ITCS5156 - Applied Machine Learning Project Presentation

Prasanna Shanmugakumarasamy
pshanmu4@uncc.edu

Project Title and Sources

- Title: **Predicting Airbnb Listing Price Across New York**
- Research Paper Source:
  - Journal Name: **Researchgate.net**
  - Publication Year: **2020**
  - Author: **Brahmaih, Kala**
- UNCC Student doing this project:
  - **Prasanna Shanmugakumarasamy**

Problem and challenges

Problem:

- Airbnb is an online marketplace for property owners seeking tenants and renters.
- Pricing a rental property right is a difficult task for the owner due to various factors such as location, market demand , neighbourhood influences etc.,.
- Customers also assess the offered price with very limited knowledge of the property's optimum value.
- Need for a more scientific approach to arriving at optimum prices that benefits both the owner and the customers who rent the property.
- Use machine learning techniques to predict the prices

Motivation

- Worldwide rental service
- Every city is different and the factors influencing pricing strategies are different.
- New York City (NYC) is a unique metropolitan city that has tourists, residents and floating populations and a fast paced environment.
- Geographical location, neighborhood influences and amenities and property type are very different within this one city and that makes it very interesting to analyze.
- Paper also discusses some modeling techniques that are new to me as an ML student.

Related Works

Primary Work (to Duplicate):

- Brahmaih, Kala. "Predicting Airbnb Listing Price Across New York". *Researchgate.net*. (2020).

Reference Papers:

- Rezazadeh Kalehbasti, Pouya et al. "Airbnb Price Prediction Using Machine Learning and Sentiment Analysis". *Machine Learning and Knowledge Extraction*. (2021): 173–184.

  Why? *This paper **discusses sentiment analysis in addition** to the Price Prediction using Machine learning techniques. Text Mining and Analysis is an area that is covered additionally. It uses the customer reviews dataset to deduce this information. Also, the reviews dataset can be condensed and combined with the listings dataset.*

- duygut. “Airbnb NYC Price Prediction | Kaggle.” *Kaggle: Your Machine Learning and Data Science Community*, Kaggle, 22 Oct. 2019, https://www.kaggle.com/duygut/airbnb-nyc-price-prediction.

  Why? *This is one of the Kaggle Work that is most discussed. This uses **Ridge, Lasso and Elasticnet** regression models that were discussed in the class. It will be a good place to apply and evaluate. Also, it uses GridSearchCV algorithm that helps identify best hyperparameters for each model.*

- Dan Wang and Juan L Nicolau. Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb. com. International Journal of Hospitality Management, 62:120–131, 2017

  Why? *This is one of the most referred research paper amongs every research project that uses Airbnb dataset. It covers a wide range of cities and since each city is different the exploratory analysis are entirely different and the path to modeing and analysis are different as well. This serves as a base study and apply some techniques based on the results for the New York city that we’re considering to analyze.*

Dataset Introduction

- Sources
  - https://www.kaggle.com/duygut/airbnb-nyc-price-prediction/data

    Primary data source that was used in the research paper and in my project.

  - http://insideairbnb.com/get-the-data.html (Scroll down to New York)

    Airbnb's open data for recent periods that provides other datasets such as reviews and calendars. This provides an overall understanding of what data is captured for various analysis use.

Dataset Introduction

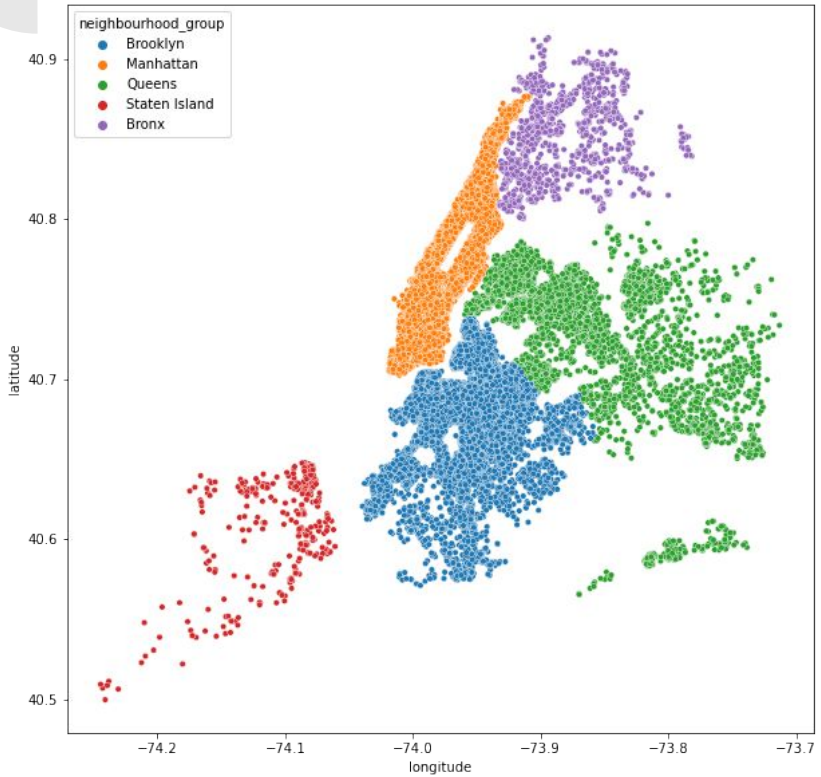
Contd..

