An Exploration of AirBnB's New York City Listings Through the Use of Data Mining Tools

Cognetti, John Carlon
College of Engineering and Computing
George Mason University
Arlington, Virginia, United States
jcognett@gmu.edu

Jayaswal, Aayushi
College of Engineering and Computing
George Mason University
Fairfax, Virginia, United States
ajayaswa@gmu.edu

Palaparthi, Phanindra
College of Engineering and Computing
George Mason University
Fairfax, Virginia, United States
dpalapar@gmu.edu

***Abstract*— Utilizing data from Inside Airbnb, this research focuses on four key aspects: predicting property prices based on listing details and customer reviews, understanding popular amenities and their influence on booking rates, exploring the impact of property availability on reviews, and analyzing price differences across various neighborhoods. Through exploratory data analysis, feature engineering, and the application of machine learning models including Lasso regression, Random Forest classifier, and K-Means clustering, this study aims to shed light on the dynamics of short-term rentals in New York City. The findings suggest correlations between listing features and rental prices, amenities and booking rates, availability and review frequency, and price disparities among neighborhoods. These insights can assist hosts in optimizing their listings and pricing strategies in the evolving regulatory landscape.**

Keywords—Airbnb, New York City, Pricing Prediction, Amenities, Short Term Rental, Booking Rates, Property Availability, Data Analysis, Machine Learning, Lasso regression, Random Forest classifier, K-Means clustering.

I. Introduction

In September of 2023, New York City enacted the controversial Local Law 18 [1], banning short term rentals in the city. A year later, and it the city has introduced a new bill [2] to walk back these restrictions. It is in this vein that a look back into the historical Airbnb listings of New York City to uncover how residents can successfully navigate the reintroduction of Airbnb to the city. Data for this project is provided by Inside Airbnb [3], a data focused and self-organizing advocacy group with the mission to provide data and advocacy on Airbnb's impact to residential communities.

This research aims to provide insights into four key aspects of Airbnb rentals. First it seeks to predict the price of a property based on its listing details and customer reviews. Second, it aims to understand the popular amenities and their influence on booking rates. Third, it explores the impact of a property's availability on its reviews. Lastly, it examines the price differences of Airbnb listings across various neighborhoods in New York City.

II. Methodology

A. Exploratory Data Analysis (EDA)

As the Inside Airbnb data set is first and foremost a website data scrape, there are a number of null values that need to be handled. Of the 37,000 records, twenty-five percent were removed due to null data. Of the data 75 features, five are outside URL's, four are on the scraping source, fourteen are about the host, sixteen are calculated columns from the data already contained, three are dates such as the *last reviewed date*, and finally two are mostly null thus leaving thirty-four features of interest. Of the features of interest, twenty are numeric, fourteen are text, and the remaining two are *longitude* and *latitude*.

B. Predicting Price

Numeric columns were first extracted from the data and a correlation matrix was performed, resulting in Fig. 1. Price data was cleaned out outliers through the removal of rows outside one and half times the interquartile range (IQR).

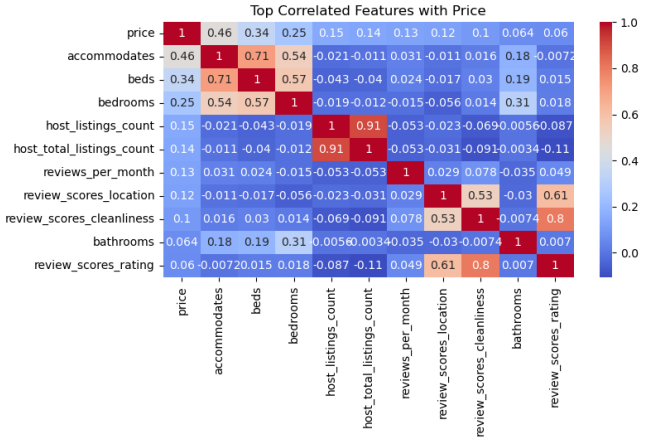

Fig. 1

Intuitively, Fig. 1 indicates that the highest correlated items with *price* are total *accommodation*, *beds*, and *bedrooms* as these can be an indication of the size of the property. Counterintuitively, the ratings fall to the lower end of this correlation and instead the host listings have a greater correlation. This could be indicative that hosts with multiple properties tend to *price* these listings around the same price.

Lasso regression [4] was chosen for this prediction both for its feature selection capabilities and for its stronger ability to handle collinearity. Collinearity is a concern as some features like *beds*, *bedrooms*, and *accommodations* are likely to be similar based on the property's size. Data was grouped by the *neighborhood group* (NYC Borough) and split into three train/test groups with varying percentages.

C. Amenities and Booking Rates

Amenities in the data set is a multi-valued feature, the first step in this analysis was to split this column into its individual features and transform this column into a feature matrix. Fig. 2 displays some of the most common amenities, of which Hot Water, Smoke Alarms, and a Kitchen were some of the most popular. *Chi-Square* [5] was used to find the correlation between the categorical variables within *amenities* with the *booking rate*.

Figure 2

Feature engineering of *luxury Amenities* and *essential amenities* was performed; where luxury listings included Pools, Hot Tubs, Gyms, Spas, and Fireplaces and those with Wi-Fi, Heating, Kitchens, Air Conditioning, and TVs were marked essential.

