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, although this information is limited only to listings of its products. In this work, an analysis is made taking into account if incidents and cultural organizations influence the prices practiced in Airbnb. Besides, the prediction of new prices is carried out that the user can come to inform taking into account information of the current listings. The price forecast has some advantages when advising the user: on the price practiced (is too large or small), the mode of advertisement, or even on the investment decisions of the market analysis. To help in this task, the present work performs an analysis using graphs, maps and different types of machine learning. The data reveals that Queens district has the more adequate trade-off among price, cultural organization, and incidents as well as the accommodates diversity. Besides, cultural organization has a high influence in to predict price.

1 INTRODUCTION

Airbnb was founded in August of 2008 and based in San Francisco, California. Airbnb is an online marketplace and hospitality service, enabling people to lease or rent short-term lodging including vacation rentals, apartment rentals, homestays, hostel beds, or hotel rooms.

Airbnb has help a lot of people, but how can we predict a price or if the cultural organizations and incidents impacts on listing price? In some moments, the guests want to a local cheap or more attractive considering local events to appreciate. Thus, predicting a cheap and safe local is arguably important. In this paper, we try and capture this characteristic using maps, charts and machine learning to help the guest in the chose.

The paper is organized as follows. Section 2 describes how the experimental evaluation was conducted. Section 3 presents and analyses the results obtained. 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 to recommend a cheap and safe local (district/borough) as well as consider the influence of these aspect in to predict listing price. To evaluate this, we consider uses maps, charts and machine learning.

According to our goals the experiment was guided by the following researches questions: **RQ1:** "How the incidents and cultural organizations impact on listing price?" and **RQ2:** "How the incidents and cultural organizations impact on regression model to predict the listing price?".

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

To answer the RQ1 we compared the result shows on maps, a collinearity matrix among the features, and how the accommodates are distributed. To answer the RQ2 we used a machine learning with 7 regression models and the results were compared using Median Absolute Error.

2.1 Target Data

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

Table 1: Data sets used in the experiment.

Dataset	Description	Size	Download
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 (NYPD) 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 in order to answer the researches questions.

3.1 Answering RQ1

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 feature. 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.

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(a) shows that Manhattan district has the highest listings mean price, and Queens and Bronx are the lowest. In Figure 2(b), we can observe that Manhattan stands out followed by Brooklyn. On the other hand, Brooklyn stands out followed by Manhattan

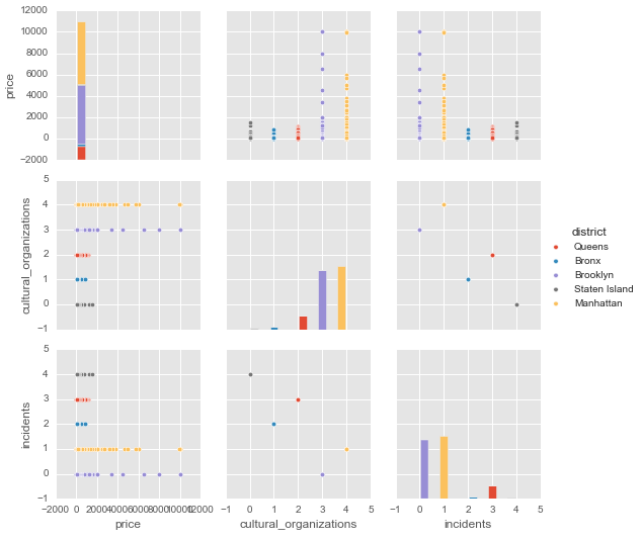


Figure 1

when we considers incidents (Figure 2(c)). In this sense, we can to affirm that Queens and Bronx have the best trade-off among mean price, cultural organizations and incidents.

Figure 3 shows mean price grouped by the number of people that can accommodate 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 highlight Queens district outperformed Bronx in the diversity of accommodates.

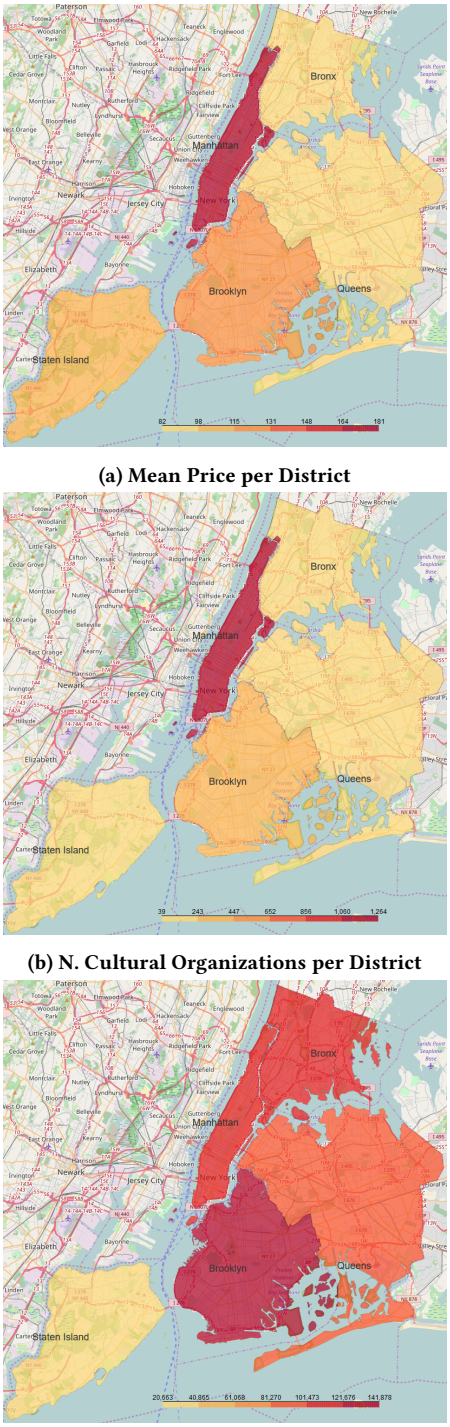
Analyzing the results, we can conclude that where it has more cultural organizations is more expensive and consequently attracts more incidents. However, we found two districts in the trade-off among price, cultural organizations and incidents: Queens and Bronx; where comparing these districts Queens outperforms Bronx in the diversity of accommodates.

