Prediction of Airbnb Data Prices and Deal Suggestions for the city of Amsterdam

Manjusha Roy Choudhury
Georgia State University

Abstract

This project deals with the prediction of prices for existing airbnbs based on various features. 'InsideAirbnb' dataset has been used to understand better the features that might result in better prediction of price. A number of factors like airbnb location, whether the host is a superhost or not, whether he or she has an image displayed, the type of property , rooms and bathrooms offered, different kinds of review ratings might help determine the cost of the property. In this paper we try to run two kinds of regression models-Linear and Random Forest to see the difference between the two in prediction of airbnb prices. Another small subpart of the project deals with classifying airbnb's into 'good', 'average', 'poor' deals based on various features mentioned above. We achieve this by performing PCA followed by unsupervised classification using K-means clustering.

1 Introduction

For most of our past, anyone who wanted to sell a good or a service had to earn some kind of certification to be able to legally do so and before people would trust them. Even 10 years ago, it would be super strange if someone was to just get into a stranger's car to go somewhere, or if someone was to randomly choose to live in a stranger's house in a new town. Companies like Uber and AirBnb have not only made this possible, but even normal and more favorable than the alternative. They have also achieved great success doing so - Airbnb has recently done an IPO.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

While removing the barrier of certification has created tons of new choices in the marketplace, it comes with its own sets of challenges and opportunities. As a person seeking to rent a space for a vacation, I want to be sure that the host is trustworthy and that I'm getting good value for my money. Meaning that I want to be sure that in a certain neighborhood, I'm paying a fair price for a property that has certain kinds of attributes and amenities. Conversely, as a host, I want to be sure that I'm getting a fair value for my property.

In this project I aim to tackle these two issues using the Airbnb dataset. Given a dataset of airbnb listings in Amsterdam and their various attributes, I demonstrate how machine-learning models can learn to predict the fair listing price of a property, given its various attributes as inputs. Where the prediction differs from the "true" listing price from the dataset, it might signal an opportunity for a host to raise their price if the difference is in their favor. If the difference is in the opposite direction, it could be used as a signal by airbnb to help the host set a more fair price for their property. For the renter's side, I aim to build a model that can guide the renter towards finding listings that could be deemed as "good deals", where perhaps the listings is being priced at a lower rate than other comparable listings in the vicinity.

2 Dataset analysis

This section in general is about all the preprocessing -including imputing missing data, deriving columns from raw data and getting insight from the data before using machine learning algorithms on it.

2.1 Data Preprocessing

a) For Regression Model: In order to be able to apply algorithms to understand prediction model for Airbnb prices, the 'InsideAirbnb' dataset for Amsterdam was preprocessed. This included dealing with missing values in columns and records, imputing them with mean or mode, coming up with derived features from the raw data that would be more useful in predicting the

target variable.

2.1.1 Dealing with Missing Values

There were around 74 features and 18906 records in the dataset to begin with. Preprocessing was performed to make it more usable for regression model. Features(columns) with more than 1/3 rd values i.e. more than 9000 values missing were dropped- 'host response time', 'host response rate', 'host acceptance rate', 'host neighbourhood'.Columns which had no values at all i.e. empty columns were also dropped-'neighbourhood group cleansed', 'bathrooms', 'calendar updated, license', 'neighbourhood'. Review scores('review scores rating') are an important factor in prediction of price of a listing. Since 2530 records did not contain 'review score ratings' and imputing these values with mode or mean could result in erroneous predictions, these records were dropped. Other kinds of missing review values such as 'review scores accuracy', 'review scores cleanliness','review scores checkin', 'review scores communication','review scores location', 'review scores value' were imputed with the mean value. Certain other missing entries in columns such as 'host since', 'host is superhost', 'host location', 'host listing count', 'host total listings count', 'host identity verified', 'bedrooms', 'beds', 'bathrooms text' were imputed with the mode value.

