2023

Predicting Airbnb Prices in San Diego

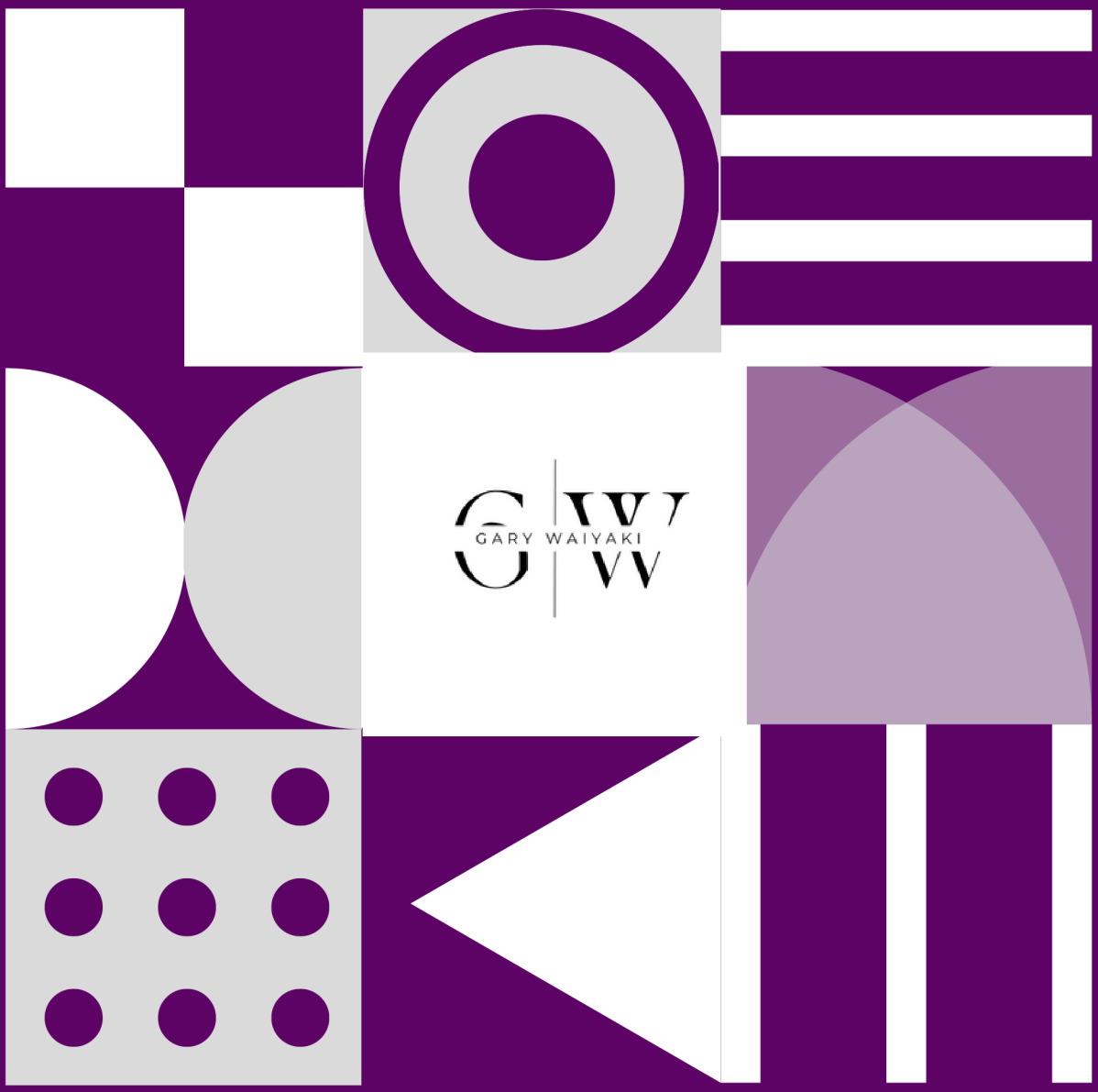