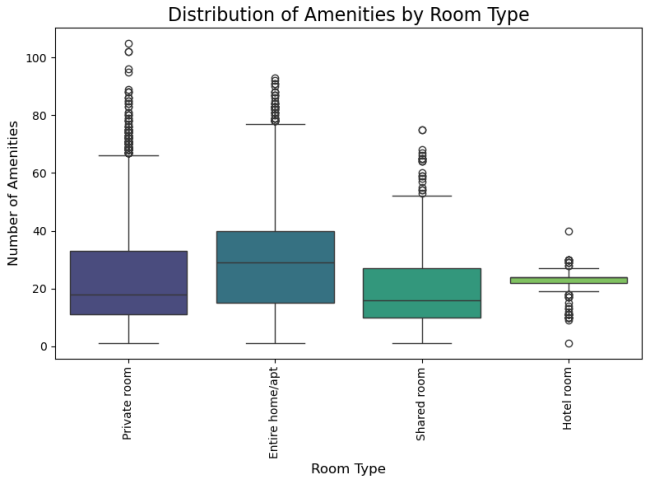

Figure 3

Amenities count and *room types* were also brought into this analysis, type was used to categorize the data by Private, Hotel, Shared, or Entire Home/Apartment. Intuitively, Fig. 3 displays, Private, Shared, and Entire Home/Apartments have drastically more amenities available as compared to Hotel Rooms.

High Booking Rate Calculation			
Table column subhead	***High Booking Count***	***Total Listings***	***High Booking Percentage***
Private Room	15285	19927	76.71
Entire Home/Apt	11803	16659	70.85
Shared Room	290	416	69.71
Hotel Room	261	539	48.42

Table 1

One-hot encoding was used on the *room types* feature to identify listings with high booking rates. Listings with an *availability thirty* of less than half were considered highly booked and marked with a binary variable. It was found that hotel rooms have high booking rate below fifty percent, as such this type was removed from the analysis.

Three types of modeling techniques, Random Forest classifier, gradient boosted [6], and logistic regression were chosen to determine the impact of amenities on the booking rates, then these models were repeated with the Synthetic Minority Over-sampling Technique (SMOTE) [7] to improve the class imbalances.

Consequentially, using SMOTE may introduce a bit of noise in the data. Hyperparameter tuning was considered but reduced the accuracy to an unacceptable level therefore we ignored this method.

D. Impact of Availability on Reviews

Two features, *availability 365* and *number of reviews* were selected to perform a logistic regression on listings with a high review frequency versus a low review frequency. A high review frequency is at least one review per month. Null values in this data were replaced by the global mean. A five-fold cross validation was performed on this model.

Additionally, a linear regression was performed on the reviews per month with the selected columns of *availability_365, number of reviews, review scores rating, price, minimum nights, accommodates, reviews per month.*

E. Price Differences of Location

To explore the differences in *price* by location, a K-Means clustering algorithm was used to predict the *neighborhood group* based on *price*. Null rows in either the *price* or the *neighborhood group* were excluded in this analysis. The data was scaled and a K value of 3 was used.

III. RESULTS

Results for these tests are separated out between the four research questions. *Price* predictions are in dollars, *Root Mean Squared Error (RMSE)* and R^2 values were used to compare the model's accuracy. *Mean Absolute Error (MAE)*, *Mean Squared Error (MSE)*, and R^2 were used in the evaluation of the linear model for availability and reviews. Additionally, Precision Recall Area Under the Curve (PR AUC) and Receiver Operating Characteristic (ROC) curve were used to compare the models of the *amenities* and booking rates. *Mean Silhouette Score* was used in the evaluation of the K-Means Clustering algorithm in price and location.

A. Predicting Price

Results for the price prediction varied quite drastically depending on the neighborhood group this method was applied to.

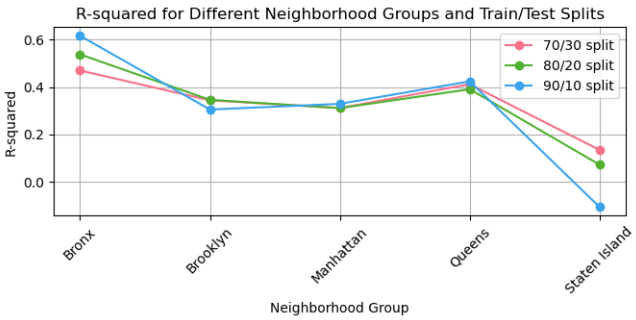

Figure 4

Fig. 4 shows that this model performed the best on the Bronx and Queens data with higher R^2 scores than the other boroughs. The model performed the worst on the Staten Island data, but of note is that the Staten Island data is the smallest number of listings. Brooklyn and Manhattan performed at the same level.