2.1.2 Derived columns from raw columns

To make the data in some of these columns more meaningful, various kinds of calculations were performed to obtain derived numeric data columns that made more sense in prediction of the target variable. Location data is very important in determining the price of the property. For this reason, the latitude and longitude from each of the record was used to calculate the 'haversine distance' from Berlin. Some of the other modifications or derivations included converting the date columns such as 'first review', 'last review', 'calendar last scraped', 'host since' to numeric values by finding the number of days by simply calculating the difference between these dates and the present date. Another important factor in determination of price could be the fact whether the host resides in Amsterdam or not. To take this into consideration, 'host locations' was label encoded to 0s and 1s depending on whether the host lives outside Amsterdam or is a local. The fact whether a customer chooses to stay in an Airbnb or not also depends on whether the host has pictures and descriptions etc. For this reason most of the urls and descriptions ('description','neighborhood overview' , 'host about', 'host name', 'host has profile pic', 'host thumbnail url', 'host picture url') were converted to binary encoder based on if the host had urls and descriptions present or not. In some cases if all urls or descriptions were present for certain columns, they were not used towards our prediction. Significant 'yes' and 'no' features such as 'host is superhost', 'host identity verified', 'has availability', 'host has profile pic', 'instant bookable' were converted to 0s and 1s using label encoder. Three important features - 'property type', 'room type', 'bathroom text' which contained more than 2 categories were not converted just to numbers for the purpose of regression because this would result in bias. Backward difference encoding was used to deal with these nominal categories. 'Backward difference encoding' is used for a feature having 'K' categories that generally enter into a regression model with a sequence of K-1 dummy variables. The mean of the dependent variable for a level is compared to that of the prior level. This type of encoding is generally used in case of nominal or ordinal variable. Thus after this particular modification , a lot more columns (132) were created than we started with. [1] b) For the clustering model: Preprocessed data from above was used.

2.2 Data Exploration

a) For regression model: For data exploration and understanding the data better, we went ahead and used correlation heat map to understand if there were any existing correlation between features and to substitute the column with high correlation (less than 0.75) with just one column instead. Specifically for the Random Forest Regression, feature importance were plotted to understand the relevance of the different features (discussed in details in the model selection section.

b) For Unsupervised k-means classification model: For this model different significant features were tested from the feature set that would result in successful clustering. Finally the following features were shortlisted for k-means clustering based on the importance of these features in classifying a listing-'host since derived', 'host location derived', 'host is superhost', 'host listings count, host has profile pic, host identity verified','distancefromberlin', 'accommodates', 'bedrooms', 'beds', 'minimum nights','has availability', 'calendar last scraped derived', 'number of reviews', 'review scores rating', 'review scores accuracy','review scores cleanliness', 'reviews per month', 'price'.

3 Model Selection and Discussion

a) For Regression Model: Using Linear Regression The entire feature set of 132 columns were used. (Remember that the number of columns were increased to a great extend because of using back difference encoders for 'room type', 'property type' and 'bathroom text').

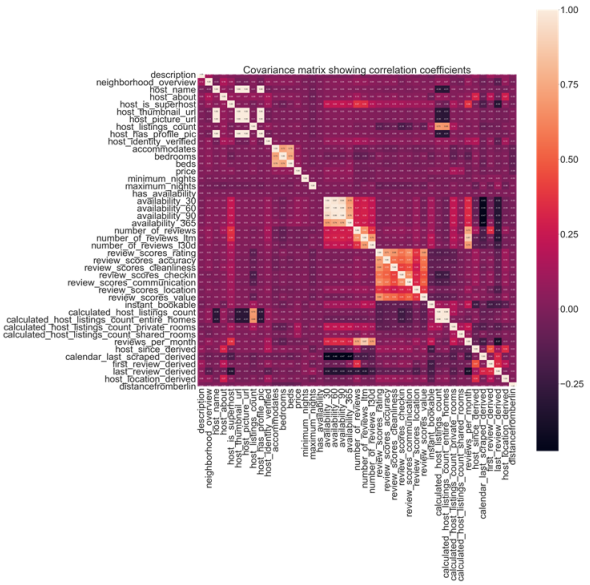

Figure 1: Correlation Plot

Though the model gave a comparatively lower MAE value of 106.53, MSE value of 106.53 and RMSE value of 10.32 for the test data, the adjusted r2 value for the model was found to be only 54 percent (Figure 4). This was actually not shocking given the complex nature of the dataset. Actual vs predicted test values as well as residuals were plotted to better evaluate the accuracy of the model.(Figure 2 and 3)

Figure 2: Actual vs Predicted for Linear Regression Model

For Regression Model: Using Random Forest Regression Initially the same set of data used in the linear

Figure 3: Plot of Residuals for Linear Regression Model

OLS Regression Results

Dep. Variable:	y	R-squared:	0.543
Model:	OLS	Adj. R-squared:	0.540
Method:	Least Squares	F-statistic:	153.1
Date:	Sat, 12 Dec 2020	Prob (F-statistic):	0.00
Time:	05:49:59	Log-Likelihood:	-1.1917e+05
No. Observations:	16351	AIC:	2.386e+05
Df Residuals:	16224	BIC:	2.396e+05
Df Model:	126		
Covariance Type:	nonrobust		