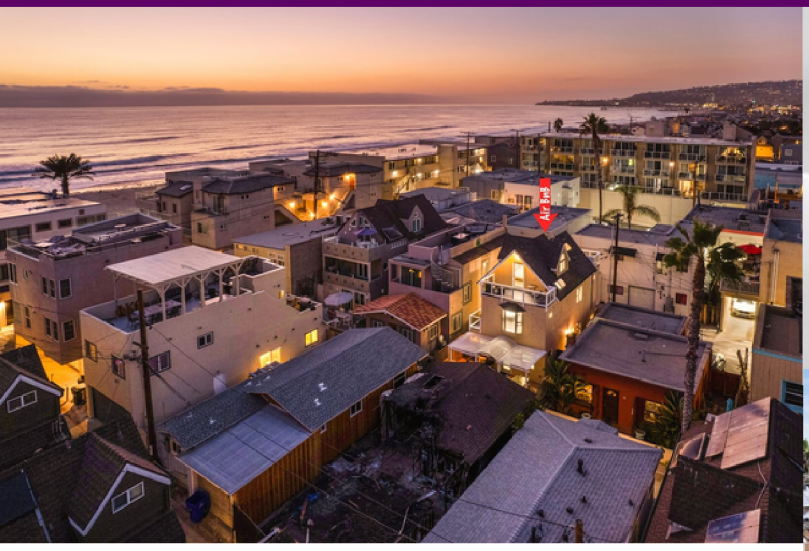
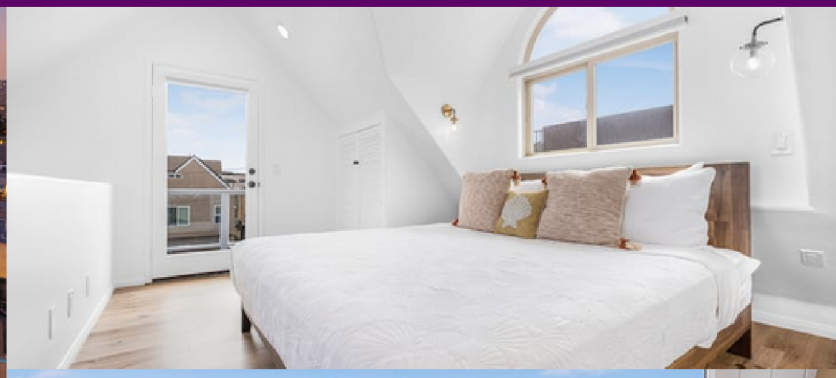
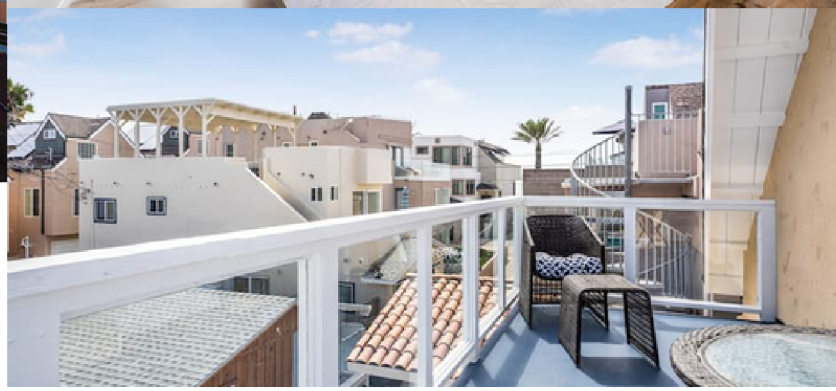