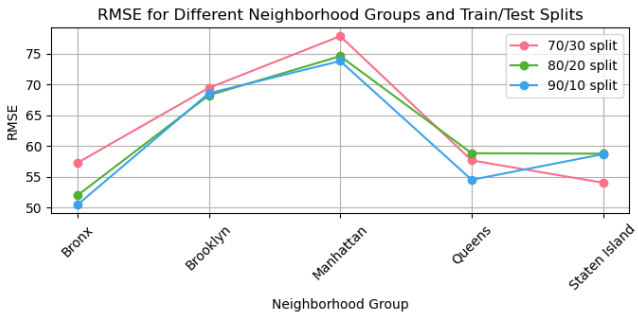

Figure 5

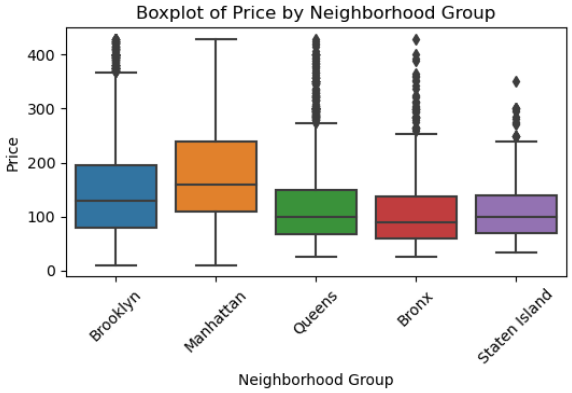

Figure 6

In Fig. 5 *RMSE* is lowest in the Bronx, Queens, and Staten Island tests. Brooklyn and Manhattan both see elevated *RMSE*, but per Fig. 6 we can verify that these groups inherently have both elevated average prices and a wider *IQR*.

Test Metrics for Lasso Regression Test		
Test	***R^2***	***RMSE***
70/30 Split Bronx	0.46965	57.27381
80/20 Split Bronx	0.53757	51.99078
90/10 Split Bronx	0.61609	50.46705
70/30 Split Brooklyn	0.34384	69.43010
80/20 Split Brooklyn	0.34587	68.23181
90/10 Split Brooklyn	0.30506	68.55839
70/30 Split Manhattan	0.31112	77.81968
80/20 Split Manhattan	0.31099	74.57897
90/10 Split Manhattan	0.32893	73.78977
70/30 Split Queens	0.41153	57.67738
80/20 Split Queens	0.39077	58.83934
90/10 Split Queens	0.42374	54.52961
70/30 Split Staten Island	0.13452	54.05576
80/20 Split Staten Island	0.07217	58.77083
90/10 Split Staten Island	-0.10630	58.69463

Table 2

B. Amenities and Booking Rates

The analysis demonstrates that basic amenities have a strong correlation with and impact on the booking rates. Features such as *amenities count* and *room type* also play a pivotal role, while luxury amenities exhibited weaker effects on the study.

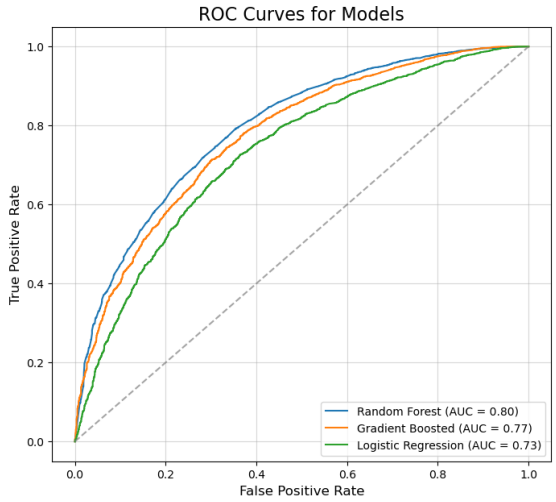

Figure 7

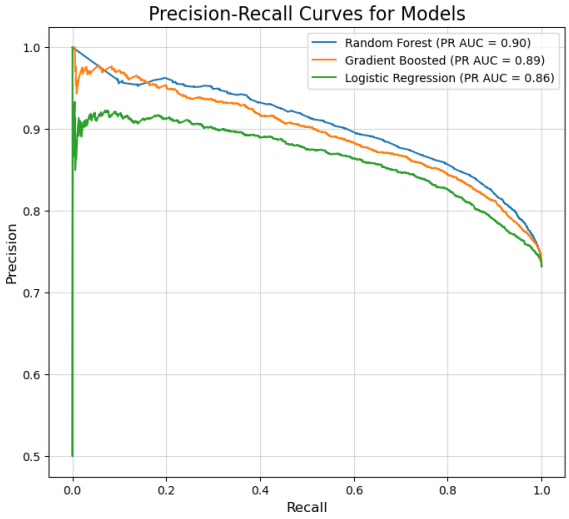

Figure 8

Random Forest Classifier Report				
Feature	***Precision***	***Recall***	***F1-Score***	***Support***
Low Booking Rate	0.70	0.33	0.45	2012
High Booking Rate	0.79	0.95	0.86	5497
Random Forest Classifier Model Metrics				
Accuracy		**0.78**		
Support		**7509**		

Table 3

Logistic Regression Report				
Feature	***Precision***	***Recall***	***F1-Score***	***Support***
Low Booking Rate	0.57	0.30	0.39	2012
High Booking Rate	0.78	0.92	0.84	5497
Logistic Regression Model Metrics				
Accuracy		**0.75**		

Logistic Regression Report				
Feature	***Precision***	***Recall***	***F1-Score***	***Support***
Support		**7509**		

