Assignment 2 **Abay Jumabayev**

Objective:

To help a company operating small and mid-size apartments hosting 2-6 guests to set the price for new apartments not on the market.

Data:

I used AirBNB data for the city of Vienna. Original data has 11,583 observations and 74 columns.

Variables:

I created/transformed several variables including the dependent variable – *log(price).* I took logs because the price distribution was skewed to the right.

Other variables and their anticipated effect on price:

- 'host_response_time': the faster host responses, the higher the price (for his services). Unfortunately, there were too many missing variables, I had to drop the variable.
- 'host_response_rate': logic very similar to the previous variable. The same problem of missing variables appeared.
- 'host_has_profile_pic': if the host has the picture, customers would trust more and be willing to give up their money.
- 'host_identity_verified': logic very similar to the previous variable.
- 'neighbourhood_cleansed': some neighbourhoods are nicer, therefore, the price would go up.
- 'latitude' and 'longitude': those variables are needed to calculate the distance from the
- 'property_type': this variable is needed to make sure that we are focusing on apartments. Possibly this variable will be helpful in predicting the price (e.g condominiums might be more expensive that apartments).
- 'accommodates': this variable is needed to make sure that we are focusing on small sized (2-6 people) apartments. Moreover, this this variable will be helpful in predicting the price as the more property can accommodate the higher the price.
- 'bathrooms_text', 'beds', 'bedrooms': the greater the number of bathrooms, beds, and bedrooms, the higher the price.
- 'amenities': the more amenities house the higher the price.
- 'instant_bookable', 'minimum_nights' and 'maximum_nights': the more possibilities there are, the higher the price.

Variable excluded from the dataset:

- Name and description of a property: 'listing_url', 'scrape_id', 'last_scraped', 'name', 'neighborhood_overview', 'picture_url', 'license', 'room_type'. Those variables do not affect price at all.
- Host information that we cannot change: 'host_id', 'host_url', 'host_name', 'host_about', 'host_thumbnail_url', 'host_since', 'host_location', 'host_acceptance_rate', 'host_is_superhost', 'host_picture_url', 'host_neighbourhood', 'host_listings_count', 'host_total_listings_count', 'host_verifications', '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'. Those are the variables that the host cannot easily change. Although, some of the variables probably have predictive powers.

- Variables that will not be available for a new apartment: 'has_availability', 'availability_30', 'availability_60', 'availability_90', 'availability_365', 'calendar_last_scraped', 'number_of_reviews', 'number_of_reviews_ltm', 'number_of_reviews_l30d', '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'. As we focus on new apartments I decided to exclude characteristics which will appear with time.
- Other variables: 'neighbourhood', 'neighbourhood_group_cleansed', 'minimum_minimum_nights', 'maximum_minimum_nights', 'minimum_maximum_nights', 'maximum_maximum_nights', 'minimum_nights_avg_ntm', 'maximum_nights_avg_ntm', 'calendar_updated'. Those are calculated/unnecessary variables that are not helpful for the analysis.

After removing unnecessary columns, I started to remove unnecessary rows. I started by deleting observations for non-apartments. Out of 56 types of properties, I chose 4 types that are considered apartments. I found that by searching for “apartment” keyword in the description. I lost about 20% of observations after I removed all non-apartments.

The next step was to remove all properties that accommodate 1 or greater than 6 people. After removing those observations, I have left with 7,861 observations.

Then I cleaned all variables. The process was cumbersome, but straightforward. The problem appeared when I tried to create dummies out of “amenities” column. There were 586 dummies. I combined dummies that were similar to each other (e.g. “Beko oven” and “oven” or “wifi” and “wi-fi”). I decreased the number of dummies in half. Then, I deleted all dummies, that had less than 300 in sum.

Variables selection:

I ran regressions of each of the variable on price and decided which one to include in the model and which one not. All excluded variables were included in the Model 3. Those are: 'f_property_type', 'f_maximum_nights', 'd_host_has_profile_pic', 'd_host_identity_verified', 'd_instant_bookable'.

Models:

Model 1: $\log(price)\ \sim n\_accommodates + n\_bedrooms + n\_dist\_from\_center + n\_minimum\_nights + f\_bathroom + f\_beds + f\_neighbourhood$

Model 2: Model 1 Variables + Polynomials + Interaction terms

Model 3: Model 2 Variables + Predictors which did not show much predictive power

Model 4: Model 3 Variables + Amenities dummies

In model 1 I put only the variables that are more likely to affect price. In model 2 I added squares and cubes of number of accommodates, number of bedrooms, and distance from city center. I also added an interaction term between the distance and neighbourhood. In model 3 I added Predictors which did not show much predictive power, like host information, maximum nights, and property type. In model 4 I added 63 dummies for amenities.

Results:

	CV RMSE	BIC
Model1	0.432786	9480.245099

Model2	0.426031	9467.292964
Model3	0.418678	5329.233973
Model4	0.390788	5203.702409

As it can be seen from the table above, cross-validated RMSE and BIC suggest that the Model 4 has the best performance. Including amenities improves the model results much. I expect that Random Forest results will be similar to the ones from cross-validated RMSE and BIC.