Data Analysis 3 : Assignment II : Business Report

Introduction

The main goal of this project is to help a company to set price for their new apartments which are not yet in the market.To build a price prediction model for a company which is operating small and mid-size apartments hosting from two to six guests in the New York, the data is taken from Inside Airbnb which can be found here. As a result of data cleaning, munging, and analysis, five price prediction models, OLS, Lasso, Cart, Random Forest and GBM, Gradient Boosting Machine, are created. As a result GBM showed the best prediction result with 65.38 USD RMSE. The major predictor features are the number of accommodates, number of bathrooms, the number of beds, neighborhood, and amenities such as availability of washer, gym, elevator are the most important. Other characteristics such as days since the first review is also an important predictor. Eventually the final goal of the project is to finalize a better prediction model measured in relative RMSE values.

Data Prepration

Before initializing building predictive models, it is vital to clean the data. The original data set in the Airbnb New York consists of 38186 observations and 74 columns. The data refer to the one night rental prices between January 6 to January 9, 2022. The target variable is price per night per person in US dollars. To transform the original data to tidy data table, the major part of transformation consists of processing *amenities* column to binary variables. For meaningful grouping similar amenities were grouped together and only binaries between 1% to 99% values are kept in the data set. After merging 3134 amenities with similar attributes it resulted into 90 amenities. For detailed description of data cleaning steps please find the codes here. The main goal of the project is to predict prices of apartments which accommodates between 2 to 6 guests, thus the data was filtered accordingly. The key variable, price, contained extreme value and missing observations, thus it was filtered to contain values below 500 USD and dropping the observation where it is missing. As a result of of data cleaning, preparing and munging the total observations are 16271 with 135 columns.

Feature Engineering

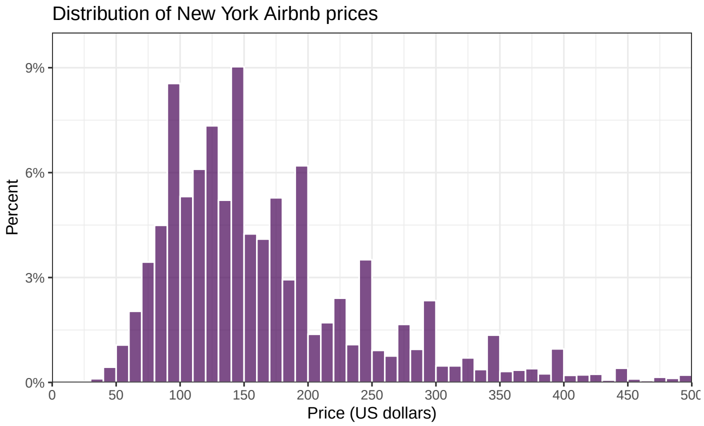
Feature engineering includes what type of predictor variables to include, and deciding about functional forms of predictors and possible interactions. The data is grouped as following:

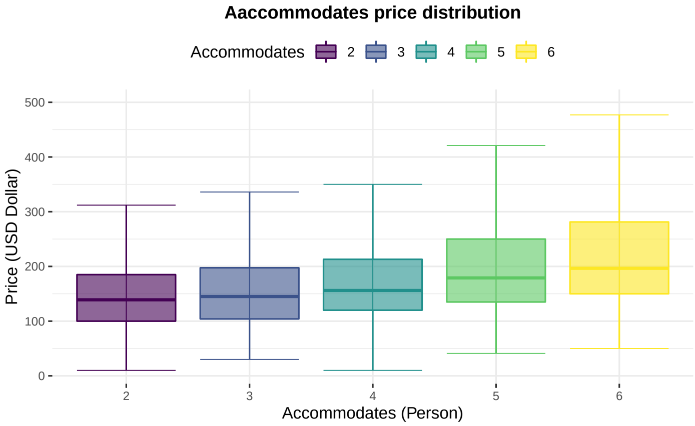
- **Basic variables** which consists of the main predictors such as: number of accommodates, property types, number of beds, number of days since the first review and flag variable of number of days since the first review to indicate missing. As the focus of project is on small and mid-size apartment thus, property types are as following: Entire home or apartment, serviced apartment, Condominium, and entire rental unit.
- **Basic addition** this includes key factorized variables, such as, neighborhoods groups, and host response time.
- **Review Variables** consists of the crucial guests reviews predictors such as total number of reviews, number of reviews per month and host review score rating and reviews flags which shows missing variables.
- **Polynomial level** consists of squared terms of guests and squared and cubic terms for days since the first review.
- **Amenities dummies** which consisted of the binary values for all of amenities.