Table 4

Gradient Boosted Report				
Feature	***Precision***	***Recall***	***F1-Score***	***Support***
Low Booking Rate	0.63	0.37	0.46	2012
High Booking Rate	0.80	0.92	0.86	5497
Gradient Boosted Model Metrics				
Accuracy		**0.77**		
Support		**7509**		

Table 5

Between the metrics for random forest, logistic regression, and gradient boosted, the random forest was the best performing model overall seen in Fig. 7 and 8 with **78.2%** accuracy, **0.90** *PR AUC*, and balanced precision-recall, and superior macro/weighted averages. This displayed robustness and generalizability. Logistic regression struggles in both scenarios due to linear assumptions, achieving **74.2%** accuracy. Limited ability to handle feature complexity. Metrics of the gradient boosted model in Table 5 show strong balance with **75%** accuracy, **0.89** *PR AUC*, and competitive recall.

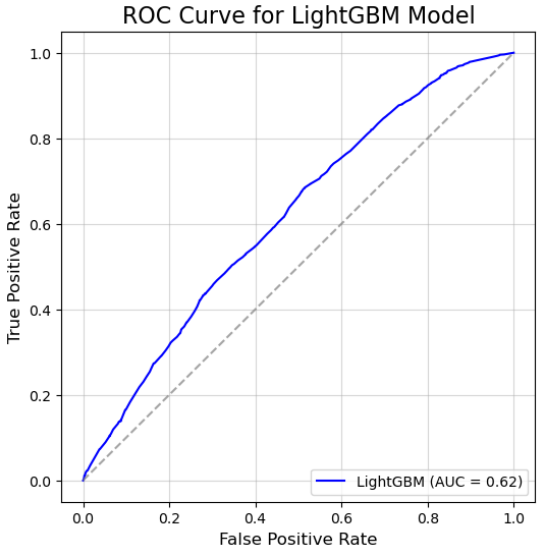

Figure 9

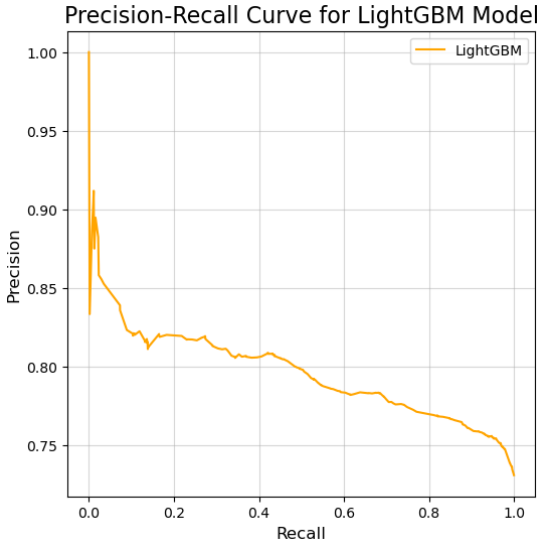

Figure 10

Light Gradient Boosted Classifier Report				
Feature	***Precision***	***Recall***	***F1-Score***	***Support***
Low Booking Rate	0.33	0.57	0.42	2021
High Booking Rate	0.79	0.57	0.66	5488
Light Gradient Boosted Model Metrics				
Accuracy		**0.57**		
Support		**7509**		

Table 6

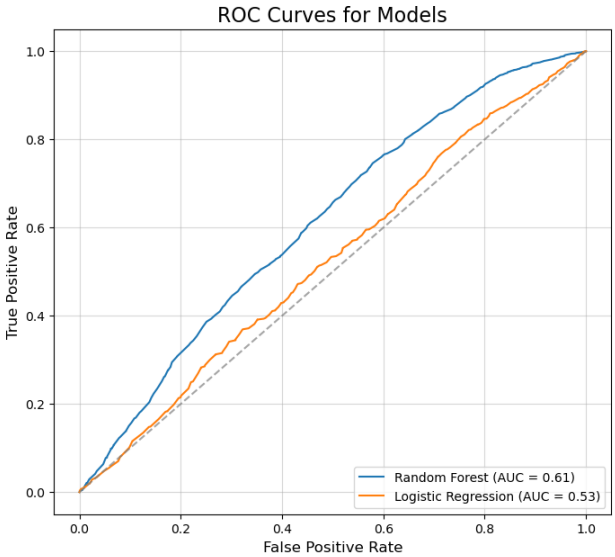

Figure 11

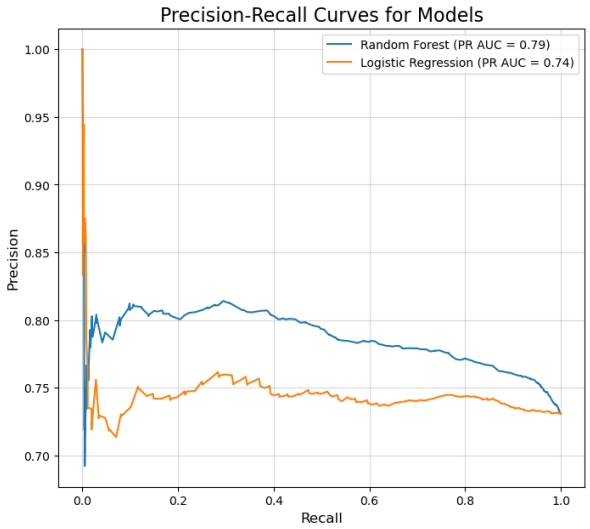

Figure 12

SMOTE Random Forest Classifier Report				
Feature	***Precision***	***Recall***	***F1-Score***	***Support***
Low Booking Rate	0.34	0.52	0.41	2021
High Booking Rate	0.78	0.64	0.70	5488
Random Forest Classifier Model Metrics				
Accuracy		**0.60**		
Support		**7509**		

