AirBnB Price Prediction Challenge

Process Doc:

- Below are the Pre Processing, Modeling and Visualization techniques that I followed for challenge.

Data Types

Different Data Types are present in the dataset. Following are the observations on Datatypes

- id, log_price, accommodates, bathrooms, host_response_rate, latitude, longitude, number_of_reviews, review_scores_rating, zipcode, bedrooms, beds are Numerical Data Type.
- property_type, room_type, bed_type, cancellation_policy, city, neighbourhood are Categorical Data Type.
- cleaning_fee, host_has_profile_pic, host_identity_verified, instant_bookable are Boolean Data Type.
- amenities, description, name are Text Data.
- host_since, first_review, last_review are Dates.
- thumbnail_url are urls of room images.

Cleaning the Boolean Data:

- True and False in host_has_profile_pic, instant_bookable, host_identity_verified is noted with t and f.
- Replacing them with True and False before changing them to bool data type.
- There are 21 missing values in host_since, host_has_profile_pic and host_identity_verified and all belong to NYC city. We can get them verified manaually if needed.
- We cannot make any assumptions in host_has_profile_pic, host_identity_verified. Though the mode for each category is True. It is better to keep them as False for security reasons.
- There is no better way to impute missing values in host_since as it does not make sence to impute with mode or median.

Dependency of first_review, last_review and review_scores_rating on number_of_reviews:

- first_review, last_review and review_scores_rating are dependent on the number_of_reviews. If number_of_reviews is 0 then it is sure that the property doesnt have a first and last review.
- Replacing 2165 missing values in first_review, last_review with 'No Review'.
- Replacing 2165 missing values in review_scores_rating with 'No Review'.

Binning of host_reponse_rate

- There are around 25% of mising values in host_response_rate.
- As per AirBnb, host_response_rate is the number of inquiries to which a host has responded to within 24 hours divided by the total number of inquiries a host has received in the past 90 days.
- There is a chance that the NaNs are caused by division by zero errors.
- Imputing them with mode or median can cause deviations in the model. Filling them with -1 and later binning the feature and marking them and 'Unknown' seems to be the best approach.
- host_reponse_rate is binned into groups named as 'Unknown','0%-25%','25%-75%','75%-99%','100%'.

thumbnail_url

- Thumbnail URL have missing values and cannot be treated.
- For further feature engineering images can be webscrapped from the urls and values like R,G,B,brightness, Sharpness can be calculated.
- For this project this feature is dropped.

Missing values in neighbourhood and zipcode.

- Zip Codes
  - Zip codes is given as object data type and it contains decimal values and few type errors. These are cleaned using string manipulation techniques.
  - Distance matrix is calculated using euclidean_distances method on latitudes and longitudes and missing zipcodes are imputed with nearest values.
- Neighbourhood
  - Distance matrix is calculated using euclidean_distances method on latitudes and longitudes and missing zipcodes are imputed with nearest values.

Missing values in bathrooms, bedrooms, beds.

- These usually depend on accomodates. Replaced the missing values with the observations from heatmaps between accomodates and respective features.
- Bedrooms and Beds are highly correlated. So dropped beds in model building.

Model With Limited Features

- Numerical columns are scaled with minmax scaler.
- pandas get dummies are used for dummyfying the categorical variables.
- XGBoost model is built with limited features.
- After further tuning and cross validations.
  - Train RMSE - 0.28
  - Validation RMSE - 0.41

Amenities

- Amenities are cleaned and number of anemities features is created.
- Word clouds on Amenities are performed.
  - Most basic essentials are listed by many properties.
  - Fire safety features are aslo listed by many properties, which is a good sign.
- Frequency Distribution of Top 50 frequently listed amenities are ploted.
- Frequency Distribution of Top 50 rarely listed amenities are ploted.

Description

- Descriptions are cleaned and processed by removing contractions, stop words and lemmatizing.

Topic Modelling on Descriptions

- LDA is performed on Descriptions.
- Plot between coherence nad number of topics showed that coherence is higest for 4 topics.
- pyldavis is used to visualize the topics.
- After Analyzing the plot following topics are interpreted.
  - Topic 1: Friendly neighbourhood with parks to walk around.
  - Topic 2: Neighborhood with easily accessible transit places like Trains.
  - Topic 3: Private are with some accessibity to close by places.
  - Topic 4: Private and luxury.
  - NOTE: Results may vary if re-run LDA

City Wise Scatterplots

- Patterns of high priced neighbourhoods are observed in each city location.

Models

- Neural Net
  - Train RMSE - 0.43
  - Validation RMSE - 0.46
- XGBoost CV
  - Train RMSE - 0.31
  - Validation RMSE - 0.38
- XGBoost Grid Search
  - Train RMSE - 0.34
  - Validation RMSE - 0.34

Other Methods that I tried that did not improve the score are

- Used glove pretrained word embeddings on description and used an LSTM network and tried to predict log_price. It did not improve the score.
- LSTM
  - Train RMSE : 0.29
  - Validation RMSE : 0.76
- NOTE: This model is excluded from the current code as it requires glove vectors file to run without errors.

Further improvements that could be done.

- Clusters can be formed by analyzing the patterns in scatter plots of cities.