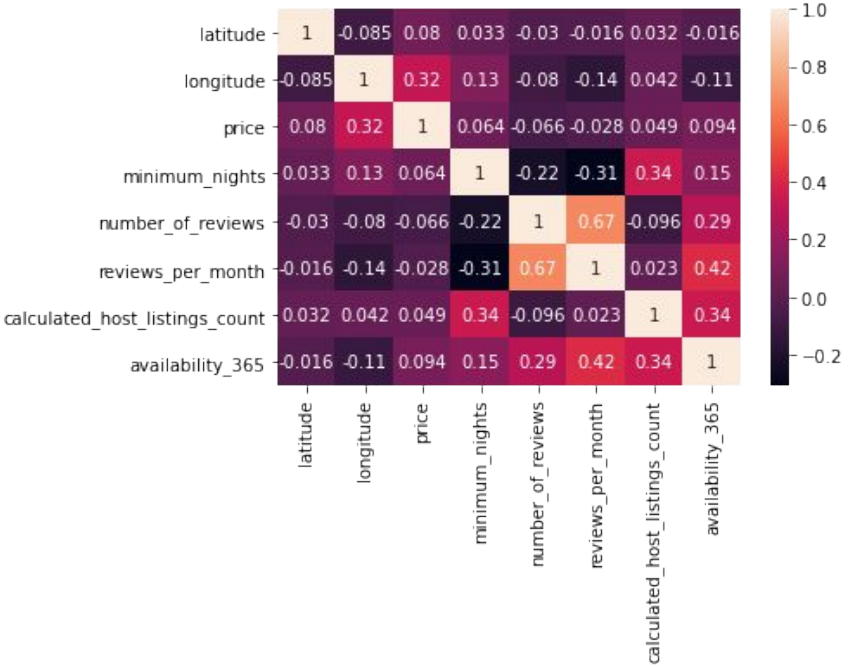
	0	1	2	3	4	5	6
id	2539	2595	3647	3831	5022	5099	5121
name	Clean & quiet apt home by the park	Skylit Midtown Castle	THE VILLAGE OF HARLEM....NEW YORK !	Cozy Entire Floor of Brownstone	Entire Apt: Spacious Studio/Loft by central park	Large Cozy 1 BR Apartment In Midtown East	BlissArtsSpace!
host_id	2787	2845	4632	4869	7192	7322	7356
host_name	John	Jennifer	Elisabeth	LisaRoxanne	Laura	Chris	Garon
neighbourhood_group	Brooklyn	Manhattan	Manhattan	Brooklyn	Manhattan	Manhattan	Brooklyn
neighbourhood	Kensington	Midtown	Harlem	Clinton Hill	East Harlem	Murray Hill	Bedford-Stuyvesant
latitude	40.6475	40.7536	40.809	40.6851	40.7985	40.7477	40.6869
longitude	-73.9724	-73.9838	-73.9419	-73.9598	-73.944	-73.975	-73.956
room_type	Private room	Entire home/apt	Private room	Entire home/apt	Entire home/apt	Entire home/apt	Private room
price	149	225	150	89	80	200	60
minimum_nights	1	1	3	1	10	3	45
number_of_reviews	9	45	0	270	9	74	49
last_review	2018-10-19	2019-05-21	NaN	2019-07-05	2018-11-19	2019-06-22	2017-10-05
reviews_per_month	0.21	0.38	NaN	4.64	0.1	0.59	0.4
calculated_host_listings_count	6	2	1	1	1	1	1
availability_365	365	355	365	194	0	129	0

```
id                                int64
name                             object
host_id                           int64
host_name                        object
neighbourhood_group              object
neighbourhood                    object
latitude                        float64
longitude                       float64
room_type                        object
price                             int64
minimum_nights                    int64
number_of_reviews                 int64
last_review                      object
reviews_per_month               float64
calculated_host_listings_count    int64
availability_365                  int64
```

