Technical Report

In this assignment, I predicted the prices of Airbnb apartments in New York, extracted from insideairbnb.com. The codes and outcomes can be found at https://github.com/Dilnovoz/Data-Analysis-3/tree/main/Assignment_2.

In the first step, I did feature engineering using the given listings.csv dataset. Irrelevant symbols ($, %) in some columns are removed to make as a readable number for analyses. Categorical and dummy features are marked down as categorical and boolean otherwise python could assume them just as a string values. Furthermore, unbalanced categories and categories with really few observations are recategorized or merged, so it is more balanced and comparable.

New features are also created. By using information about first review date, I calculated how long the apartment is in the market as it can be important determinant of the price determination. Furthermore, amenities that apartment includes were given as a list form in one column. To construct valuable features from this information, first, number of amenities offered is counted and used as a feature. Further, around 100 most common amenities that are mentioned by more than 1000 hosts are created as separate dummy variables.

Missing values in the data is solved in 3 ways based on its relevance and number of the missings. Variables including only a few missing observations are filled with logical values without flagging them out. To illustrate, 20 number of bathrooms were missing, and I changed them to 1 as I assumed each apartment should have at least 1 bathroom and if it had many bathrooms, the host would point it out. Second a set of features where many observations are missing but can be represented by other features were dropped from the dataset. For example, the review scores for different characteristics of apartment were given but all of them had many missing observations. I kept only the one with fewer missing values and I recoded them to the median value. I also flagged it out too.

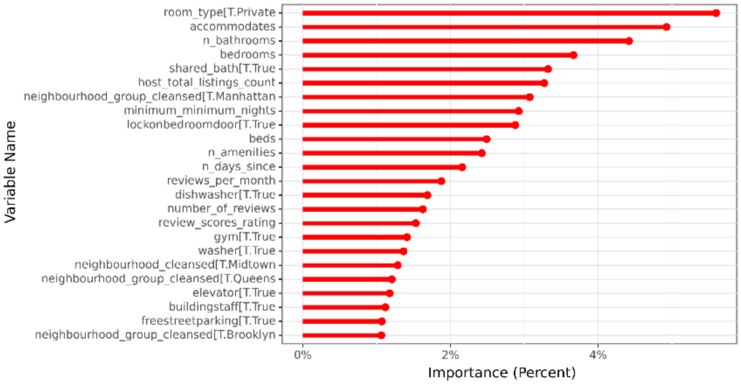
After the feature engineering is completed, the data is saved as csv form and prediction analyses are estimated based on it in the second Jupyter notebook. Before starting our model choice and prediction, two features related restrictions are introduced to make is more similar to the apartments that are planned to enter to the market. Apartments that can host 2-6 individuals are only kept in the analyses. However, it was a bit challenging to keep only apartments in the analyses. Host mentioned different type of property types and defined them quite ambiguously. I removed houses and hotels from the dataset as we are interested in only apartments and assumed rental unit, home, loft, condo, serviced apartments and guest suite categories can show different apartment types. Hence, there may be some measurement error and some unsimilar properties maybe kept in the analyses.

Additionally, before predicting the price of the apartments – target variable, some extreme values or really expensive prices are removed from analyses as they are quite unlikely, outlier values. These observations constitute less than 1 percent of the dataset.

Models	RMSE
OLS	87.15
Lasso	88.45
CART	114.1
Random Forest	85.47

As a model choice, I compared 4 different models (OLS, Lasso, CART and Random Forest) and chose based on their root mean squared errors (RMSE) statistics. For OLS and Lasso regression estimation, I included log form and polynomial form of some features as some of them are not normally distributed and the association could be not linear. Furthermore, interaction terms are introduced to the model so Lasso select the most relevant features among the big range of features. For the case of CART and Random Forest, the features are included as given since there is no closed form pre-assumed association form between target variable and features. According to RMSE values, Random Forest is the best model to predict the prices of apartments.

Further, diagnostic analyses showed that most important features to predict the rental price of the apartments in New York is the type of the room, followed by the number of accommodates it can guest and number of bathrooms in the apartment. As these features are a bit hard to interpret and discuss further grouping of the features and better visualization forms are presented.


Variable Name
room_type[T.Private
accommodates
n_bathrooms
bedrooms
shared_bath[T.True
host_total_listings_count
neighbourhood_group_cleansed[T.Manhattan
minimum_minimum_nights
lockonbedroomdoor[T.True
beds
n_amenities
n_days_since
reviews_per_month
dishwasher[T.True
number_of_reviews
review_scores_rating
gym[T.True
washer[T.True
neighbourhood_cleansed[T.Midtown
neighbourhood_group_cleansed[T.Queens
elevator[T.True
buildingstaff[T.True
freestreetparking[T.True
neighbourhood_group_cleansed[T.Brooklyn
Importance (Percent)