Table 7

SMOTE Logistic Regression Report				
Feature	***Precision***	***Recall***	***F1-Score***	***Support***
Low Booking Rate	0.33	0.50	0.40	2021
High Booking Rate	0.77	0.63	0.69	5488
SMOTE Logistic Regression Model Metrics				
Accuracy		**0.59**		
Support		**7509**		

Table 8

As noted, these tests were repeated with the use of the SMOTE technique. SMOTE effectively balanced the class imbalance thus enhancing the minority class predictions. This did not induce overfitting into the model. This technique was useful for logistic regression where the precision and recall improved for the minority class seen in Table 8. Logistic regression outperforms LightGBM in terms of overall accuracy, recall, and F1-score for the majority class, making it the better choice in this case. However, LightGBM has a slight edge in identifying minority class instances, as seen in its better recall and F1-score for the Low Booking Rate class. Based on the *ROC* and precision curve, LightGBM is the preferred model for maximizing both *AUC* metrics, making it ideal for imbalanced datasets where Recall (minimizing false negatives) is important. Excels in *PR AUC* (**0.80**), outperforming others in precision-recall tradeoff but accuracy reduces to **57.4%**. Random Forest is also a strong alternative, particularly if consistent Precision is required at higher Recall values. In this case, *PR AUC* remains strong (**0.79**), but accuracy drops to **60.4%** due to noise from synthetic samples. Logistic regression should be avoided for this dataset due to its inability to handle the complexity of the features effectively.

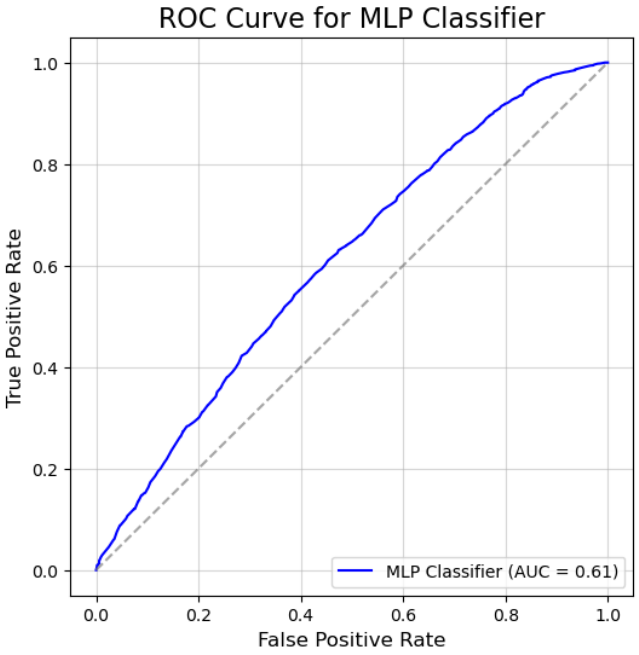

Figure 13

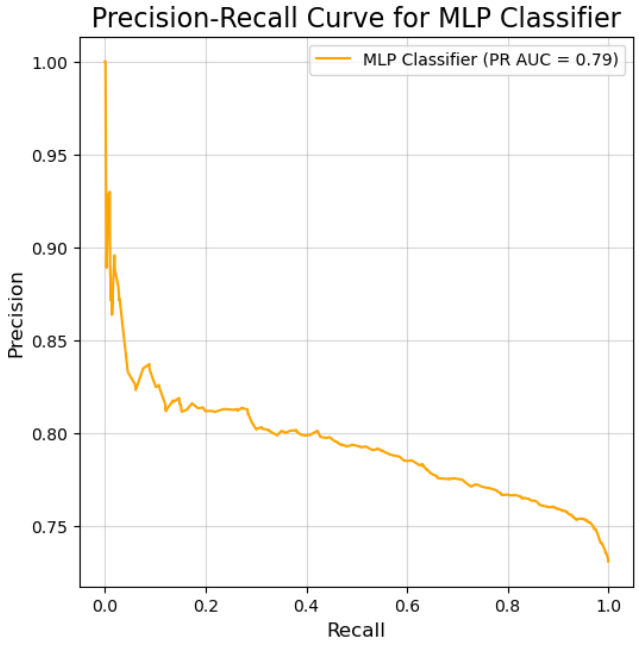

Figure 14

Neural Network Multilayer Perceptron Report				
Feature	***Precision***	***Recall***	***F1-Score***	***Support***
Low Booking Rate	0.33	0.60	0.43	2021
High Booking Rate	0.77	0.55	0.65	5488
MLP Model Metrics				

Neural Network Multilayer Perceptron Report				
Feature	*Precision*	*Recall*	*F1-Score*	*Support*
Accuracy		**0.56**		
Support		**7509**		

Table 9

Results for the Multilayer Perceptron classifier did not give satisfactory results as seen in Fig. 13 and 14 as other models outperformed this algorithm.

