**PROJECT REPORT**

**AirBnB Price Prediction**

**“submitted towards partial fulfilment of the criteria for award of PGPDSE by GLIM”**

Submitted by:

**R.Aanand**

**M.Adithya**

**S.U.Ajeet Adeteea**

**R.Pragatheeswaran**

**R.Sidharthen**

**Batch: DSE\_CHENNAI\_JULY\_2019**

**Mentor: Mr. Vikas Chandra**





**Abstract & keywords**

Airbnb, an online platform for peer-to-peer short-term accommodation rental, is growing with a spectacular speed in most countries of the world. This creates a great regulatory challenge, as empirical evidence suggest that Airbnb may have a significant impact on the traditional hotel industry and on the housing market. The aim of this analysis is to find the price of the new hosts based on their listings of Airbnb to traditional services and to examine the standard between the different business models. The empirical analysis is based on a unique dataset of scraped data on the listings of Airbnb in City of Barcelona, Spain. The offers are compared using descriptive statistics and KDE analysis. The paper supports that Airbnb is providing a cheaper alternative for hospitality services in all price segments. Controlling for common attributes (e.g. distance from the city centre).

**Keywords**: Price Prediction, Machine Learning, Imputation and Regularization

**Acknowledgement**

At the outset, we are indebted to our **Mentor Mr. Vikas Chandra** for his time, valuable inputs and guidance. His experience, support and structured thought process guided us to be on the right track towards completion of this project.

We are extremely gifted and fortunate to have **Ms. Anne Grace** as our General Manager – Academic Delivery. Her in-depth knowledge coupled with her passion in delivering the subjects in a lucid manner has helped us a lot. We are thankful to her for her guidance towards entire coursework.

We are thankful **to Ms. Meena Vardhini**, Jr.Data Scientist, Acad-Ops , DSE Program for her unflinching and unabated help extended to us always.

We also thank all the course faculty of the DSE program for providing us a strong foundation in various concepts of analytics & machine learning.

Last but not the least, we would like to sincerely thank our respective families for giving us the necessary support, space and time to complete this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

**R.Aanand**

**M.Adithya**

**S.U.Ajeet Adeteea**

**R.Pragatheeswaran**

**R.Sidharthen**

Date**: 20th Nov 2019**

Place: **Chennai**

**Certification of completion**

I hereby certify that the project titled “**Airbnb Price Prediction**” was undertaken and completed under my guidance by R.Aanand, M.Adithya,S.U.Ajeet Adeteea, R.Pragatheeswaran, R.Sidharthen, students of the July 2019 batch of the Post Graduate Program in Data Science & Engineering, Chennai.

Mr**. Vikas Chandra**

Date: **20th Nov 2019**

**Table of contents**

**Executive summary:**

**Background:**

**Airbnb, Inc.**is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking.] The company is based in San Francisco, California, United States.

Shortly after moving to San Francisco in October 2007, roommates and former schoolmates Brian Chesky and Joe Gebbia could not afford the rent for their loft apartment. Chesky and Gebbia came up with the idea of putting an air mattress in their living room and turning it into a bed and breakfast. The goal at first was just "to make a few bucks".In February 2008, Nathan Blecharczyk, Chesky's former roommate, joined as the Chief Technology Officer and the third co-founder of the new venture, which they named Air Bed & Breakfast.

 They put together a website which offered short-term living quarters, breakfast, and a unique business networking opportunity for those who were unable to book a hotel in the saturated market. The site *Airbedandbreakfast.com* officially launched on August 11, 2008. The founders had their first customers in town in the summer of 2008, during the Industrial Design Conference held by Industrial Designers Society of America, where travelers had a hard time finding lodging in the city.

**Objective:**

The objective of this project is to do a research and build models for Airbnb Price Prediction. The scope of the project is that the host might not know how much they should price their listing and this affects the company's revenue. By analyzing and considering the aspects of the listing types, location, reviews, price near neighbourhood and the amenities in their listings, we guide the new hosts to set an optimal pricing for their listings. There are some cases where the hosts are new to the locality and don’t know the optimum price of the listing and end up setting up some random prices. Hence we alert the hosts to change the price accordingly.

**Methodology:**

The data used in this post has been downloaded from the http://insideairbnb.com/ website and was collected in September 2019. After preprocessing the dataset, we use various regression algorithms to predict the price of the listing using features like location, host details, beds, bedrooms and bathrooms and much more. The models are evaluated using relevant model performance measures to arrive at the most robust models for prediction.

**Chapter 1 – Project Overview**

**Problem statement:**

Airbnb wants its hosts to set their own prices. Most of the hosts don’t know how to fix a price for their listing and fix some random price which would make the customers to churn. So we use data science and use the features to our benefit and predict the optimum price for our listing.

**Problems faced by Airbnb:**

1. From the suppliers side, long term rentals are converted to short term rentals.  
2. From the demand side, people know how much an AirBnB costs outside so they tend to accept higher rents.  
3. Anti tourist march.

4. Two different pricing for listings in the same locality with very less or no distinguishing difference.

**Data Source:**

The data used in this post has been downloaded from the http://insideairbnb.com/ website and was collected in September 2019.