Table of Contents

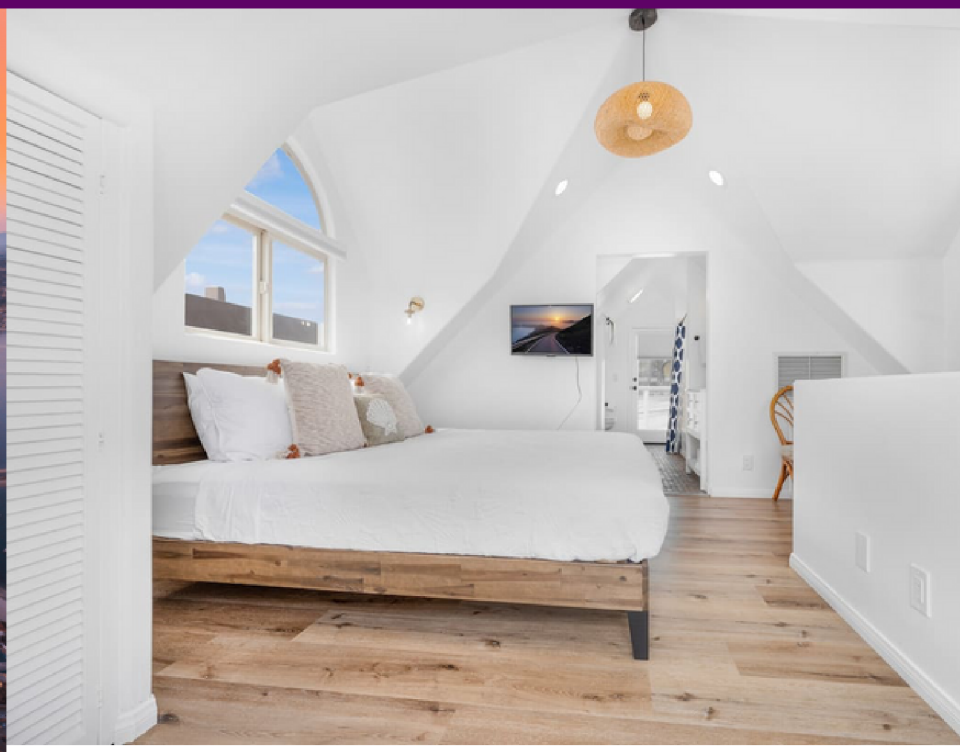






Introduction

Airbnb has become a popular alternative to traditional hotel accommodations. With over 7 million listings worldwide, it is disruptive to the hotel industry. However, the pricing of Airbnb listings varies significantly based on a plethora of factors such as location, amenities, availability, etc. Predicting the price of Airbnb listings will help hosts to price their properties competitively and guests to choose the best options within their budget.



Data

Our dataset is publicly accessible from Inside AirBnB.

- **listings.csv:** Summary information and metrics for listings in San Diego (good for visualizations). The file gives insights about a host, location, review rating score, room, and property type.
- **calendar.csv:** Summary information about availability and price per day in 2022-2023 years. Based on 'listing_id' this file will be merged with listings.csv.
- **reviews.csv:** Summary review data for the listings. This dataset won't be used in this analysis.

Methodology

We chose the San Diego dataset and will take the following steps to make conclusions about what factors affect price listings:

- First Look at the Data:
  - What information is present? What information is missing? Discover general facts such as host listings count.
- Initial Data Preparation:
  - Remove irrelevant information, reformat the data, and input missing values.
- Find high level trends and correlations.
- Use Machine Learning for further analysis:
  - Encode categorical features correctly.
  - Train more than one model and look for feature importance and model prediction.

Data Wrangling

Overview of Data Cleaning Issues and Solutions

1. **Fix and reformat data types**
The amenities column was formatted as a long string separated by commas.

- **Solution part 1**: Use ***.str.split(“,”)*** to split the values of the amenities column by commas, returning a pandas Series of lists of amenities by row.
- **Solution part 2**: Create a new feature with the number of listed amenities in each row of the amenities column.
- **Solution part 3:** Create a new column/feature “total_amentities” that contains the number of amenities for each row in the original amenities column.

2. Fix and reformat the date column

- **Solution part 1**: Create a month column/feature. This column will contain the month as an integer of the corresponding data in each row in the date column.
  - Achieve this using ***.apply()*** method to apply lambda function on the date column.
  - ***Lambda function*** splits the date string on the ***“-”*** character and returns the second element on the resulting string (i.e. month).

3. Fix and reformat Price Feature and Target variable

- **Solution part 1**: Remove the dollar sign from the price column/feature but first convert the ‘object’ data type to string using ***.astype()***.
- **Solution part 2**: Use ***.str.replace(“$”, “ ”)*** to replace the ***“$”*** with a space ***“ ”***.
- **Solution part 3**: Remove commas from prices with place values in the thousands.

