Capstone Two Report

1. Problem statement

Maximize profit of Airbnb in NYC
Assumption: high review_score_rating can improve occupation rate
Profit=Price*Occupation rate, Cost=Cleaning Fee
Goal: Improve review_score_rating
Find Feature and take action to improve review_score_rating

2. Data Preprocessing

Select rows in NY state, as we only want to make prediction for NYC.
Learn about the business meaning of feature, delete unrelated features
Calculate missing percentage for each column
Delete columns>80% missing
Correlation Matrix, delete highly-correlated feature
Delete rows<=20% missing
Impute mean/null/0 with <=5% missing

3. Datatype and Outliers

Learn about Data Type of Each Column
Change Boolean to 0&1
One hot encoding for categorical type
String to Numerical/Datetime type
Use zipcode as index, simplify city according to zipcode
Delete outliers

4. EDA

Check Price Related Features
Check Review-Score Related Features
Chi-square test for categorical variable correlation, all test pass
Anova for correlation between categorical variable and numerical variable, all test pass
T-test for Boolean variable, delete the one can not pass test
Histogram check normal distribution for independent variables
Scatter plot check correlation between independent variable and target variable
Use standard scaler to

5. Model Development

Linear Regression, Random Forest Regressor, XG Boosting Regressor

6. Model Evaluation

Mean Absolute Percentage Error
ROOT Mean Square Error
Linear Regression: (MAPE: 4.4, RMSE:5.79)
Random Forest Regressor: (MAPE: 4.3, RMSE:5.85)
Xgboost Regressor: (MAPE: 5.37, RMSE:6.29)

Model Selection: Linear Regression & Random Forest Regression

7. Model Interpretation

Use p-value select significant feature,
Use coefficient from linear regression result to evaluate the importance of feature
Use random forest feature importance to assist the feature analysis
(important features)

- **host_since**
- **host_response_rate**
- **host_is_superhost**
- **host_identity_verified**
- **accommodates**
- **price**
- **security_deposit**
- **minimum_nights**
- **minimum_nights_avg_ntm**
- **availability_30**
- **availability_365**
- **number_of_reviews**
- **review_scores_communication**
- **review_scores_location**
- **instant_bookable**
- **require_guest_phone_verification**
- **review_diff**
- **host_response_time_within a few hours**
- **host_response_time_within an hour**
- **cancellation_policy_moderate**
- **cancellation_policy_super_strict_60**

8. Conclusion

Different location (not review score) affect price
Object factors: Location/Communication/Price
Market factors: Review Frequency/#of Reviews/Host_since
Subject factors: Availability/Accommodates/Min nights/Cancel Policy Moderate/ Host Response Few Hours
Non-related factors: Property Type/Room Type