C. Impact of Availability on Reviews

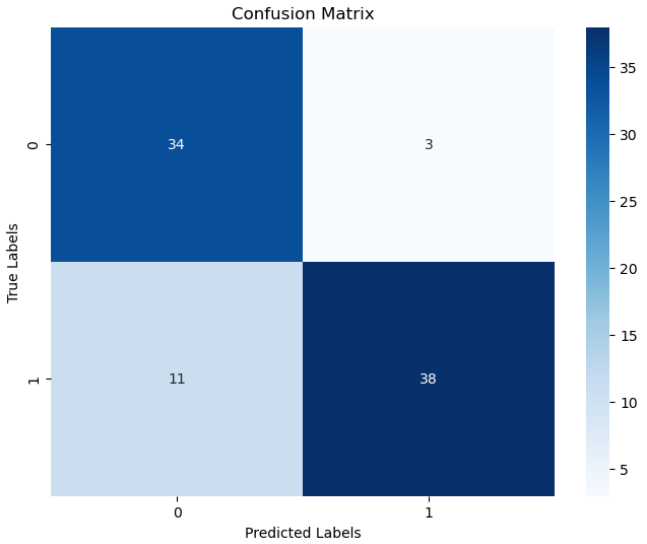

Figure 15

Five-Fold Cross Validation Results						
Test	*T1*	*T2*	*T3*	*T4*	*T5*	*Mean*
Accuracy Score	0.860	0.894	0.812	0.812	0.753	0.826

Table 10

Fig. 15 displays the confusion matrix for the logistic regression. The classification report showed a precision of **0.76** for predicting low-frequency reviews and **0.93** for predicting high-frequency reviews, with an overall F1-score of **0.84**. Across five-fold cross validation results in Table 10 the model had an accuracy score of **0.83**. Based on these results, the model can accurately predict if a property will have a high review frequency.

D. Price Differences of Location

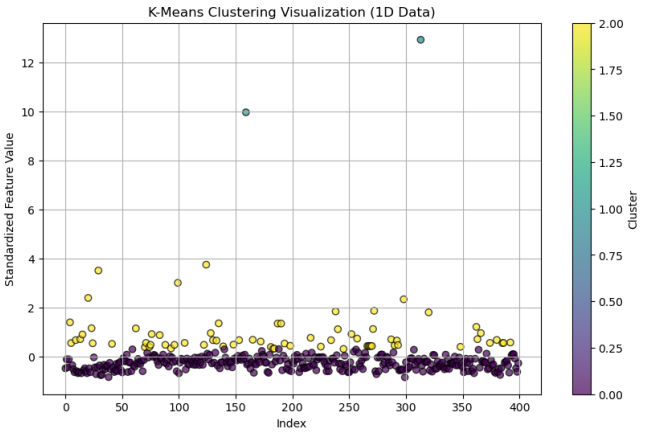

Figure 16

The *Mean Silhouette Score* yielded a value of **0.69**, indicating a reasonable separation of clusters. Fig. 16 shows distinct clusters of properties based on price, with most properties concentrated in the lower price range and a few outliers representing higher-priced properties. The color-coded clusters highlight how certain neighborhoods share similar pricing characteristics, providing insights into regional price trends. This analysis helps identify areas where properties are similarly priced, assisting hosts in setting competitive prices and helping guests make informed decisions.

IV. Bias, Conclusions, & Further Research

A. Data Biases

This work began by seeing a level of selection, reporting, and omission biases within an Airbnb data set initially found on Kaggle [8]. To remove these concerns from our analysis, the referenced source from Inside Airbnb website for New York City was used. This data set is a more accurate representation of the listings made by Airbnb containing web scrape sources. This data source reduced bias, but some bias was still observed, such as the high booking rate samples significantly outnumbered the low booking rate samples leading to class imbalance. This impacted the model performance while looking into how the amenities affect the booking rate. The lack of adequate presence of the listings with low booking rate made it difficult to find out the reasons for the underperformance of the low booking rate listings. This led to reduced correlation between the variables and poor performance of a few models. This also pushed us into using the SMOTE technique to address the skewed distribution of data.

B. Conclusions

The performance of the models used in this study was fair, providing valuable insights into the factors influencing Airbnb rental prices, booking rates, and review frequency in New York City. The lasso regression model effectively handled collinearity and identified key predictors of property prices, although its performance varied significantly across different boroughs. The random forest classifier and gradient boosted models demonstrated strong capabilities in predicting booking rates based on amenities, while logistic regression and linear regression models offered valuable perspectives on the impact of availability on reviews.

C. Future Research

Models discussed in this paper could be improved in a variety of ways, including data quality, feature engineering, model selection, tuning, and incorporating user feedback.

1) Data Quality: If regulations are relaxed, the influx of new short-term rentals could drastically increase from the approximately 35,000 records analyzed in this report. More data points could improve the models.

2) Feature Engineering: Developing additional features, such as interaction terms between amenities and neighborhood variables, could improve model performance. Further exploration of temporal patterns, like seasonality effects on bookings and prices, would add valuable insights.