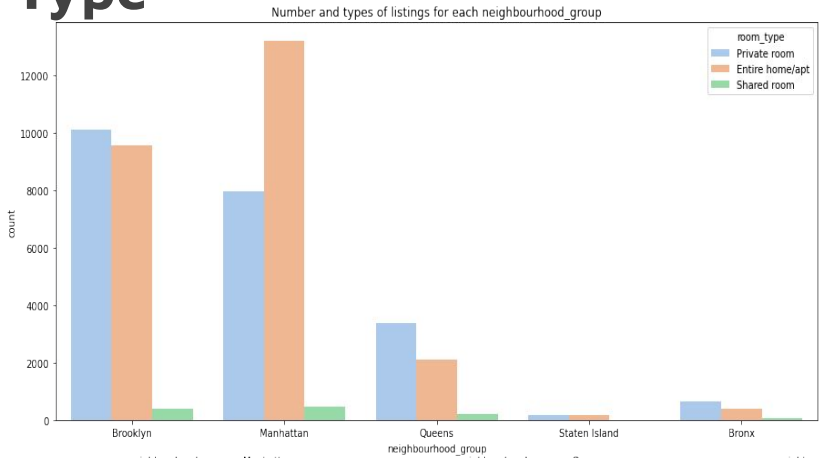
48895 rows

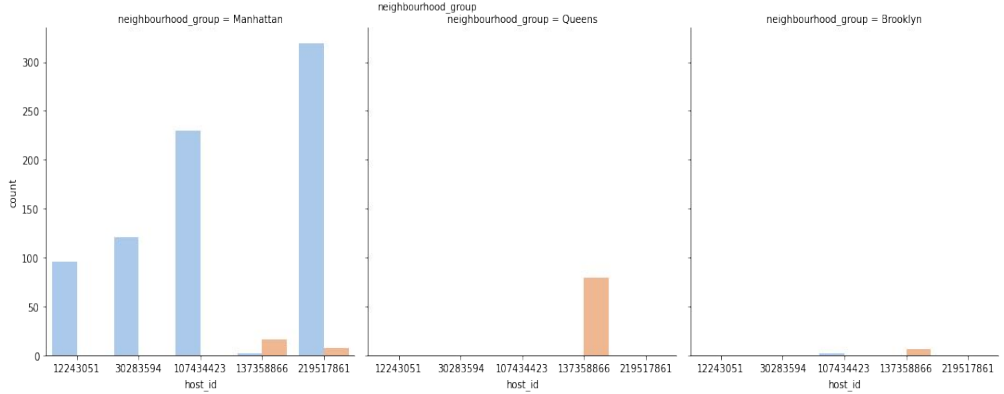
Data Exploration - Geo Dist & Heatmap





Data Exploration - Neighbourhood & Property Type



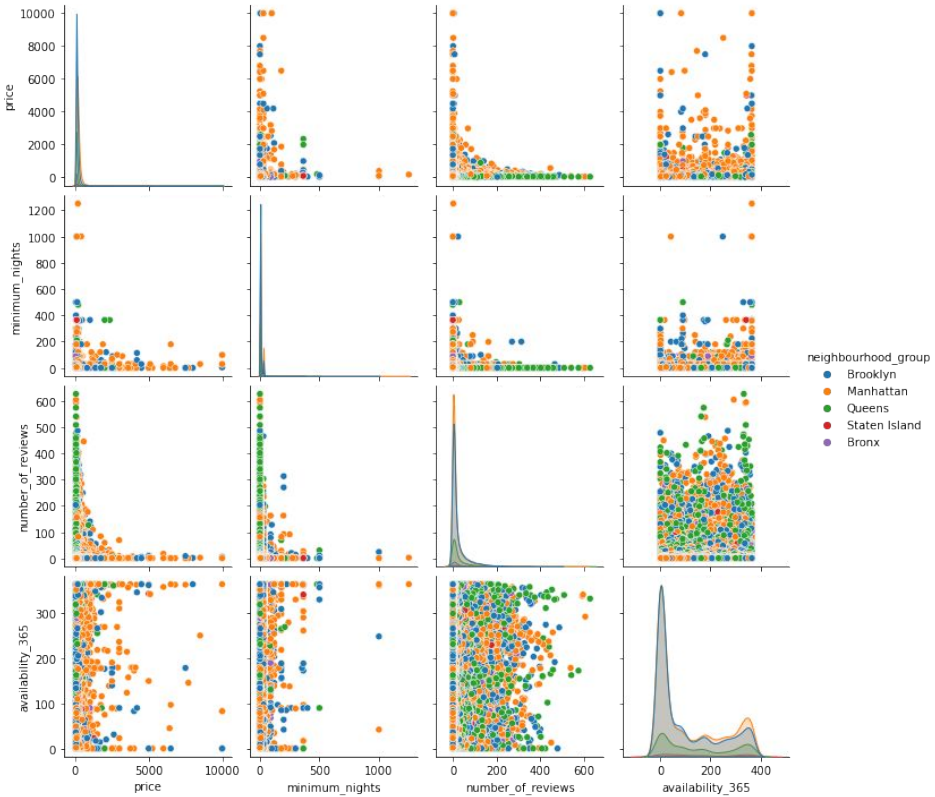


Contd..

- Entire Home/Apartment is listed more in Manhattan and Brooklyn.
- Bronx, Staten Island and Queens have significantly lower listings in comparison to Brooklyn and Manhattan.
- Top 5 Hosts seem to all have predominantly Entire Home room type with one exception and most are in Manhattan.

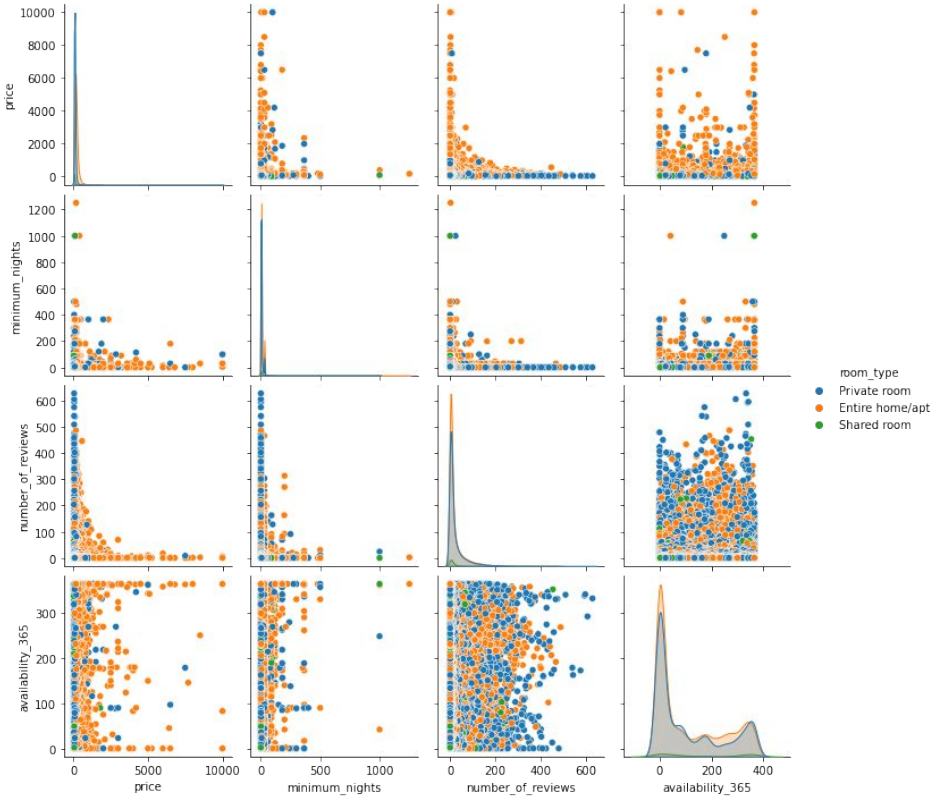
Data Exploration - Variable Distribution and Correlation

Contd..



- All of the variables in the scatter matrix histogram appear highly skewed.
- Manhattan and Brooklyn properties seem to be the most expensive.

Data Exploration - Variable Distribution and Correlation



Contd..

- Most high priced properties are Entire home/Apartment and their distribution is dense compared to private rooms.
- Reviews seems to be very popular in Airbnb.
- Only very few are available for booking beyond 100 days.
- The low minimum nights criteria seem to be popular as most listings fall in this criteria.