Figure 4: OLS for Linear Regression Model

regression model was used in the random forest model to see if there was a possibility to improve r2 as well as RMSE values. Though the data fit the model better with r2 value of 96 percent for the training set, the r2 value for the testing set was found to be 70 percent. The accuracy of the model however decreased when compared to the linear regression model with MAE value of 56.98, MSE value of 73556.46 and RMSE value of 271.2. Actual vs predicted test values as well as residuals were plotted to better evaluate the accuracy of the model.(Figure 5 and 6). Feature Importance were plotted for all of the RF models(Figure 7, not all are shown). Some of the important features for this model were -'property type', 'host listing count', 'accommodates','distancefromberlin', 'first review dervied, calendar last scraped', 'review score location', 'review score cleanliness', 'review score accuracy', 'number of reviews ltm' among. Feature subset selection was performed from the above feature set and the model ran on it but none of them could improve on the accuracy or r2 values that we got from the model (not shown in this paper).

Next, we tried to use a feature set containing more independent variables and got back some of the columns that we had rejected before such as 'host name, 'minimum minimum nights', 'maximum minimum nights','minimum maximum nights','maximum

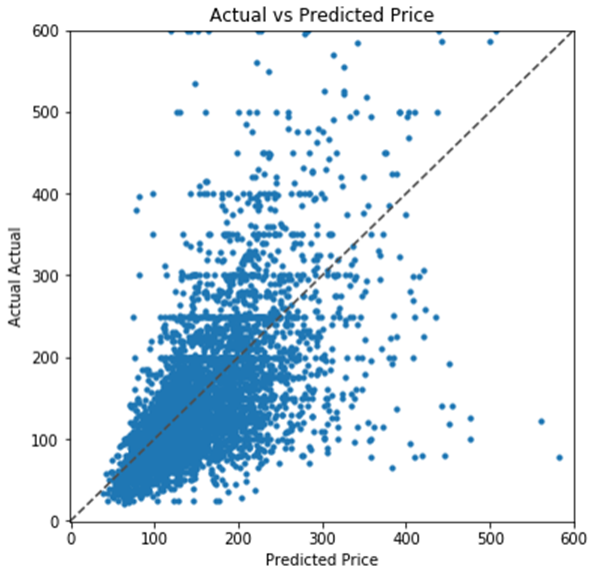

Figure 5: Actual vs Predicted for Random Forest Model-1

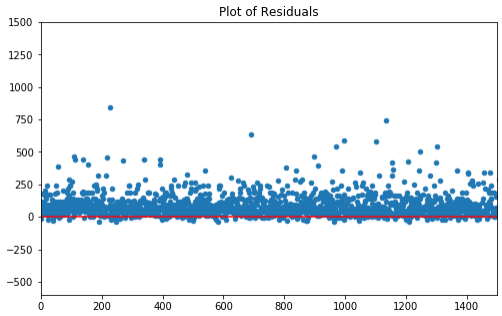

Figure 6: Plot of Residuals for Random Forest Model-1l

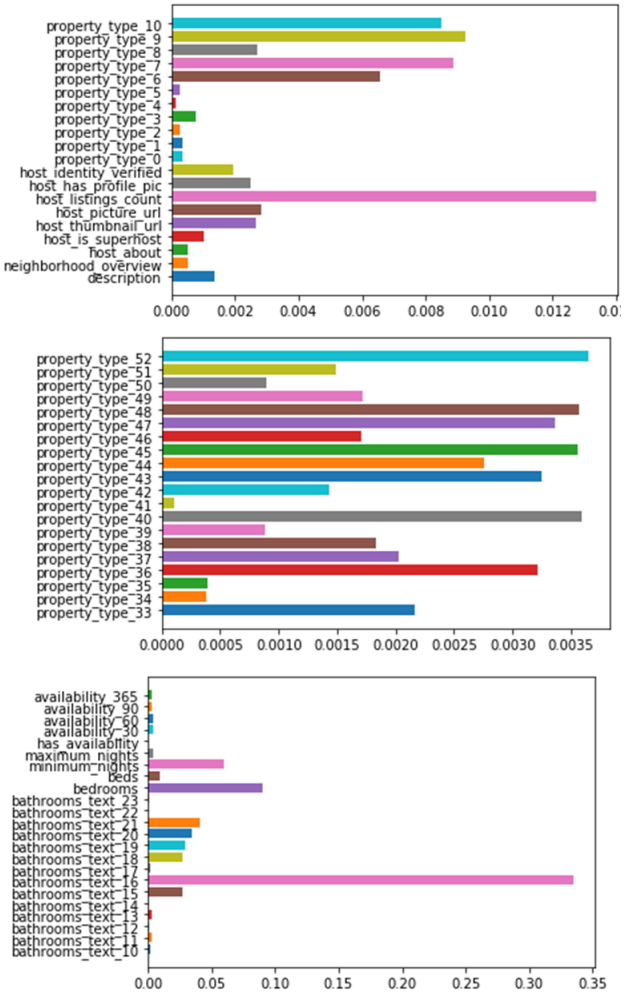

Figure 7: Feature Importance for RF Model-1:Does not show all features

maximum nights'.This feature set now contained 137 features. This was one of the better performing models. The data fit the model well with r2 value of 97 percent for the training set, while the r2 value for the testing set was found to be 87 percent. There was a substantial improvement in the r2 for the test set.The accuracy of the model however did not improve substantially when compared to the RF- model-1 :MAE value of 54.31, MSE value of 32029.61 and RMSE value of 178.96. Actual vs predicted test values as well as residuals were plotted to better evaluate the accuracy of the model.(Figure 8 and 9).The beautiful fitting of the training data was also plotted (Figure 10)

