classify them using algorithms for classification.

We implemented and evaluated the New York City in relation with the Airbnb listings. The results show that where it has a more cultural organization, it has more incidents probability. Using Gradient Boosting Regression, we were able to fit a model that obtained a Median Absolute Error of $23.91 for all listing data. Besides, the cultural organization was the third feature more influent in the prediction of the price. In relation the classification using Decision Tree, we were able to classify the features with the minimum test error.

For a possible predictor to be used in practice, future work will need to be done to explore further and build more suitable models. Predicting a significantly skewed right response variable, price, yields a set of challenges that need to be addressed rather than predicting a particular price. Future works could consider population density, property price, and analysis by neighborhood.

REFERENCES

[1] Inside Airbnb. 2017. Listings. https://goo.gl/dT6m7H. (2017). Acessed Apr 2017. Date Compiled April 2, 2017.
[2] BetaNYC. 2015. NYC Borough. https://goo.gl/gWP1yp. (2015). Acessed Apr 2017.
[3] New York City Police Department. 2016. Incidents. https://goo.gl/gzZFg2. (2016). Acessed Apr 2017.
[4] Department of Cultural Affairs. 2017. Cultural Organizations. https://goo.gl/5jHzxa. (2017). Acessed Apr 2017. Last Update April 7, 2017.