After the key filters and grouping variables there were 15 variables with missing values. Missing values were addressed as following: first assumption is there is at least one bathroom in each apartment, second assumption is if the number of guests are less than 4 then impute 1, otherwise imputing 2. Missing number of beds were replace with half of number of accommodates, assuming there are double beds in the apartments. It is assumed that the minimum number of nights is one and minimum number of reviews is also 1. Flags were created to indicate the missing values in the each predictor. As the goal of project is to build price prediction model, it is crucial to check price and log of price distribution. Price distribution shows Airbnb apartment prices is skewed with a long right tail and the log price is close to normally distributed. In this project, log of price is not considered, prediction is carried out with price for all of the models.

Exploratory Data Analysis

Some of the important predictor variables are related to the size: for example the total number of guests an apartment can accommodate and number of beds, number of bathrooms. The below box plots show the average price per number of guests. On the other hand, it can be seen that apartments with many guests have high average prices.





Modeling

Regressions The best model gives the best prediction in the live data. Before turning to the modeling part of the project, it is worth mentioning that in order to avoid over fitting, the original data is split into two random parts by 20% to 80% ratio. Holdout set contains the 20% and the rest is work data set. In addition the k-fold cross-validation is a good way to find a model which gives the best prediction for the original data. For the purpose of this project 5-fold cross validation is used. This means splitting the data into five random samples and calculating and deciding based on the average of 5 CV RMSE result. Eight basic OLS regression models from simplest to the most complex one were used to find a better model to use for further analysis. The below model, Model 7 regression, had the lowest RMSE value. Further details are provided in the technical report. Model 7 consist of basic variables, basic additional variables, reviews variables, polynomial levels, amenities and interactions

Models

It is important to run and evaluate different models for a given data set. In order to predict apartment prices, the following models and algorithms were used. Naming them as following:

- **OLS** and **LASSO** using model 7 based CV-RMSE result
- **CART**, **Random Forest** , **GBM** using basic level variables, basic additions, review variables and amenities as dummy variables.

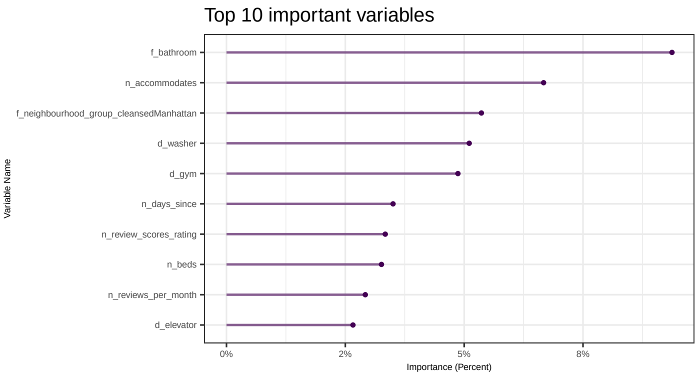
Based on the below table of models, it is can be seen that GBM model has the best performance. 5-fold cross validation RMSE for the data is 65.38 USD RMSE which is 1.683 USD RMSE less than the second best model which is Random Forest. Moreover, The 5-fold cross validation RMSE for the GBM model using Holdout set is 64.38 USD RMSE indicating a better performance than other models illustrating a better performance from second best model of random forest by 1.671 USD RMSE. GBM model tend to be robust, thus, the selected model for this project is GBM BASIC TUNING model.

