KAGGLE PRESENTATION

Predicting Rental Price For AIRBNB

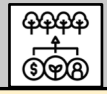
Data Exploration and Visualization

- Exploring the type and values of existing 91 variables in AnalysisData
- Exploring outliers of some variables such as price, minimum days using ggplot
- Adding some levels missing in train data
- Scoping the data for additional potential variables

Data Cleaning

- Finding the multiple NA, N/A or “” values present in data
- Changing the dates in the data to more useful formats
- Counting the characters of the factors with multiple levels
- Logically replacing the various types of NA with mean, median or 0

Data Analysis and Modelling



- Trying out the linear regression model to fit all variables
- Testing the submission file for NA values to see if all data is cleaned properly o no
- Using random forest (turned out to be so time consuming!!)
- Hyper parameter tuning of **gradient boosting model** (gbm) – number of trees, interaction depth and shrinkage

Feature engineering

- Extracting the count of “luxury” & “luxurious” from 5 variables
- Extracting the transit modes “bus”, “train” and “subways” from transit
- Extracting latitude and longitudes from zip codes
- Extracting number of amenities with help of commas

Things that went wrong

- Methodology of selecting the variables
- Using the parameters of trees to fine tune the various models
- Running models for hours and then realizing some unclean data
- Trying to utilize all variables

Final results and learning

- RMSE public : 55.98079
- RMSE private : 57.43220

Don’t give up and have enough time to run models or to test them and probably learn NLP and SQL for faster and shorter codes