b) For studying clustering of the data into 'good', 'average' and 'poor' deals, principal component analysis was performed on the data to reduce the dimensionality of the data. A derived feature called 'Deal Type' which was basically 'price' divided by 'review score rating' was computed as a measure of classification for the different listings and the PCA plot was color coded accordingly. A record having a 'deal' score of less than 1.5 was classified as 'good deal', a value between 1.5 and 4.5 was classified as an 'average deal' while all other values higher than 4.5 were classified as 'poor deal'. No clear cut segregation was observed in the PCA plot(Figure 11). Though initially, this lead us to belief that clustering with k-means could be difficult for this dataset, an important fact that was being ignored is that a deal depends on much more than just the price and review ratings and thus we included the significant features mentioned in the data exploration section and went ahead and carried out k-means clustering. K-means clustering could successfully divide

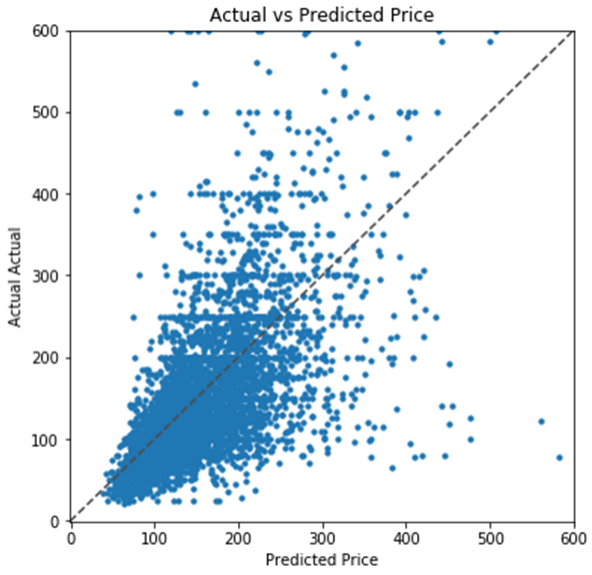

Figure 8: Actual vs Predicted for Random Forest Model-2

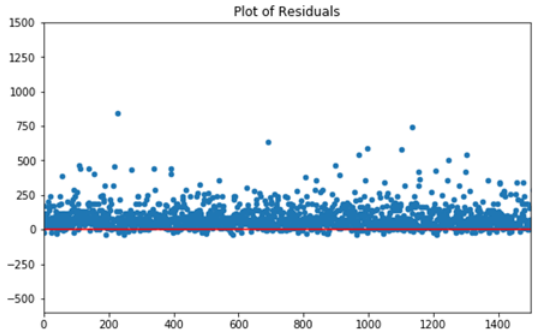

Figure 9: Plot of Residuals for Random Forest Model-2l

the listings into 3 classes. To further label these classes we went through the actual datasets and realised that the clustering with k-means was more or less accurate if we take into consideration various factors like whether the host is a super host or not, whether the host is verified, the price of the property, the different kinds of ratings, number of reviews etc. Thus the k-means plot could be color coded based on the table (Figure 12) where k-mean classification of '0' meant a 'good deal', '1' was an 'average deal' and '2' belonged to 'poor deal'.(Figure 13)