**Dataset Description:**

|  |  |
| --- | --- |
| host\_response\_time | Time taken by host to respond to a particular query |
|  |  |
| host\_response\_rate | Total no.of responses/Total no. of queries |
|  |  |
| host\_is\_superhost | Superhost status of a host(Yes/No) |
|  |  |
| host\_listings\_count | Total no. of listings provided by a particular host |
|  |  |
| host\_verifications | No of id’s needed to verify the host/customer |
|  |  |
| host\_has\_profile\_pic | Whether or not the host has a profile picture |
|  |  |
| host\_identity\_verified | If the host is a verified person or not |
|  |  |
| is\_location\_exact | If the actual and mentioned location go hand in hand |
|  |  |
| property\_type | Type of property(Apartment/House/Boat/Villa....etc) |
|  |  |
| room\_type | Type of room(shared/privat4e/hotel...) |
|  |  |
| accommodates | No. of accomodates available |
|  |  |
| bathrooms | No. of bathrooms available |
|  |  |
| bedrooms | No. of bedrooms available |
|  |  |
| beds - No. of beds available | No. of beds available |
|  |  |
| bed\_type | Type of bed(Real/Air/Sofa) |
|  |  |
| amenities | A list of facilities available |
|  |  |
| price | Price for stay(target) |
|  |  |
| security\_deposit | part payment for a particular stay(with the promise to fully pay the money after the stay) |
|  |  |
| cleaning\_fee | amount charged for maintaining the place |
|  |  |
| guests\_included | Extra people who might or will join the intial group |
|  |  |
| extra\_people | Charge per extra person |
|  |  |
| minimum\_nights | minimum nights stayed by a guest |
|  |  |
| maximum\_nights | maximum nights stayed by a guest |
|  |  |
| minimum\_minimum\_nights |  |
|  |  |
| maximum\_minimum\_nights |  |
|  |  |
| minimum\_maximum\_nights |  |
|  |  |
| maximum\_maximum\_nights |  |
|  |  |
| minimum\_nights\_avg\_ntm |  |
|  |  |
| maximum\_nights\_avg\_ntm |  |
|  |  |
| availability\_30 | availability for the next 30 days |
|  |  |
| availability\_60 | availability for the next 60 days |
|  |  |
| availability\_90 | availability for the next 90 days |
|  |  |
| availability\_365 | availability for the next 365 days |
|  |  |
| number\_of\_reviews | no. of reviews |
|  |  |
| number\_of\_reviews\_ltm | no. of reviews in last twelve months |
|  |  |
| review\_scores\_rating | The ratings given by the customer |
|  |  |
| review\_scores\_accuracy | review score for accuracy of details provided by the host |
|  |  |
| review\_scores\_cleanliness | review score for cleanliness |
|  |  |
| review\_scores\_checkin | review score for checkin experience |
|  |  |
| review\_scores\_communication | review score for the way in which the host communicated |
|  |  |
| review\_scores\_location | review score for the accuracy of location |
|  |  |
| review\_scores\_value | Value for money/Pricing review rating |
|  |  |
| instant\_bookable | Whether the place is instantly bookable or not |
|  |  |
| cancellation\_policy | Flexibility of cancellation(strict\_14\_with\_grace\_period,moderate,flexible,super\_strict) |
|  |  |
| require\_guest\_profile\_picture | Whether the guest needs a profile picture in his profile |
|  |  |
| require\_guest\_phone\_verification | Whether the guest mobile no. is verified by AirBnB |
|  |  |
| calculated\_host\_listings\_count | Host\_listings(AirBnB) |
|  |  |
| calculated\_host\_listings\_count\_entire\_homes | Total no. of entire home listings provided by a particular host |
|  |  |
| calculated\_host\_listings\_count\_private\_rooms | Total no. of private room listings provided by a particular host |
|  |  |
| calculated\_host\_listings\_count\_shared\_rooms | Total no. of shared room listings provided by a particular host |
|  |  |
| reviews\_per\_month | No.of reviews per month |

**Percentage of missing values:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No |  |  | Features | Percent of Missing values | Number of Missing values |
| 0 |  |  | host\_response\_time | 12.39953 | 2530 |
| 1 |  |  | host\_response\_rate | 12.39953 | 2530 |
| 2 |  |  | host\_is\_superhost | 0.107822 | 22 |
| 3 |  |  | host\_listings\_count | 0.107822 | 22 |
| 4 |  |  | host\_total\_listings\_count | 0.107822 | 22 |
| 5 |  |  | host\_verifications | 0 | 0 |
| 6 |  |  | host\_has\_profile\_pic | 0.107822 | 22 |
| 7 |  |  | host\_identity\_verified | 0.107822 | 22 |
| 8 |  |  | is\_location\_exact | 0 | 0 |
| 9 |  |  | property\_type | 0 | 0 |
| 10 |  |  | room\_type | 0 | 0 |
| 11 |  |  | accommodates | 0 | 0 |
| 12 |  |  | bathrooms | 0.034307 | 7 |
| 13 |  |  | bedrooms | 0.039208 | 8 |
| 14 |  |  | beds | 0.161733 | 33 |
| 15 |  |  | bed\_type | 0 | 0 |
| 16 |  |  | amenities | 0 | 0 |
| 17 |  |  | price | 0 | 0 |
| 18 |  |  | security\_deposit | 29.513821 | 6022 |
| 19 |  |  | cleaning\_fee | 20.324446 | 4147 |
| 20 |  |  | guests\_included | 0 | 0 |
| 21 |  |  | extra\_people | 0 | 0 |
| 22 |  |  | minimum\_nights | 0 | 0 |
| 23 |  |  | maximum\_nights | 0 | 0 |
| 24 |  |  | minimum\_minimum\_nights | 0 | 0 |
| 25 |  |  | maximum\_minimum\_nights | 0 | 0 |
| 26 |  |  | minimum\_maximum\_nights | 0 | 0 |
| 27 |  |  | maximum\_maximum\_nights | 0 | 0 |
| 28 |  |  | minimum\_nights\_avg\_ntm | 0 | 0 |
| 29 |  |  | maximum\_nights\_avg\_ntm | 0 | 0 |
| 30 |  |  | has\_availability | 0 | 0 |
| 31 |  |  | availability\_30 | 0 | 0 |
| 32 |  |  | availability\_60 | 0 | 0 |
| 33 |  |  | availability\_90 | 0 | 0 |
| 34 |  |  | availability\_365 | 0 | 0 |
| 35 |  |  | number\_of\_reviews | 0 | 0 |
| 36 |  |  | number\_of\_reviews\_ltm | 0 | 0 |
| 37 |  |  | review\_scores\_rating | 22.294648 | 4549 |
| 38 |  |  | review\_scores\_accuracy | 22.338757 | 4558 |
| 39 |  |  | review\_scores\_cleanliness | 22.338757 | 4558 |
| 40 |  |  | review\_scores\_checkin | 22.368163 | 4564 |
| 41 |  |  | review\_scores\_communication | 22.319153 | 4554 |
| 42 |  |  | review\_scores\_location | 22.358361 | 4562 |
| 43 |  |  | review\_scores\_value | 22.35346 | 4561 |
| 44 |  |  | requires\_license | 0 | 0 |
| 45 |  |  | instant\_bookable | 0 | 0 |
| 46 |  |  | is\_business\_travel\_ready | 0 | 0 |
| 47 |  |  | cancellation\_policy | 0 | 0 |
| 48 |  |  | require\_guest\_profile\_picture | 0 | 0 |
| 49 |  |  | require\_guest\_phone\_verification | 0 | 0 |
| 50 |  |  | calculated\_host\_listings\_count | 0 | 0 |
| 51 |  |  | calculated\_host\_listings\_count\_entire\_homes | 0 | 0 |
| 52 |  |  | calculated\_host\_listings\_count\_private\_rooms | 0 | 0 |
| 53 |  |  | calculated\_host\_listings\_count\_shared\_rooms | 0 | 0 |
| 54 |  |  | reviews\_per\_month | 21.206626 | 4327 |
| 55 |  |  | listing\_id | 0 | 0 |