3.2 Answering RQ2

For this analysis using machine learning we split our dataset into a train and test (80/20). We run 7 regression models to calculate the Train and Test Score: Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Bayes Ridge Regression, Orthogonal Matching Pursuit (OMP), and Gradient Boosting Regressor (GBR). In the analysis, we used 6 regression models and compared with GBR using an exhaustive "grid search" which simply tries all the supplied parameter combinations and uses cross-validation folding to find the best one.

Figure 4 shows the Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Bayes Ridge Regression, OMP and 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 relative to price.

When analyzing these six estimators by their Median Absolute Error, we see that they appear to be roughly the same with most of the estimators being able to predict the price with a median error



(a) Mean Price per District

(b) N. Cultural Organizations per District

(c) N. Incidents per District

Figure 2: Data Distribution per District

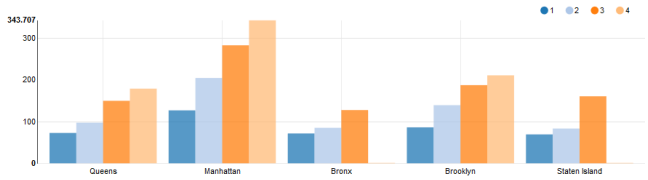


Figure 3

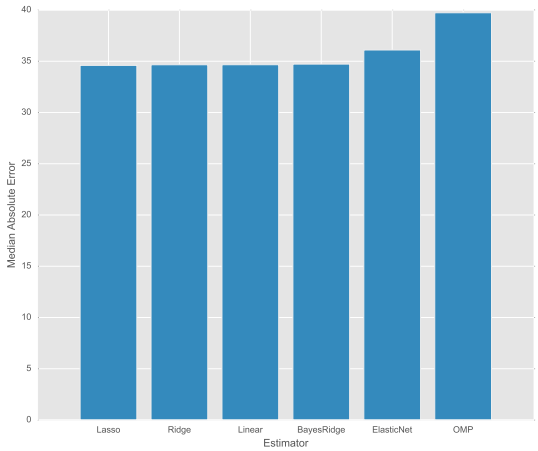


Figure 4: Price - Median Absolute Error

around 30-35 dollars, with Lasso regression outperforms other ones with small margin and a Median Absolute Error of $34.59.

We next fit an ensemble method, GBR, and compare its results to that of our previous models. We run it for 300 iterations and the result is a Median Absolute Error of $23.91, approximately 31% less than Lasso Regression. We see that GBR, in regards to Median Absolute Error, outperforms Lasso Regression.

After, we calculate the feature importances to see which features were most influential in predicting the listing price, Figure 5. This show a relative scoring of how important each feature is relative to the feature with the most importance.

Clearly some of the variables have more influence than others, and the most influential 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 was not so influent. In this sense, we can conclude that the local is not so influent in to predict the price, but the room type.

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) Organizing the data to create a possible research question and answer it; and (iii) How

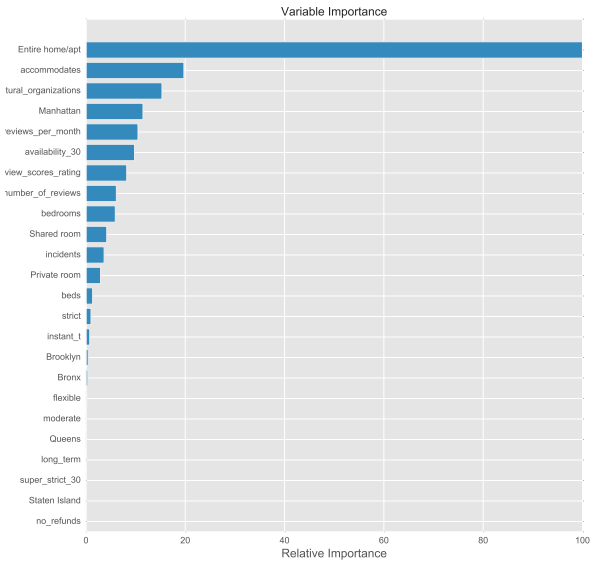


Figure 5: Gradient Boosting Regressor - Feature Importance

to use machine learning to show something that could not be answered, i. e., using charts, maps or display the raw data in the DataFrame provided by Apache Spark.

5 CONCLUDING REMARKS

This work introduces an analysis trying to discovery if the cultural organizations and incidents influence on listing price. The ideia is to check the collinearity the features, try to predict the price using regression models, and to check the level importance of the features chose.

We implemented and evaluated the New York City in relation the Airbnb listings. The results shows that where it has more cultural organization it has more incidents probability. Besides, the cultural organization was the third feature more influent in to predict the price. Using Gradient Boosting Regression we were able to fit a model that obtained a Median Absolute Error of $23.91 for all listing data.

For a practical predictor to be used in practice, future work will need to be done to further explore 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 specific price. Future works could consider population density, property price, and an 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.