Data Preparation for Modeling

- **Dropping Attributes:**

  Name, ID, Host_Name, Last_reviews were not relevant for the modeling.

- **Cleaning:**

  Reviews_per_month had missing values. The numbers were very few and therefore they were replaced with zero.

  LabelEncoder method was used to transform categorical variables to represent in a numerical format. Encoded attributes were Neighbourhood_group, neighbourhood, room_type

- **Transformation:**

  Most of the input variables and the target variable Price were highly skewed. Therefore, a $\log_2$ transformation was applied to all the variables before passing to the models.

Modeling Techniques

Target Variable: ***Price*** (Numeric -Continuous)

Models Applied in this Project:

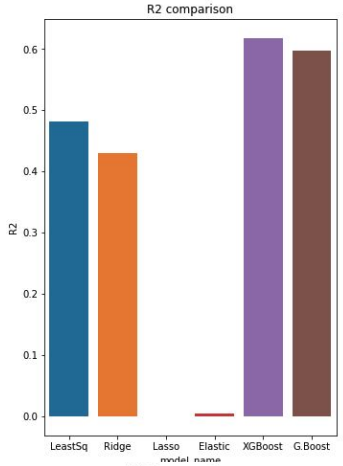
1. Linear Regression
2. Regularized Linear Models
    a. Ridge
    b. Lasso
    c. Elastic Net
3. Gradient Learner Models
    a. Gradient Boost (GradientBoostingRegressor)
    b. Extreme Gradient Boost (XGBRegressor)

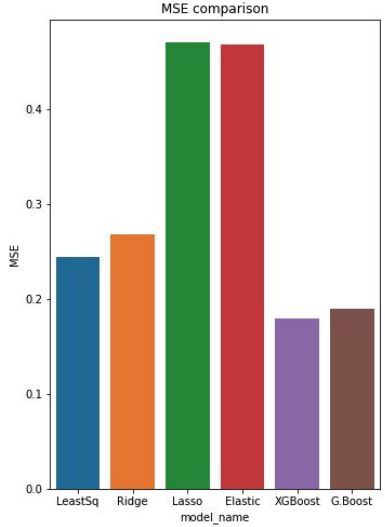
Evaluation Criteria and Results

Evaluation criteria :

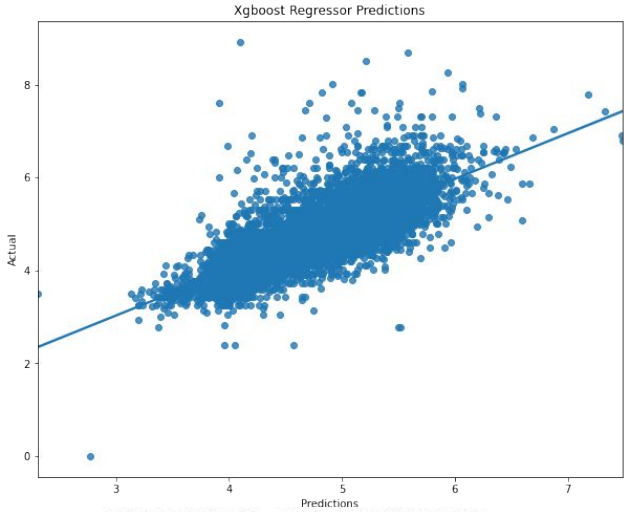
A high adjusted R^2 value and a low MSE is considered the criteria for evaluating performance.

	MAE	MAPE	MSE	R2	max_error	model_name
0	0.359502	1.873290e+12	0.243934	0.482056	4.688803	LeastSq
1	0.383310	1.898115e+12	0.268375	0.430161	4.630931	Ridge
2	0.545706	2.182146e+12	0.471014	-0.000100	4.738256	Lasso
3	0.544089	2.185378e+12	0.468545	0.005144	4.745274	Elastic
4	0.303667	1.274781e+12	0.179878	0.618067	4.825093	XGBoost
5	0.312292	1.788277e+12	0.189264	0.598138	4.790546	G.Boost