**Data Preprocessing:**

**Label Encoding:**

We have label encoded the categorical data. The columns that are label encoded are host response time, property type, room type, bed type, cancellation policy, bathrooms, bed, bedroom.

**Missing values:**

Iterative Imputation: Multi variate imputer

A strategy for imputing missing values by modeling each feature with

missing values as a function of other features in a round-robin fashion.

The order in which the features will be imputed can be determined by the user either by features with fewest missing values to most and vice-versa.

The various estimators used to impute values are:

Linear Regression

Ridge Regression

KNeighborsRegressor

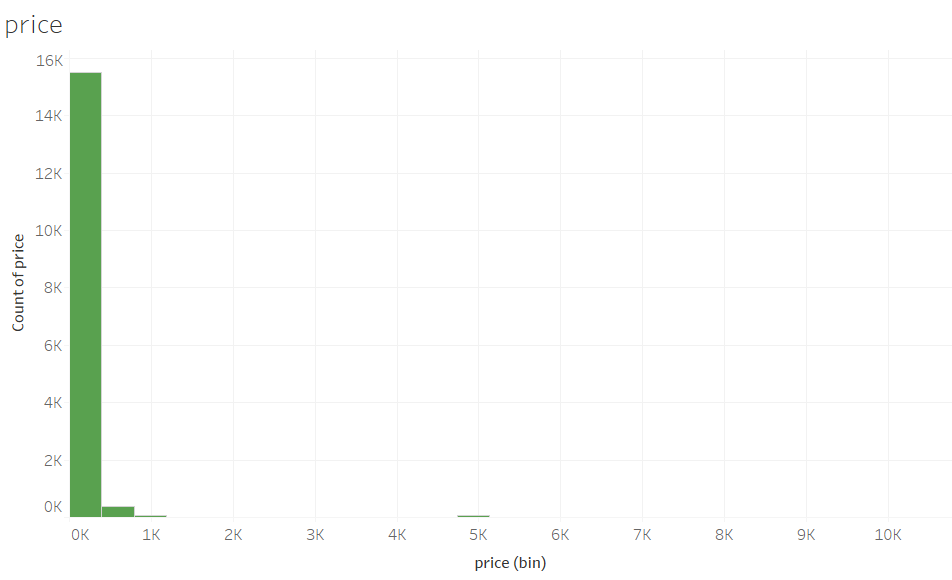
DecisionTreeRegressor

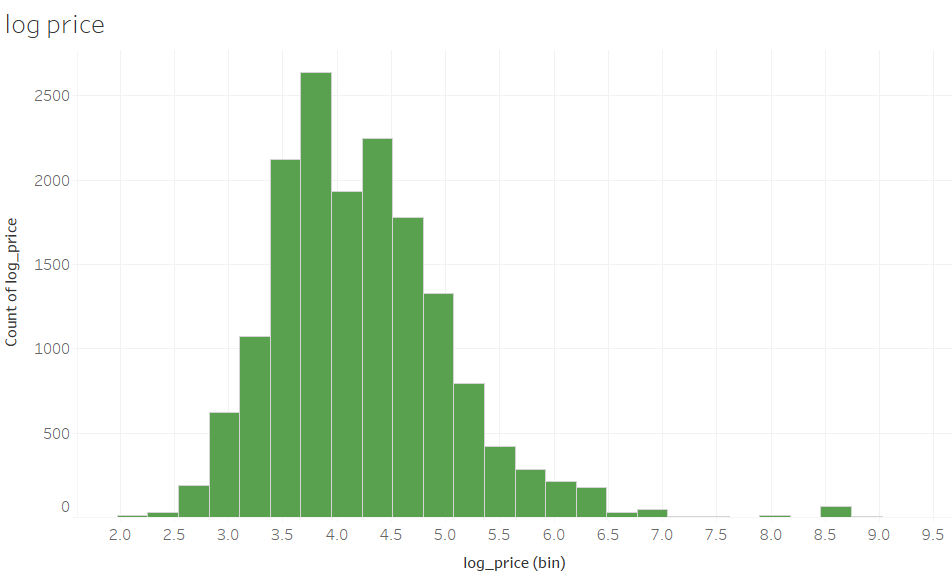
**Outlier Treatment:**

After analysing the boxplot of the target column(price) we found that there were few outliers.

We came to know that there were 1640 outliers but then failed to remove it as the data were real and it is possible to have outliers in price column.

**Transformation:**

****



The above graph shows distribution of target feature before and after transformation. The transformation that we have performed is log transformation.

