KAGGLE PROJECT: PREDICTING AIRBNB RENTAL PRICES

EXPLORING DATA

As part of the project, we were given access to a dataset of 25,000 Airbnb rentals with over 90 features. The goal of this project was to create a model that accurately predicts the rental prices. I have never stayed in an Airbnb hence the initial phase of this project involved a lot of research to understand the dataset. The first step was to read through the Airbnb website to get a sense of their business, offerings and researching on terms such as “superhost”.

PREPARING DATA & MODELLING

Rather than cleaning the entire dataset at one time, I only cleaned data for the features I was experimenting with. I ran an str() on to understand the data, different levels, types of features, and the number of NA. I initially started by eliminating:

- Features that had all N/A such as “host_acceptance”,
- Features that had majority N/A such as “square_feet”, and
- Other features that were the same across of listings such as “market”, “country_code”, and “country”

I then went on to select features that I would personally look for when renting an Airbnb like all review scores, property type, room type, bathrooms, beds, security deposit, guests included, etc. Several features contained NA values which I replaced with either the mean value or zero depending on the nature of the feature.

My first few submissions were a trial run using cross-validation and a linear regression model. This was extremely fast and improved the RMSE. I continued with this model adding and removing simple features that did not need to altered, to see the impact on the RMSE and stuck with some important features such as “neighborhood_group_cleanse”.

Once I was done eliminating all the simple numeric and boolean features, I spent ample time on stackoverflow in attempting and failing on how to re-engineer features to incorporate into the model. I ended up converting string data into date for “host_since”, “first_since”, and “last_since” features and performed a word count of text columns. This in combination with a boosting model significantly improved my RMSE.

Further, I used the grepl function to identify key features such as Wi-Fi, heating, and breakfast among others in amenity columns. This slightly improved the RMSE but after a point, it was pretty stagnant so I moved on to random forests. These were taking too long run so I installed the ranger package that tremendously improved the speed with which it ran random forests.

One of the most valuable feature engineering was to use the zip codes for latitude and longitude. I cleaned the zip code column to change listings with grouped zip codes to one zip code, removed unnecessary information from the zip code cells, converted all zip codes to 5 digits and replaced all blank and uncharacteristic rows with the most common zip code. This was a last-minute success that drastically improved my RMSE.

CONCLUSION

This project provided a good opportunity to apply and experiment with the various concepts taught in class. As I hone my skills in R, I would like to employ feature selection techniques such as correlation and lasso to identify the key features rather than using a trial and error method, and also create functions and loops to clean data that helps avoid repetitive coding on the two data sets.