3) Model Selection and Tuning: Experimenting with a wider range of machine learning algorithms and conducting extensive hyperparameter tuning could lead even greater performing models. Further use of cross-validation and

implementing grid search could optimize the models parameters.

4) User Feedback: Since Airbnb prices are user-set, incorporating user feedback on predicted prices (perhaps by presenting them as "recommended prices") could enhance the model's accuracy and usability.

REFERENCES

[1] A. Hoover, "New York cracked down on Airbnb one year ago. NYC housing is still a mess," Wired, https://www.wired.com/story/new-york-city-airbnb-law-one-year-results/ (accessed Dec. 8, 2024).

[2] A. Pohle, "New York Made Airbnbs Harder to Find. Now It's Reconsidering.," The Wall Street Journal, https://www.wsj.com/lifestyle/travel/new-york-city-airbnb-crackdown-bill-1caa0bc5 (accessed Dec. 8, 2024).

[3] M. Cox, J. Morris, and T. Higgins, "New York City, New York, United States," Sep. 4, 2024, https://insideairbnb.com/get-the-data/ (accessed Sept. 28, 2024)

[4] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html

[5] SciPy v1.14.1, "chi2_contingency," https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html, Accessed: Dec. 11, 2024.

[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, Aug. 2016, pp. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[7] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017. [Online]. Available: https://jmlr.org/papers/v18/16-365.html

[8] A. Azmoudeh, "Airbnb Open Data: New York Airbnb Open Data." Feb. 28, 2022, https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata/data (accessed Sept. 3, 2024)

APPENDIX

As discussed in the biases section, this project began with [8] where though EDA it was found that this data was likely manipulated and contained strong biases. The following figures 17-21 describe and discuss the issues found in [8].

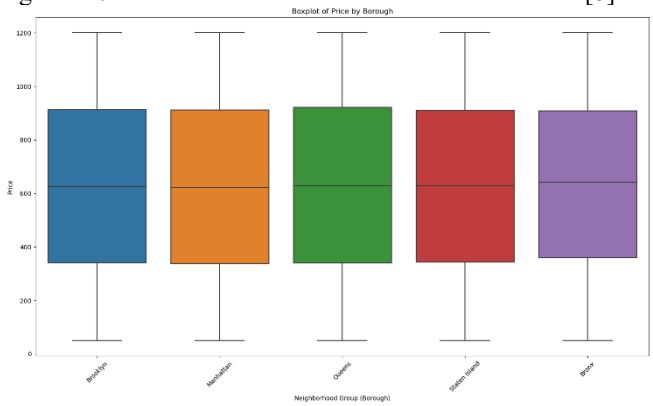

Figure 17

Fig. 17 first shows the impressively consistent price data across the five boroughs, the initial mark of manipulated data.

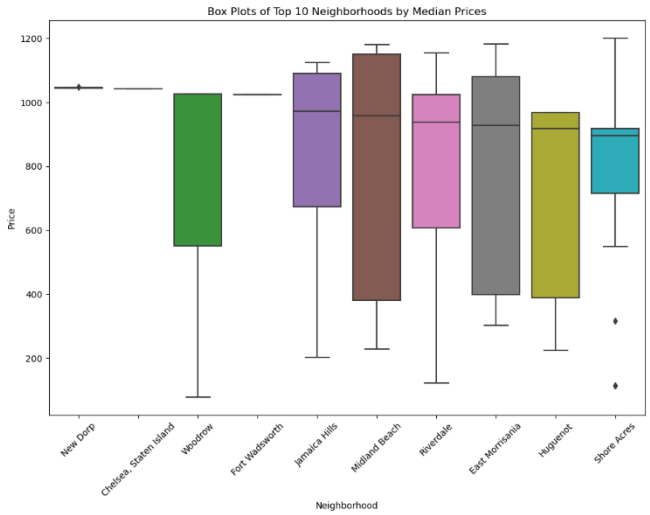

Figure 18

Unlike the boroughs, this behavior was not seen in the distinct neighborhoods of Fig. 18.

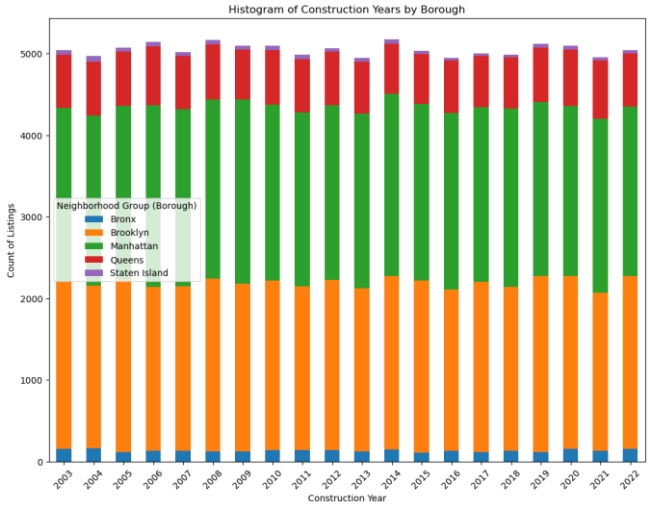

Figure 19

Counting the listings by borough made intuitive sense, Manhattan and Brooklyn are two popular destinations and as such would have more listings. However, seeing approximately 5,000 listings for every construction year in Fig. 19 was another indication that the data had been manipulated.