**Statistical tools & techniques:**

The Regression algorithms used for predicting the Price include

* Linear regression
* Random Forest Regressor
* DecisionTree Regressor
* Ada Boosting
* XG Boost

**Linear Regression:**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task to compute the regression coefficients. Regression models a target prediction based on independent variables. Linear Regression performs the task to predict a dependent variable value (y) based on a given independent variable (x).

After building the base model, we check whether our data satisfies the assumptions.

**1) No autocorrelation:**

The presence of correlation in error terms drastically reduces model’s accuracy. This usually occurs in time series models where the next instant is dependent on previous instant. Test needed : Durbin- Watson Test. Its value ranges from 0-4.

**2) No Multicollinearity :**

Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables. Pair plots and heatmaps (correlation matrix) can be used for identifying highly correlated features.

Test needed : VIF Test

**3) Homoscedasticity:**

Homoscedasticity describes a situation in which the error term is the same across all values of the independent variables.

Test needed :Breusch-Pagan and Goldfeld-Quandt test

**4) Normality of Residuals:**

If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Presence of non – normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.

Test needed :JarqueBera test. The Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution.

**5) Linearity of Residuals:**

If you fit a linear model to a non-linear, non-additive data set, the regression algorithm would fail to capture the trend mathematically, thus resulting in an inefficient model.

Test needed: Rainbow test.

**Random Forest Regressor:**

The random forest approach is a bagging method where deep trees, fitted on bootstrap samples, are combined to produce an output with lower variance.

Sampling over features has indeed the effect that all trees do not look at the exact same information to make their decisions. Another advantage of sampling over the features is that it makes the decision-making process more robust to missing data.

**Decision Tree Regressor:**

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decisionnodes and leafnodes.

**Hyperparameter tuning:**

We had some good results with the default hyperparameters of the Random Forest regressor. But we can improve the results with some hyperparameter tuning. There are methods available for this: Grid search You have to provide a parameter grid to these methods. Then, they both try different combinations of parameters within the grid you provided. The Grid search tries all the possible combinations with the grid you provided.

**Adaptative boosting:**

Adaptative boosting (often called “adaboost”), is a boosting ensemble model and works especially well with the decision tree. Boosting model’s key is learning from the previous mistakes, e.g. misclassification data points. AdaBoost learns from the mistakes by increasing the weight of misclassified data points.Inada boost, we try to define our ensemble model as a weighted sum of L weak learners.

**XGBOOST:**

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.

**Cross-validation:**

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

**K - FOLD:**

K-Fold CV is where a given data set is split into a **K** number of sections/folds where each fold is used as a testing set at some point. Lets take the scenario of 5-Fold cross validation(K=5). Here, the data set is split into 5 folds. In the first iteration, the first fold is used to test the model and the rest are used to train the model. In the second iteration, 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated until each fold of the 5 folds have been used as the testing set.

**R-squared**

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R-squared is the percentage of the response variable variation that is explained by a linear model.

**Adjusted R-squared**

Adjusted R-squared penalizes you for adding variables which do not improve your existing model. It is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative.

**F static:**

The F-statistic is simply a ratio of two variances. Variances are a measure of dispersion, or how far the data are scattered from the mean. Larger values represent greater dispersion.

**Probability of F statistic:**

In statistics, the p-value is the probability of obtaining the observed results of a test, assuming that the null hypothesis is correct. It is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

**Sentiment Analysis:**

Sentiment analysis is the automated process of analyzing text data and classifying opinions as negative, positive or neutral. Usually, besides identifying the opinion, these systems extract attributes of the expression e.g.:  
Polarity: If the speaker express a *positive* or *negative* opinion,  
Subject: The thing that is being talked about,  
Opinion Holder: The person, or entity that expresses the opinion.

**Voting Regressor:**

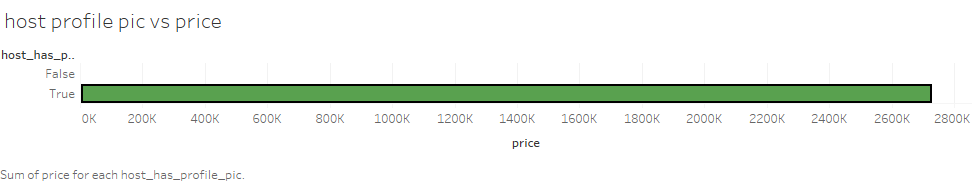
A voting regressor is an ensemble meta-estimator that fits base regressors each on the whole dataset. It, then, averages the individual predictions to form a final prediction.

**CHAPTER 2 – EXPLORATORY DATA ANALYSIS**

The purpose of Exploratory data analysis is to understand the data and get insights of various features.

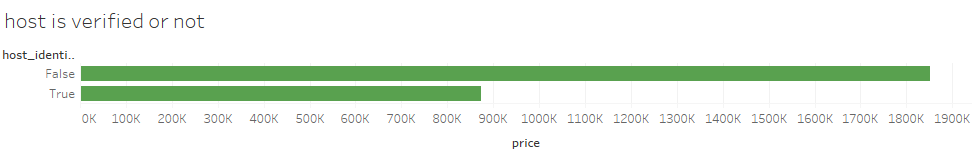
Let us check how much revenue is generated from each feature.

**Host profile picture vs. price:**



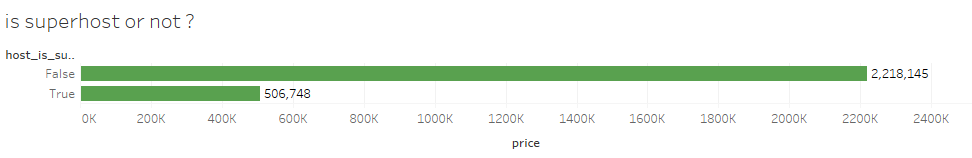
We can see that people prefer hosts with a profile picture than without a profile picture making the whole transaction transparent.

