DSC 540

Final Paper on Airbnb dataset, New York City
Paripon Thanthong

Abstract

The goal in this paper is to classify and predict availability of the listing room specifically in New York City areas including all five neighborhoods which may potentially help travelers finding the right area to rent on Airbnb. In this research, multiple machine learning algorithms are applied including decision tree (QUILAN, J.R. (1986)), random forest (Breiman, L.,2001) and boosting for both regression and classification to compare results. Wrapper feature selection is also applied throughout the research to find the best set of the model. As a result, both regression and classification show low accuracy around 30% and high error which I conclude that this approach and method I used, may not bring up the suitable model to predict or classify accurately.

Introduction

Background

In the hospitality world, advance technology company like Airbnb has played an impactful role to make it easy for travelers to explore the world. Airbnb has connected people around the world through their platform which help people access to local areas. Since it was founded in San Francisco in 2008, it has rapidly growth the community with now over 500 million guests arrivals all-time and over 7 million and listing room worldwide in over 191 countries (Airbnb, 2019). Travelers have selected Airbnb to expand on traveling possibilities and opportunities to experience the more unique and personalized way of vacation or trip. Furthermore, since the technology is an important part to people especially young people who eager to travel, Airbnb is disrupted the hospitality world which it takes a huge part of the market share from the leader in this hotel business in the past years.

The pricing and location are the most intriguing that make people abandon their safety and convenience living in hotel and seek to connect with strangers. However, Airbnb still has a huge problem that the rate of host cancelling reservation at the last minute is rising and also the safety problem and poorly property management which sometimes led to tragic (New York Times). Yet, it is still irresistible by some specific location and style that make it a huge different even if it may come with risk.

Objective

Since New York city is one of the biggest cities in the world, the problem of listing availability seems not to be reliable. For example, some host have never updated their number of availabilities of their renting property or those number are made up. Moreover, the knowledge of the area that sometimes cause traveler a problem.

In this paper, the dataset is about host, geographical availability and metrics number in New York City which is represented from Airbnb within 2019. As for the prediction and analysis through the process, machine learning algorithms are applied to this dataset to try to learn about the trend of traveler using Airbnb from the past year and try to learn from the traveler's review and rating to evaluate which areas in New York city have the busiest schedule and try to predict the availability for room listing in certain area.

literature review

Over a decade, Airbnb has dominated the hospitality field that impact leader of in hotel business with reported that over 2,000,000 listing room and 60,000,000 guests in 2016 The rapidly grown popular alternative for travelers all over the world. Given the growth rate of Airbnb, many people have become host by listed their own place without knowledge of hospitality or service ability. Because of this, travelers pay close attention to other traveler review on the listing.

One study found the important features that related to Airbnb room listing price (Gibbs, Guttentag, Gretzel, Morton Goodwill ,2018). The research was focused on the five metropolitan area in Canada included over 15,000 Airbnb listings. The process tries to predict pricing on Airbnb platform by applied hedonic regression analysis. What they found was interesting that a few of their selected feature were related to my work. The author claimed that those features, review, the type of room, location, rating, also have an effect on the price. In addition, there is some limitation on the paper for example, the structure of analysis they used compare to the structure of Airbnb analysis which needed to adjust in the future work.

Additionally, other research tried to tackle some similar goal in using Airbnb dataset which was about the travelers desired destination (ULFSSON.H, 2017). This paper applied XGboost and boosted decision tree (J.R. Quilan) machine learning algorithm to multiple set of features. The dataset that they collected was more toward to the user information. The major flaw in this research was how features were selected using in their research. The author claimed that there were 3 session of dataset using in the paper, but they did not perform any type of feature evaluation or features selection method which can support final result. In conclusion, the author made 3 attempts to analyze the dataset, and got a high score at 0.877. For the second and third try, the author stated that it scored lower since they decrease the number of features. In my opinion, even if the scores seem solid the authors should state which type of matrix they use for particular regression.