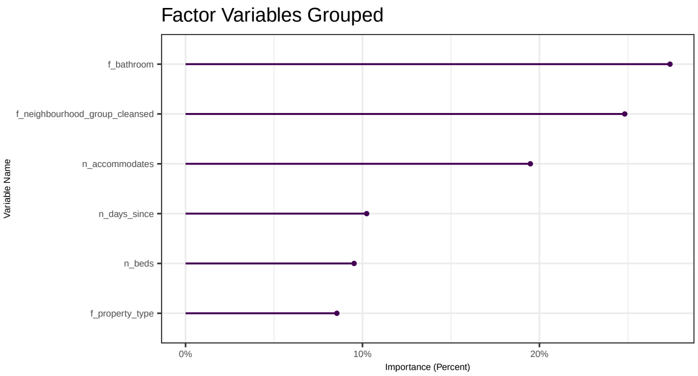
	CV RMSE	Holdout RMSE	CV Rsquared
OLS	67.99696	68.21073	0.3167363
LASSO	67.97247	68.12913	0.3171112
CART	71.88612	71.42794	0.2360636
Random Forest	67.06593	66.07839	0.3644942
GBM	65.38265	64.39938	0.3681212

Diagnostics

The best selected model, GBM, is an ensemble method which is a black box model, because it does not reveal the pattern of association that drive prediction. However, diagnostic tools can be used to uncover information about the patterns of association which drive prediction. Some of them are as following:

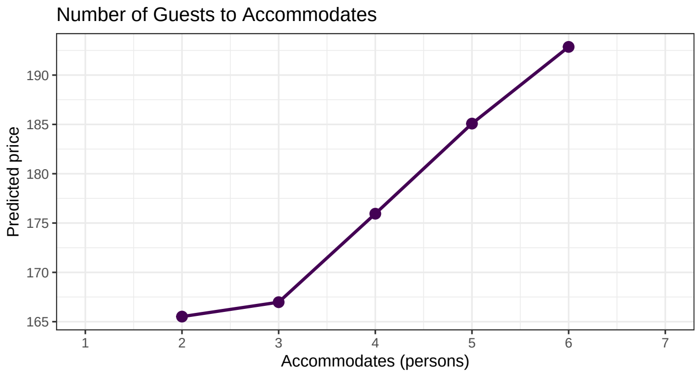
Variable Importance plot it shows the average importance of fit when we use an x variable or group of x variables. Variable importance plot for Top 10 important variables shows that number of bathrooms, number of accommodates, Manhattan neighborhood are the most important along with amenities such as washer, and gym. The grouped variable importance shows that bathrooms, neighborhoods, number of accommodates are the most important variables.

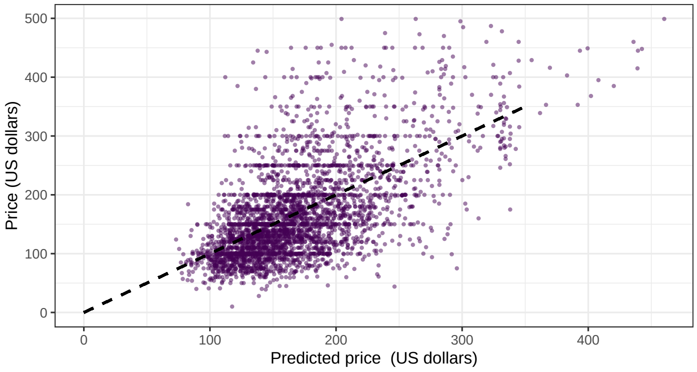




Partial Dependence Plot it shows how average y differs for different values of x conditional on all other predictor variables. Partial dependence plot is based on predictors for the holdout set. Partial dependence plot for number of accommodates and price shows that price increases as the number of accommodates.

Actual vs Predicted Price another post prediction diagnostics is comparing the predicted prices versus the actual prices. The figure below shows that prediction does a better job for lower than higher prices.





Conclusion

The goal of this report was to find a better model to predict Airbnb prices in New York for a small to mid-size apartments. Five models were illustrated to compare and contrast across models performance. GBM resulted to be the best model by 65.38 USD RMSE, besides this GBM model also shows better performance in holdout set with 64.38 USD RMSE. The second best model was basic Random Forest which has highlights meaningful characteristics about the nature of Airbnb apartments in New York. Key price drivers based on post prediction diagnostics are the number of bathroom, number of accommodates, and availability of amenities such as washer and gym. partial dependence plot also illustrated that the model better predicts Manhattan neighbourhood.