**Host verified is or not:**



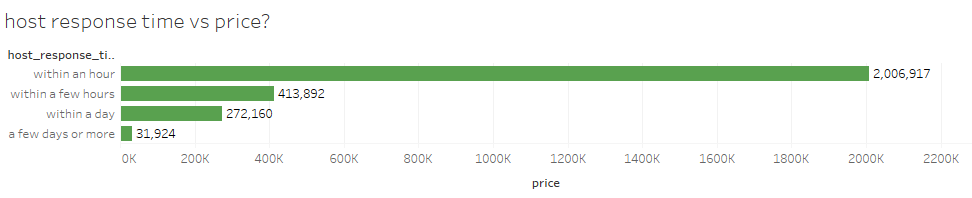
Most of the hosts aren’t verified, which hints that there might be some illegal listings and the company must work on it to verify its hosts.

**Super host or not:**



Most of the revenue generated is from non super hosts.

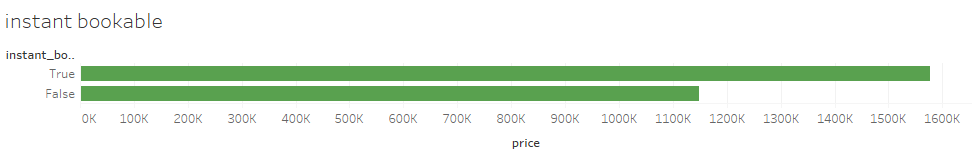
**Host response time vs. price:**



Most of the revenue generated when the hosts respond quickly. We can see that there is a significant difference between hosts responding quickly and taking time to respond.

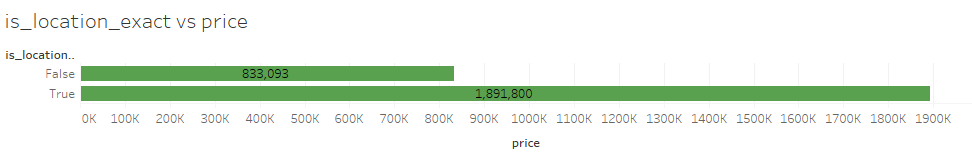
Hence we can suggest the hosts to respond as soon as possible to increase the revenue.

**Instant bookable:**



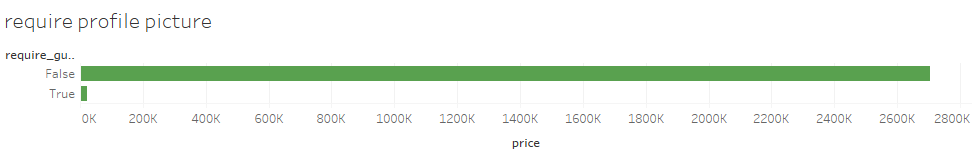
Customers bring in more revenue when they have an option of instant bookable.

**Location exact vs. price:**



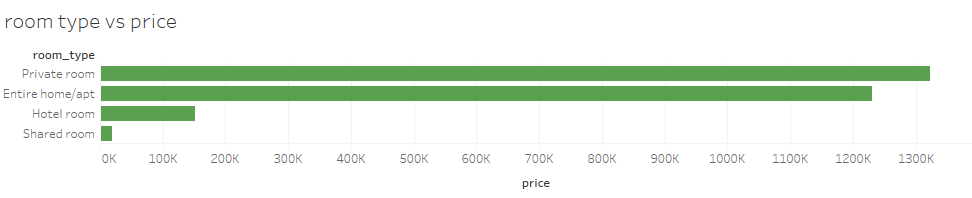
Customers bring in more revenue when the location of the listing is exact.

**Profile picture verification:**



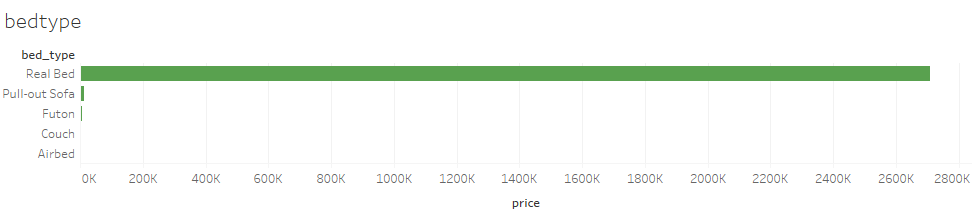
Customers bring in more revenue when the hosts don’t ask for a profile picture, which shows that the customers care for their privacy.

**Room type vs. Price:**



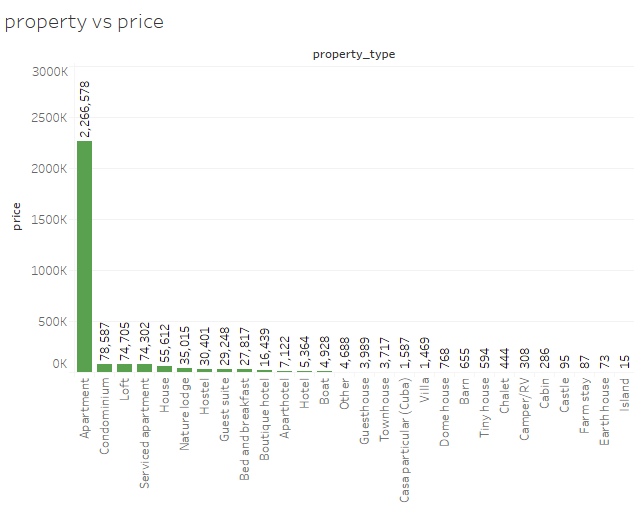
Private rooms bring more revenue when compared to other room types. Hence based on this we can increase the number of listings with private rooms.