Furthermore, users rating has been used heavily for many analysis purposes but sometimes that reviews is made up. In Zhao, Dong and Yang research, sentiment analysis, which was text analysis, to extracted Chinese hotel review. The classifiers that used in this research were Naïve Bayes, ME (Maximum Entropy) and SVM (Support Vector Machine) to find the suitable algorithm for the aspect extraction from review. Overall performance showed high scored with ME have the highest at 93.47%, and 91.27 % for ME combined with TF-IDF method. Thus, the author claimed that algorithms use in this research was all suitable with score above average. With the high performance of text analysis come in to play, one of the researchers applied sentiment analysis to predict the Airbnb price using machine learning (Kalehbasti, Nikolenko and Rezaei, 2019). The dataset was and Airbnb in New York city including with 50,000 data points and 96 features. 5 machine learning algorithms were applied predict room listing price. Sentiment analysis was also applied to customer review which it analyses text, and the assign score. The matrixes they presented was error terms, MAE and MSE, also R square score. Thus, support machine learning performed the best compare with others algorithm.

Finally, in socio-economic analysis of Airbnb in New York City (Dudas, Vida, Kovalcsik, Boros, 2017) the goal of their research was to analyze the spatial distribution of Airbnb listing by indicate the socio-economic condition in the area. Linear regression was the main analysis they used and select feature manually from the p-values threshold. In conclusion, they stated that Airbnb accommodations (Airbnb dataset) and number of reviews were

concentrating in the area that has young population, number of housing and number of points of interest.

Methodology

Data

The dataset is Airbnb specifically in New York City area from Kaggle.com without timeline constraint within 2019, but the last date travelers review. It represents information about Airbnb information in New York City neighborhoods and some metrics such availability number and review information. It includes 16 features which is mostly about the host information, geographical, type of room, range of price, review metrices with 48,895 data points. The dataset has both categorical and numerous data points which some of them may not mean anything. For example, name of the listing and host name cannot be interpreted anything useful for the analysis.

Table 1. Description of Variables

Variable	Description
Id	Listing Id
Name	Name of the listing
Host_id	Host id
Host_name	Name of the host
Neighbourhood_group	Neighborhood area
Neighbourhood	Area (sub neighborhood)
Latitude	Latitude coordinates
Longitude	Longitude coordinates
Room_type	Listing space type
Price	Price in dollars
Minimum_nights	Amount of minimums night
Number_of_reviews	Number of reviews
Last_review	Latest Review
Reviews_per_month	Number of Review per month

Data Cleaning and Preprocessing

To begin with, there is around 20,000 missing data points which the majority of it is in 2 features. Features are initially selected based on intuitive knowledge of the dataset that will be used in the analysis. One of the features that has about 10,000 missing is dropped which is 'last_review'. It represents the last date travelers review the listing room. After that, another 10,000 missing from 'review_per_mounth' is filled with zeros to keep the most data for analysis. Also, features such as 'id', 'name', 'host_id', which is an individual information are also drop from using for analysis. Because there are two features that indicate the neighborhoods which are 'neighbourhood_group' and 'neighbourhood', so I decided to drop out the 'neighbourhood' since it has 221 neighbor names. Moreover, feature 'availability_365' is the target variable in this analysis which will be assign to a separate variable and dropped out of the main data frame. As a result, 7 total features will be used for both exploration and analysis (applied in to machine learning).

Since, 2 features are categorical text term which It cannot use in the model. These two features are converted to numerical.

Data Exploration

In data exploration, geospatial visualization, statistical plotting and exploration will be computed to see pattern, relation and the insightful information with in this set of features. First, the five number of summaries in statistic is computed to see initially how the data look like. The Pearson's correlation table is plotted to see the relation between feature only for the numerous values. Next, histogram plots for 'price' and 'availability_365' are computed for variable to see the distribution of those features if it needs to be transformed. Another plot for distribution is pie chart which the focus is on feature 'neighbourhood_group' and 'room_type'. The neighbourhood group shows Manhattan and Brooklyn neighbor hoods dominating the New York City about 40% each from the whole dataset. For room type, listing entire room and private room taking about 45% and 52% respectively from the whole dataset.

Additionally, I used latitude and longitude combined with other features to see the spread through the shape of New York City such as the spread of room type and price. Also, heatmap through folium package in python is computed with latitude and longitude with host listing count and review per month to see the density in the map. Still, the most density of both plot point to Manhattan and Brooklyn areas.

