Chenning Zhang
Rober Simione
December 14, 2020

Kaggle Project Final Report

Initial exploration:

This project aims to construct a model using the dataset supplied a use it to predict the price of a set of Airbnb rental included. Before we started cleaning the data, we need to identify all parameters' category. The metrics in this data can be simply separated into the following categories, description, multiple-choice, date, logical, and numerical. We noticed that all categories of data have some kind of missing values, and categories of description and multiple-choice have the most missing value. Some data in date and numerical have partially missing data, and logical data have the lowest missing rate. To clean the data of description and multiple-choice have the highest difficulty, relatively, data in date and numerical will be easier to clean. Those missing values bring trouble in the prediction model, the machine is unable to learn from empty value, to find out the best model, we need to have a cleaned data than test data by different models.

Data Cleanning:

For the data related to description and multiple-choice, it is hard to cleaning. Thus, we focused on the data related to logical and numerical. If we find out a logical missing value, we will replace it with "TRUE". If we find out a numerical missing value, we will replace it with the mean value of all existing data. Most numerical data we have are easy to have their mean value which helps us to find models in the future.

Models and Feature selection:

The first model we chose to test is the linear regression model with the feature of “minimum_nights” and “review_scores_accuracy”. Because it is the basic model with a default feature selected by the question, the RMSE is very large. Thus, we select more features to update our linear regression model. To select correct features, we use Forward Selection to find out. After we use Forward Selection, we find out the following features are critical in the linear regression model. The second model we chose to test is also linear regression model with feature of bedrooms + guests_included + bathrooms + review_scores_accuracy + minimum_nights+beds. We need to have an upgrade on model selection, the third model we use is the regression tree model, and we use the same feature in model 2. Meanwhile, we also select the boosting model as the fourth model to predict, and the features we selected are the same as the previous two models. At last, we find out the random forest model has the largest potential to be the best model, then, we select all features we could select which is been data cleaned.

(bedrooms + guests_included + bathrooms + review_scores_accuracy + minimum_nights+beds+ accommodates+ square_feet + weekly_price+ monthly_price + security_deposit+host_neighbourhood + host_listings_count+host_total_listings_count+ host_is_superhost + host_acceptance_rate+ host_response_time + host_response_rate +host_has_profile_pic+ host_identity_verified + is_location_exact+ host_since + host_name+ neighbourhood+ neighbourhood_cleansed+ neighbourhood_group_cleansed+city+ state+property_type+room_type+ bed_type+cleaning_fee+extra_people+minimum_nights+maximum_nights+minimum_minimum_nights+maximum_minimum_nights+minimum_maximum_nights+minimum_nights_avg_ntm+ maximum_nights_avg_ntm+has_availability+availability_30+availability_60+availability_90+availability_365+number_of_reviews+number_of_reviews_ltm+first_review+last_review+review_scores_rating+review_scores_accuracy+review_scores_cleanliness+review_scores_checkin+review_scores_communication+review_scores_location+review_scores_value+instant_bookable+requires_license+is_business_travel_ready+cancellation_policy+require_guest_profile_picture+require_guest_phone_verification+calculated_host_listings_count+calculated_host_listings_count _entire_homes+calculated_host_listings_count_private_rooms+calculated_host_listings_count_shared_rooms+reviews_per_month)

Model Comparison:

Model from previous selection	RMSE on training data	RMSE on Kaggle	Other notes
Model 1: linear regression	109.9649	104.52749	
Model 2: linear regression	95.72844	90.15401	Same features selected for model 1
Model 3: regression Tree model	90.63579	86.03443	Same features selected for model 1
Model 4: Boosting model	87.24839	84.02730	Same features selected for model 1
Model 5: Random Forest model	69.30943	63.45526	The best model we chose, largest feature selection we had

Discussion:

After we select multiple models and cross-compare those different models, the random forest model has the smallest RMSE in the same feature selection. Besides of select the best model, we also need a good selection of features. After data cleaning, we selected all features that have no empty value, and this process can bring as many parameters as we can to increase the accuracy of the model. As the final result, the best RMSE on Kaggle we have is 63.455. Compare with the range of price is from $50 to $500, 60 RMSE is not a good result we should accept, we don’t think this model can make any inferences about the relationships between predictor variables and Airbnb prices.

Future Directions:

This model still has some limitations. Firstly, features selected in the model are only limited to logical and numerical, if we have a better data cleaning technique to clean the data of description and multiple-choice, there will be more features can be selected into a model. Under the current situation, the random forest model is the best model we have, thus, the selection of features became a key fact. If we can select the keywords in the description and find the relationship between its price, it may improve our model.