**Bed type vs. price:**



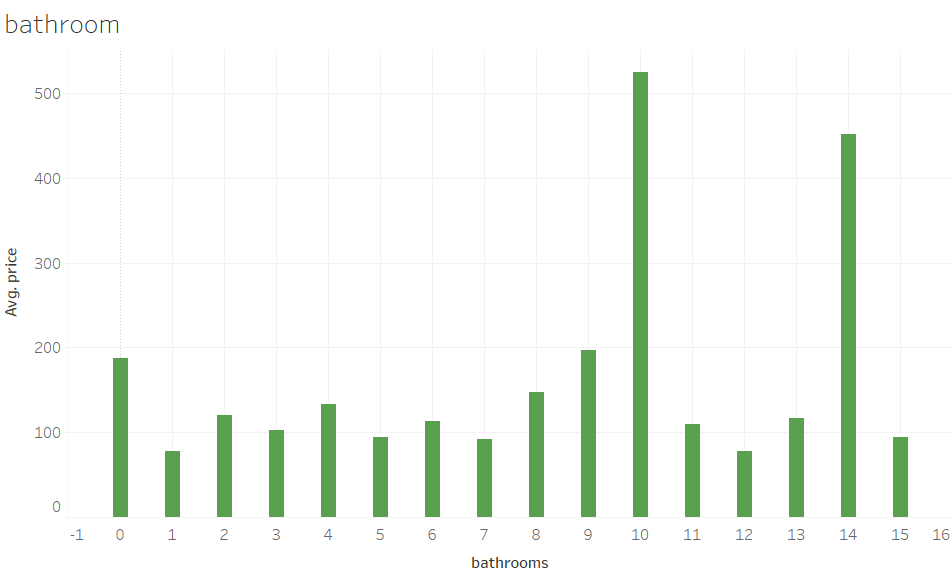
Real bed brings in more revenue when compared to other types of bed.

**Property vs. Price:**



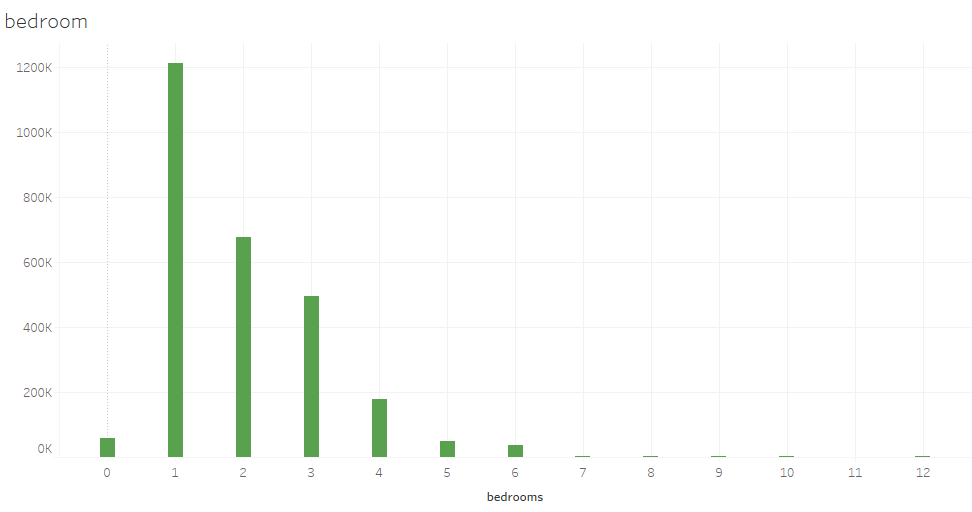
Apartments bring in more revenue when compared to other type of properties. Hence we can promote more apartments.

**Bathroom vs. Price:**



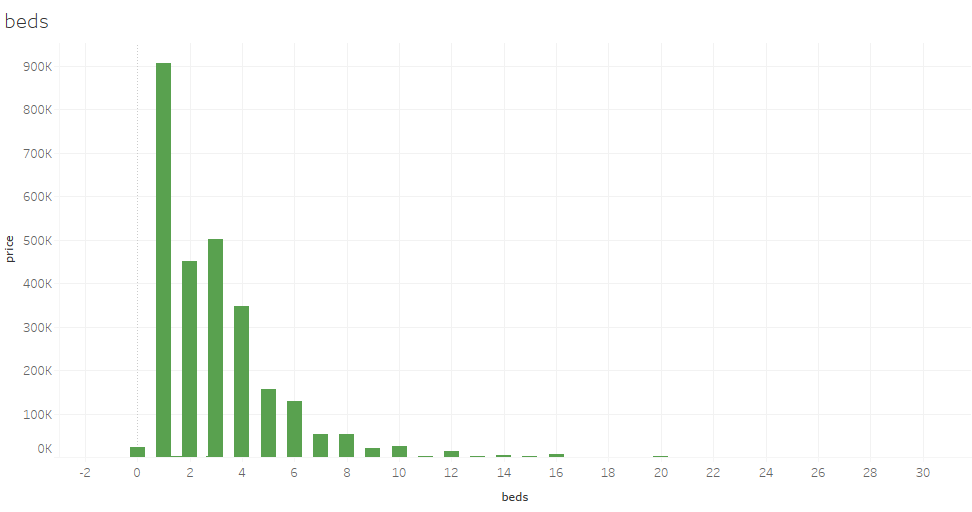
As we can see that the average price is high for 10 bathrooms

**Bedroom vs. Price:**



Less number of bed rooms bring more revenue.

**Beds vs. Price:**



From the above graph we can infer that the lesser number of bed bring more revenue.