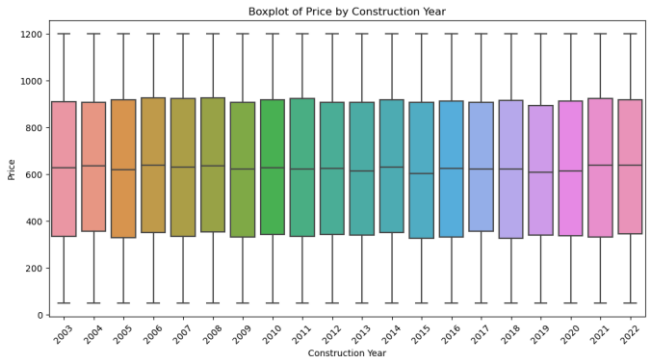

Figure 20

Again, boxplots of the price by construction year showed an uncanny consistency.

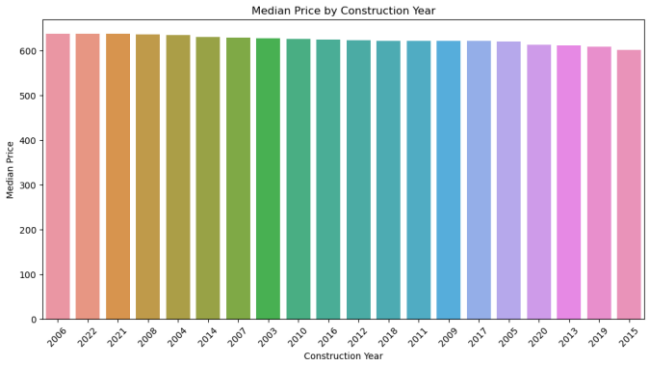

Figure 21

Fig. 21 displays the median price of all construction years with nearly all years about $600. This is another example where the Kaggle data set did not fit common intuition. Newer properties should tend to have higher costs associated.

With a renewed focus on [3], a lasso regression and a backwards selected linear model were fitted on the entire data set regardless of neighborhood group.

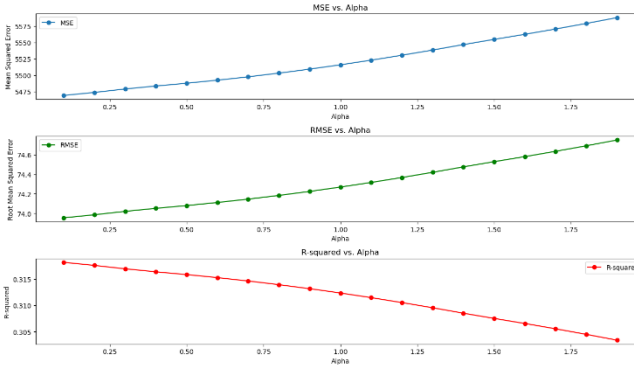

Figure 22

Fig. 22 Shows the performance of the lasso model over a range of twenty alpha values. The best performance was seen with an alpha of 0.1, *RMSE* of **80.73** and an R^2 of **0.214**.

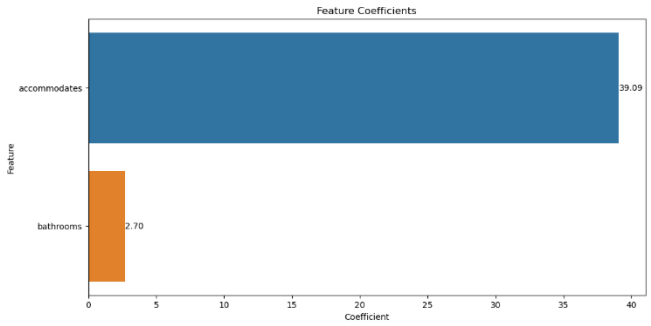

Figure 23

The fitted linear model with backwards selection picked out accommodates and bathrooms as its two features. This model performed marginally worse than the lasso, with an *RMSE* of **80.77** and an R^2 of **0.21**.

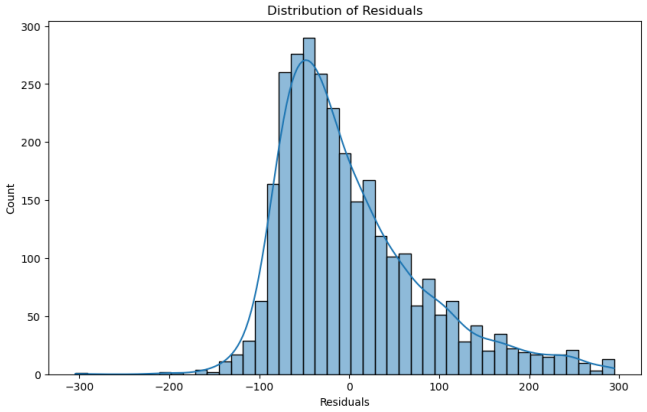

Figure 24

Residuals in Fig. 24 of the linear model showed a normal distribution with a slight left skew, indicating the model does have a tendency to underpredict.