Finally, pair plot is computed with all features to try to find the pattern through scatter plot which it shows all feature plotting against each other. As a result, we can barely see any possible pattern and trend that could be useful in this set of features.

Low Variance Filtering

From the feature selection in sci-kit learn module in python, VarianceThreshold method is applied to the pre-selected features. The variance threshold is set at 0.7 which mean the computational will show the feature that has variance lower than 70%. Thus, two features show less than 70% of the variance which is neighbourhood_group and room_type. Thus, 5 total features with variance 70% are applied to feature selection process.

Feature Selection

In this research, Wrapper feature selection method is applied through algorithm. The wrapper is one of feature selection type which it applies the selection through machine learning algorithm. So, the wrapper feature selection is applied through 6 machine learning algorithms including decision tree regressor/classifier, random forest regressor/classifier and boosting regressor, gradient boosting and ada boosting.

Applied Machine Learning and Model Evaluation

The data evaluation tests on 2 type of methods which are train/test evaluation and cross validation. For the train/test, preprocessed dataset is split to 80% train and 20% test with feature availability_365 as a target variable. And, cross validation performs 10-folds cross validation.

For the algorithm conducting in this paper, Decision Tree, Random Forest and Boosting are applied to dataset with wrapper feature selection. For the Decision Tree (QUILAN, J.R. (1986)) and Random Forest, both regression and classification type of algorithm will be used verifying the main purpose of this project and hope to learn insightful information from this dataset. In addition, boosting algorithms, Gradient Boosting Regressor and Ada Boosting Regressor, are used only with regression method to see the change in different algorithm.

Matrixes

As for the performance score, at least 2 scores matrixes are computed in each algorithm. With the regression algorithms, RMSE (Root Mean Square Error) and explained variance are computed to see the performance of the algorithm. For the classification algorithms, accuracy score and confusion matrix are included throughout the research.

Result

Decision Tree Performance

To begin with, decision tree regressor and classifier are applied to predict the availability_365 variable. As for the decision tree regressor, three features are selected with wrapper feature selection, including price, review_per_month and calculated host listing count. In this method, the dataset is evaluated in both splitting dataset and cross validation. RMSE shows score in the same range with variation which is around 143.46 for train/test split dataset, and 144.79 for cross validation. For the explained variance, both model evaluation type has significantly low explained variance at -.18 and -.23 with variation. Another point that it is worth to mention is the CV runtimes in train/test split dataset is much faster compared to the cross-validation method.

Secondly, decision tree classifier with wrapper method selected 2 features for analysis which are price and review_per_month. Since, the class in target variable has not been binned before, there is 365 for this method. Accuracy score is computed, and it depict the low accuracy around 0.28. Confusion matrix is also included for F1 score for the multi-class situation. In this case, it's not quite reliable since there are too many classes in this problem.

Random Forest and Ensemble Method

Random Forest regressor and classifier are being applied to see if there is any improvement in term of performance and ability to predict. With wrapper feature selection, same features from decision tree classifier which is price and review_per_month. For the regressor method, bagging (Dietterich, T. ,2000) with split data evaluation and random forest with cross-validation show improvement on the error term and explained variance which RMSE at 122.68 and 123.15 with variation respectively. Explained variance also improve to .13 and .12 respectively. Between these two methods, runtimes on random forest with cross-validation is significantly higher than decision tree ensemble, at 52 second.

Moreover, random forest classifier with wrapper feature selection select the same set of features from decision tree regressor with 3 features. The performance in this method is not adequate, with the accuracy score is only .29.

Boosting Performance

For boosting, two type of boosting, gradient and ada boosting regressor, are applied. To begin the analysis, the dataset is pre-processed with normalization method in both features and target variable. Also, the target variable is binned in to 2 bins. After that low variance filtering and wrapper features selection are applied and select 2 features to be used in the analysis. As a result, both gradient boosting and ada boosting show an improvement from decision tree and random forest at RMSE 115.94 and 117.84, explained variance at 0.23 and 0.21 respectively. The runtimes is relatively close at 3.6 second and 2.3 second.