Chapter 3 - Feature Selection & Model Building

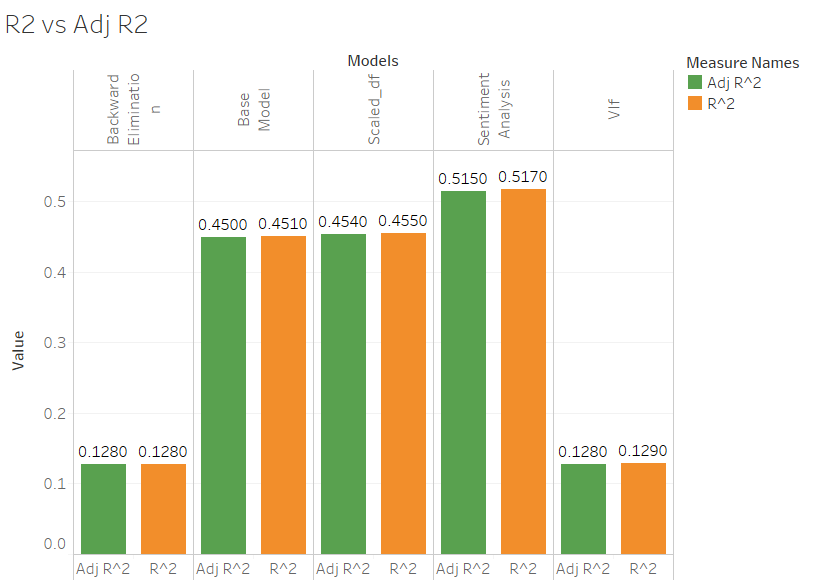
Feature selection is the process of selecting a subset of relevant attributes to be used in making the model in machine learning. Effective feature selection eliminates redundant variables and keeps only the best subset of predictors in the model which also gives shorter training times. Besides this, it avoids the curse of dimensionality and enhance generalization by reducing over fitting.

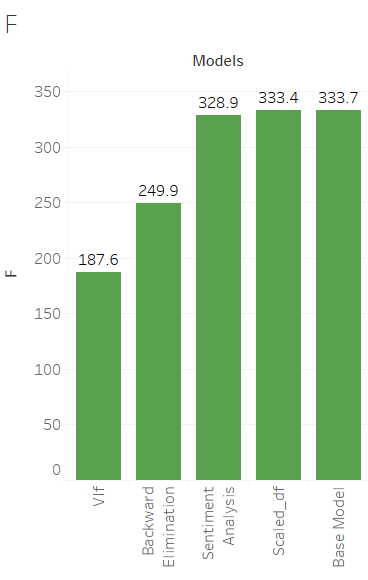
In this project, feature selection techniques are applied to improve the prediction performance and/or scalability of the system. Thus, we aim to investigate if better or similar classification performance can be achieved with a smaller number of features. An alternative of feature selection is the use a feature extraction technique such as Principal Component Analysis for dimensionality education. However, in this

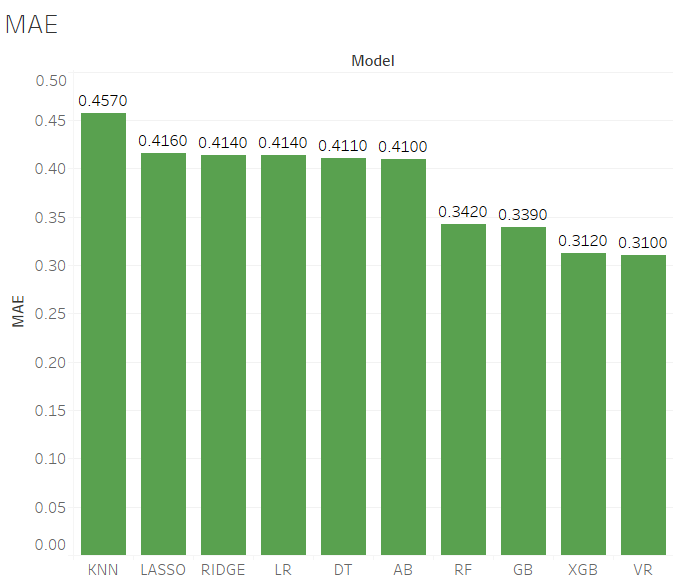
case, we use lasso and ridge technique to fine tune.

Thus, maximum prediction accuracy is aimed to be obtained with minimal subset of features. Here since the assumptions of linear regression are not satisfied we wont be bothered about the multicollinearity and PCA is not required in this case as we try to incorporate tree based approach to our data to build the model.