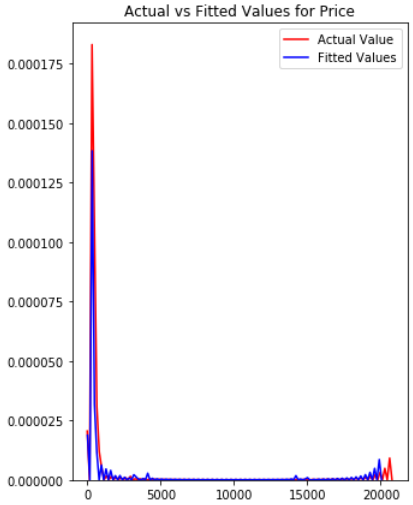

Figure 10: Plot o Actual vs Fitted Values for pricel

4 Visualization

a) With the regression model there is less chance of visualization-it would just be helpful in predicting or advising on price for an existing listing. b) Visualization for the model would require further work. The 'good'. 'average' and 'poor' deal could be plotted on the actual geographic map of Amsterdam. Plotting these listings using latitude and longitude data and then classifying them using the k-means classification could be a very useful visualization for a customer who is deciding to a book an airbnb.

5 Comparison Tables

a) A table comparing the accuracy and prediction values for Linear Regression Mode, Random FOrest Model-1 and Random Forest Model-2 are shown below.

Table 1: Regression Model Comparison

Model	r2 train	r2 test	MAE	MSE	RMSE
Linear	54	–	106.53	106.53	10.32
RF-1	96	70	56.98	73556.46	271.21
RF-2	97	87	56.9	32039.61	178.96

Conclusion

In conclusion, we were able to model the 'InsideAirbnb' dataset for prediction of price using Linear and Random Forest Regression (Table 1). Best accuracy was achieved by RF-2: r2 for train was 97 per-

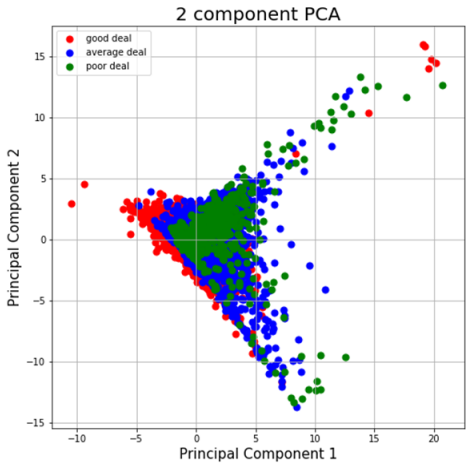

Figure 11: PCA plot classified using Deal score(Price/Review Ratings)

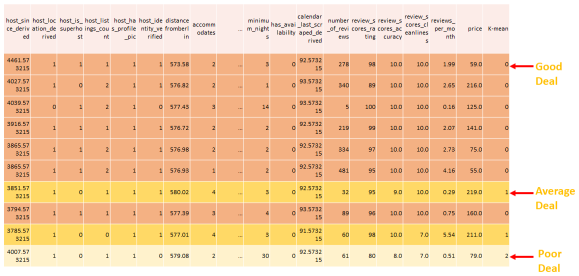

Figure 12: Table determining the classification and labelling of k-means plot

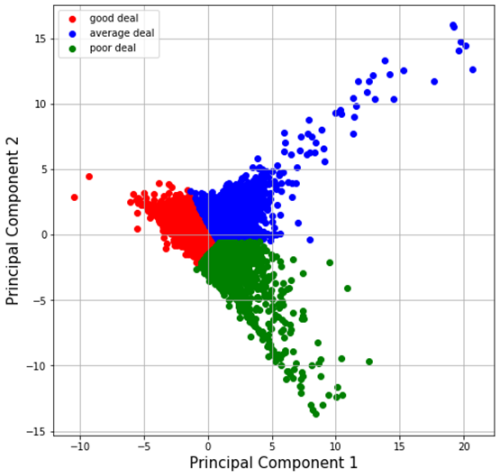

Figure 13: k-means clustering plots with labels

cent and that for test was found to be 87 percent. We were also able to classify the existing listings into 'good', 'average' and 'poor' deals using PCA and k-means clustering.

only in the camera-ready papers.

Acknowledgement

I would like to thank Dr. Jaya Krishna Mandivarupu for his valuable advice during class and especially during the presentation session. I have tried to include the changes that he advised during the presentation session.

References

[1] Backward differe encoding. *https://www.kdnuggets.com/2015/12/beyond-one-hot-exploration-categorical-variables.html*.