Table 2. Result

	RMSE	Explained Variance	CV-Runtimes
Decision Tree (Split Data)	143.46	-0.18	0.12
Decision Tree (CV)	144.79	-0.23(+/- 0.23)	1.16
Decision Tree (Bagging)	122.68	0.136	1.4
Random Forest (CV)	123.15	0.12(+/- 0.18)	52.86
Gradient Boosting (CV)	115.94	0.23(+/- 0.16)	3.66
Ada Boosting	117.84	0.21(+/- 0.13)	2.32

Discussion & Conclusion

The goal of this project is to predict and classify the availability of room listing in New York City by applied multiple machine learning methods. In regression, the purpose is to find and average availability date and classification for classify if the listing will be available. Wrapper selection is applied in this project to get an optimal set of features. Through performance matrixes including in this project, all algorithm does not perform well with less accuracy, high error and low explained variance with this set of features.

In regression method, all four algorithms do not generate any convincing matrix performance that can be used to build a model for prediction. It shows consistency high error and low explained variance. This may be the target variable does not suitable enough to use since it is number of dates that listing room is available, so it can be 1 to 365. Moreover, the features in this dataset seem to have a close meaning such as ‘number of review’ and ‘review per month’. In addition, over half of the feature cannot be used for analysis since there is no meaning behind it, for example, ‘id’, ‘name’ and ‘host name’. And, the correlation between a pair of feature shows that there is no strong relation between any of features that being selected for analysis. Because the structure of each algorithm, it obviously shows consistency improvement at the end.

As for the classification, same target variable as in regression is used and the performance matrix shows relatively the same trend. It has a low accuracy score around 29% in both random forest ((Breiman, L.,2001)) and decision tree. The problem is I have not discretized the target variable before putting in the algorithm. So, it will try to classify 365 class which I believe might be the cause of low accuracy. Also, confusion matrix is computed, since the target variable is not binary class, to observe the F1 score which it also has a very low weight.

Through the process, finding and result in this project, the result may not be adequate to predict and classify as of now. However, I still see the potential of an improvement and do it in the different direction.

Future Work

For the future work, there is a few things that I think it may improve the performance for this project. First, collecting more information both data points and useful features. Second, conducted a feature engineering for a meaning full one to see if it can tell and insightful information. Next, applied another type of feature selection and pre-processed data such as another approach of dealing with missing value. Additionally, set another project goal on predicting room listing price, and implement a model for prediction.

Kaggle (2019) : New York City Airbnb open data https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data (downloaded : October 2019)

Airbnb (2019) : About Us https://news.airbnb.com/about-us/ (downloaded : November 2019)

Gibbs, C., Guttentag, D, Gretzel, U., Morton, J. & Goodwill, A. (2018) Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings, Journal of Travel & Tourism Marketing, 35:1, 46-56,
DOI: 10.1080/10548408.2017.1308292

ULFSSON, H. (2017) Predicting Airbnb user's desired travel destinations, *Predicting user Intentions*

Zhao, Y., Dong, S., Yang, J., (2015). Effect Research of Aspects Extraction for Chinese Hotel Review Based on Machine Learning Method, International Journal of Smart Home, 9, No. 3

Kalehbasti, P., Nikolenko, L., Rezaei, H. (2019). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis

Dudas, G., Vida, G., Kovalcsik, T., Boros, L. (2017). A socio-economic analysis of Airbnb in New York City, Regional Statistics, Vol., p. 135-151.
DOI : 10.15196/RS07108

Chen, B. (2019). Booking With Airbnb? Here's Your Survival Guide. Nytimes.com. Retrieved From https://www.nytimes.com/2018/06/06/technology/personaltech/booking-with-airbnb-heres-your-survival-guide.html

QUILAN, J.R. (1986) Induction of Decision Trees, Machine Learning 1, p. 81-106.

Breiman, L. (2001). Random Forest, Machine Learning, Vol. 45 Issue 1, p. 5-32.

Dietterich, T. (2000). Ensemble Methods in Machine Learning, MSC '00 Proceeding of the First International Workshop on Multiple Classifier Systems, p. 1-15