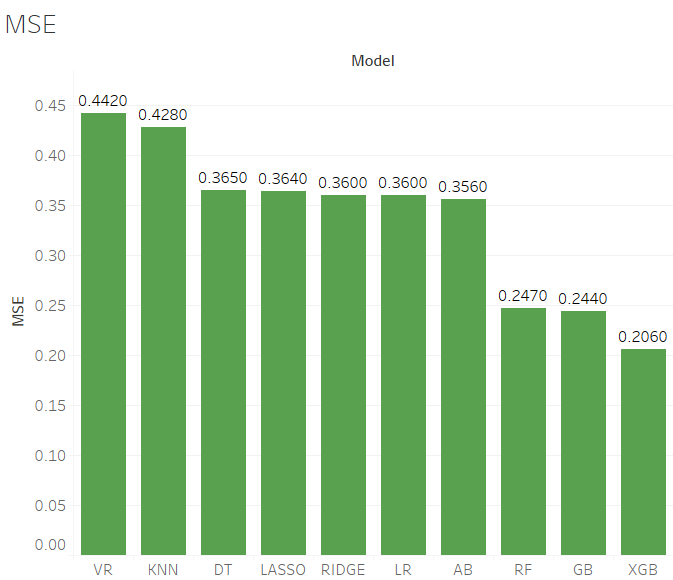
**Model Result:**

****

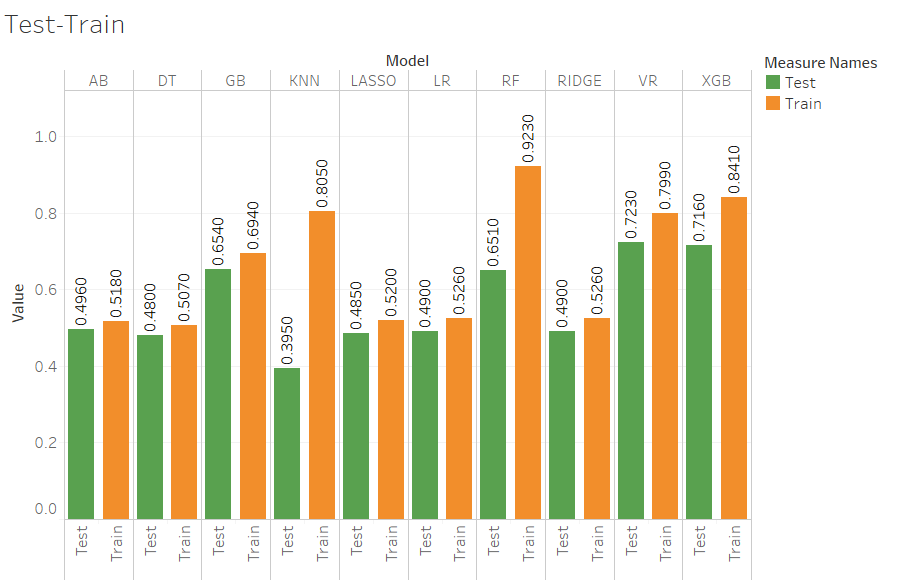
****

****

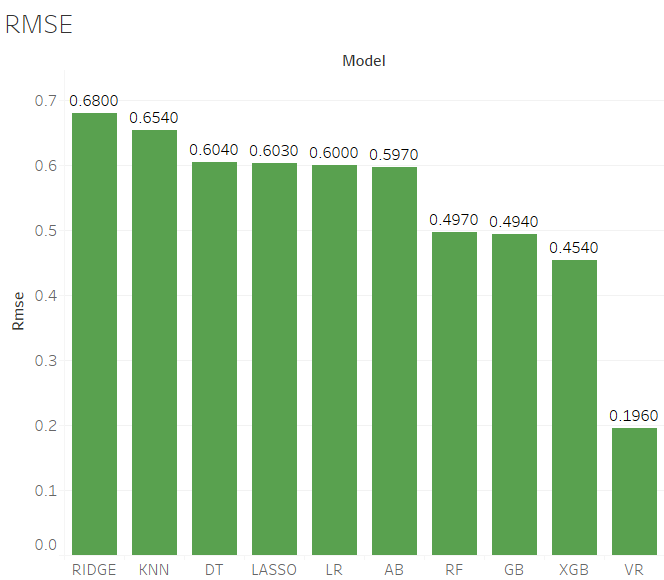
MAE stands for Mean Absolute Error. Voting regressor has the lowest MAE when compared to others

****

MSE stands for Mean Square Error. Xgboost has the lowest MSE score.

****

As we can see that train score is higher in random forest regressor but the test score is highest in voting regressor, but the random forest model is overfit, so we go with voting regressor where the score of test and train is almost similar and comparatively higher when compared to other models.

****

The RMSE score is the highest for ridge regression and smallest for voting regressor when compared to other models. RMSE stands for Root Mean Square Error. The lower RMSE score the better the model is.

**Results:**

In this project we analyse the price of the new host based on their listings and group the host based on their revenue to find and give some high priority. After data cleaning, the datasets are fed to linear regression. Since the linear regression fails in accepting the five assumptions, we gone to random forest regressor, gradient boosting regressor and adaptive boosting regressor using K-Fold and Grid search cross validations. After doing Regression algorithm, the dataset has been feature transformed by the standard scalar and then it fed to K-Means clustering for grouping the hosts based on their revenue. The Accuracy of both training and test split, R-Squared, Root Mean Squared Error are presented for each regressor.

In this project we predict the price of the listing listed by the host. After the pre-processing the dataset is fed to linear regression. After that we checked for the linear regression assumptions and some assumptions are fulfilled. Later we checked for multi collinearity and removed multicollinearity but still the variance explained was reduced far below than before. Then we did feature selection and compared the OLS model. The dataset which explains the better variance is taken.

Since, the multicollinearity exist PCA is carried out and linear regression model is built and metrics are checked. The model performance was not up to the mark.

We have dataset of the reviews given by the customer about the experience on the stay. From that we grouped the reviews with respect to the listing id and performed sentiment analysis using text blob .The output of the sentiment analysis will be polarity and subjectivity value for each review. These values are then joined to the main dataset and and OLS model is built. This model seemed to explain better variance. Thus, it is finalized and the other models are built.

**SUMMARY:**

We tried to optimize the pricing of AirBnB listings in the city of Barcelona. Our project helps hosts to earn more through AirBnB and thereby increasing the revenue of the company itself. We suggest hosts to either increase or decrease their price for the listing by predicting the optimum price for the place.

The sentiment analysis for reviews given by the customer is done and the subjectivity and polarity of the reviews is appended back to the dataset.

The OLS model for the dataset was built and this model seemed to have explained the variance better than other models.

Henceforth we built models based on other algorithms like Decision trees regressor and KNN regressor etc. Ensemble techniques like bagging, boosting and stacking were performed to get a better result.

We found that the voting regressor had the least RMSE and best accuracy scores. Finally, we compare the predicted price with the actual price.

Barcelona is infamous for its protests and marches against AirBnB, as the people believe the high prices set for listings has a direct effect on the rents of houses, our project may have a solution to this problem as we optimize prices for the listings.

And with the help of sentiment analysis on reviews, we club the review of each listing to get an impact factor for the pre-existing review scores and this helps AirBnB to categorize the hosts efficiently.

Bibiliography:

1. https://scikit-learn.org/stable/auto\_examples/impute/plot\_iterative\_imputer\_variants\_comparison.html
2. https://www.sciencedirect.com/science/article/abs/pii/S0261517717301036
3. https://www.sciencedirect.com/science/article/abs/pii/S0261517717301036
4. Do short-term rent platforms affect rents?-Munich Personal RePEc Archive
5. d Barcelona: now anti-tourism marches spread across Europe-Guardian
6. https://www.emerald.com/insight/content/doi/10.1108/JTF-11-2015-0048/full/html
7. When Tourists Move In: How Should Urban Planners Respond to Airbnb?**-**https://www.tandfonline.com/doi/full/10.1080/01944363.2016.1249011
8. https://towardsdatascience.com/random-forest-and-its-implementation- 71824ced454f
9. https://monkeylearn.com/sentiment-analysis/