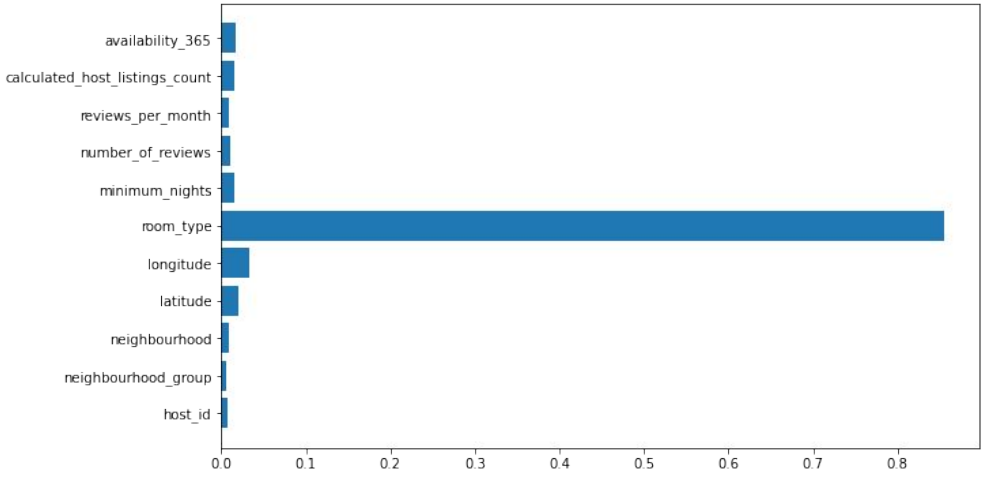


Results - XGBoost Metrics



	Actual Values	Predicted Values
0	63.732680	27.321270
1	30.883404	29.644995
2	38.803494	30.312424
3	27.776454	20.496761
4	22.449888	18.500448

Feature Importance Graph



Conclusion

- XGBoost Modeling Technique produced the best results.
  - The **Room type** was identified as the most important feature in predicting the price. The room types comprise of ‘Private Room’, ‘Entire Home/Apartment’ Or ‘Shared Room’.
  - The **Longitude** (geographical location of the property) was the second most feature that influences the price decisions.
- Price Prediction
  - The Model under predicted the prices in comparison to the actual values indicating that the optimum prices are expected to lower.
- Neighbourhood Group
  - Prices of New York City listings highly vary for each neighbourhood groups. The ones that are tourist destinations and business locations in and around Manhattan and Brooklyn are priced high. The average predicted price around these neighborhoods is $129.

Future Work

- From the modeling techniques perspective, RandomForest, Neural Networks and Deep learning methods can be tried if more time and resources are available.
- Also, the Airbnb Open Data has other attributes in the listings data set such as Amenities, customer Ratings.
    - These additional features could have a different result with the modeling process providing us more insight towards predicting the price.
- The reviews dataset have details about the review texts, feedback and interactions between owner and guests.
    - **Text mining of Reviews data set** and Time series analysis could further help understand the performance of the different Airbnb listings, customer sentiments and the predicted growth.
- The current dataset used pertains to **2019** values and it was purposefully decided not to pick a new/recent dataset as the **Pandemic situation has changed the travel/rental dynamics.** A lot of exceptions or **weights to be adjusted** for the period of 2020 - 2021 to compute those. This can be also a future work attempt.

Appendix: Duplication Vs My Work

Duplication/Original Work:

- Few Data Exploration bar plots. (Neighbourhood Distribution and Top 5 Hosts)
- LabelEncoder and Dropping Attributes
- Linear and Gradient Boosting Models
- Original work used Azure databricks and PySpark modules. My work does not include it.

My Work/Enhancements:

- Original model did not perform Log Transformation. I applied Log2 before modeling.
- Pairplots, Geographical distribution, heatmap.
- **Entire Model section was reworked** and regularized regression models were included and Gradient Models were reworked.
- Avg Vs Prediction Plots were created and log transformation was reversed to get the predicted price values.
- **Complete Results and Evaluation metrics** arrival is new as the original paper work cannot be applied. (Includes Results DF, Feature importance plot and R^2 and MSE comparison between models.

References

- Brahmaih, Kala. "Predicting Airbnb Listing Price Across New York". *Researchgate.net*. (2020).
- Rezazadeh Kalehbasti, Pouya et al. "Airbnb Price Prediction Using Machine Learning and Sentiment Analysis". *Machine Learning and Knowledge Extraction*. (2021): 173–184.
- duygut. “Airbnb NYC Price Prediction | Kaggle.” *Kaggle: Your Machine Learning and Data Science Community*, Kaggle, 22 Oct. 2019, https://www.kaggle.com/duygut/airbnb-nyc-price-prediction.
- Dan Wang and Juan L Nicolau. Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb. com. International Journal of Hospitality Management, 62:120–131, 2017