4. Drop columns with 70% missing values.

5. Replace missing values with the mean.

To clean the data, the above issues were identified and solutions were proposed. The amenities column was reformatted using the .str.split() method to create a new feature with the number of listed amenities in each row. The date column was reformatted by creating a month column/feature using the .apply() method and a lambda function. The price feature and target variable were reformatted by removing the dollar sign and commas from prices with place values in the thousands. Columns with 70% missing values were dropped, and missing values were replaced with the mean.

Exploratory Data Analysis - Part 1

In the EDA report, I was able to:

- Identify which months and neighborhoods had the highest prices
- Identify the factors that greatly influence the listing prices

Figure 1 on the right shows that San Diego prices vary depending on the time of year. The busiest months are July to September.

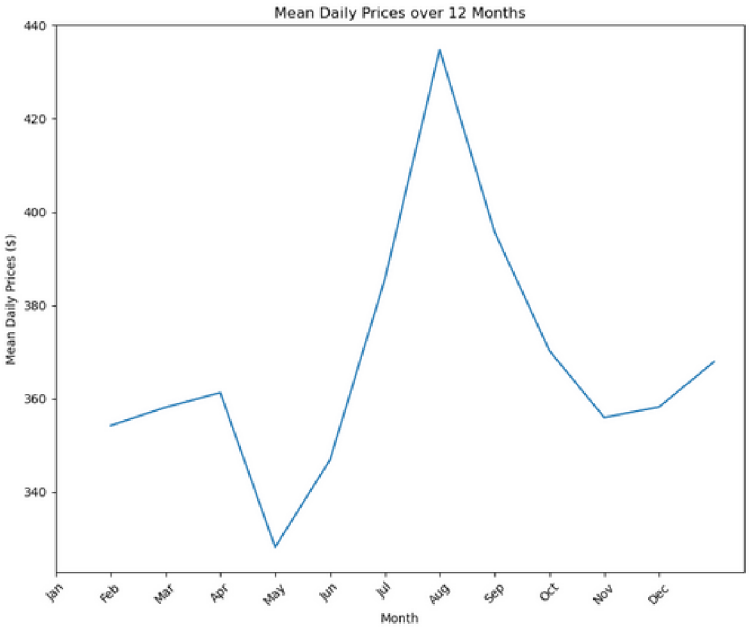


Figure 1: Mean Daily Prices over 12 Months

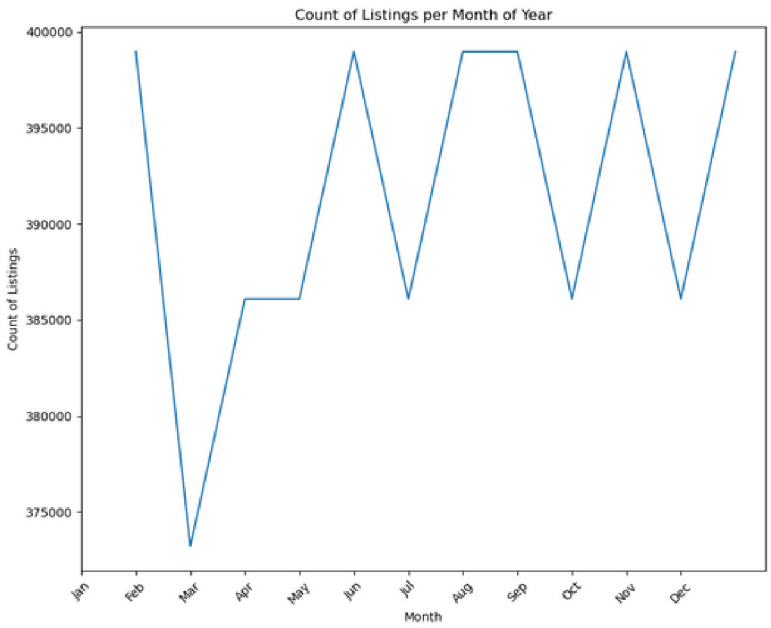


Figure 2: Number of Available Listings per Month of Year

Figure 2 on the left shows that the highest number of available listings is in August and September. However, high availability does not necessarily correlate with low rental prices. Other factors such as location and amenities may explain the high prices during these months.

Exploratory Data Analysis - Part 1 Continuation

In the EDA report, I was able to:

- Identify which months and neighborhoods had the highest prices.
- Identify the factors that greatly influence the listing prices

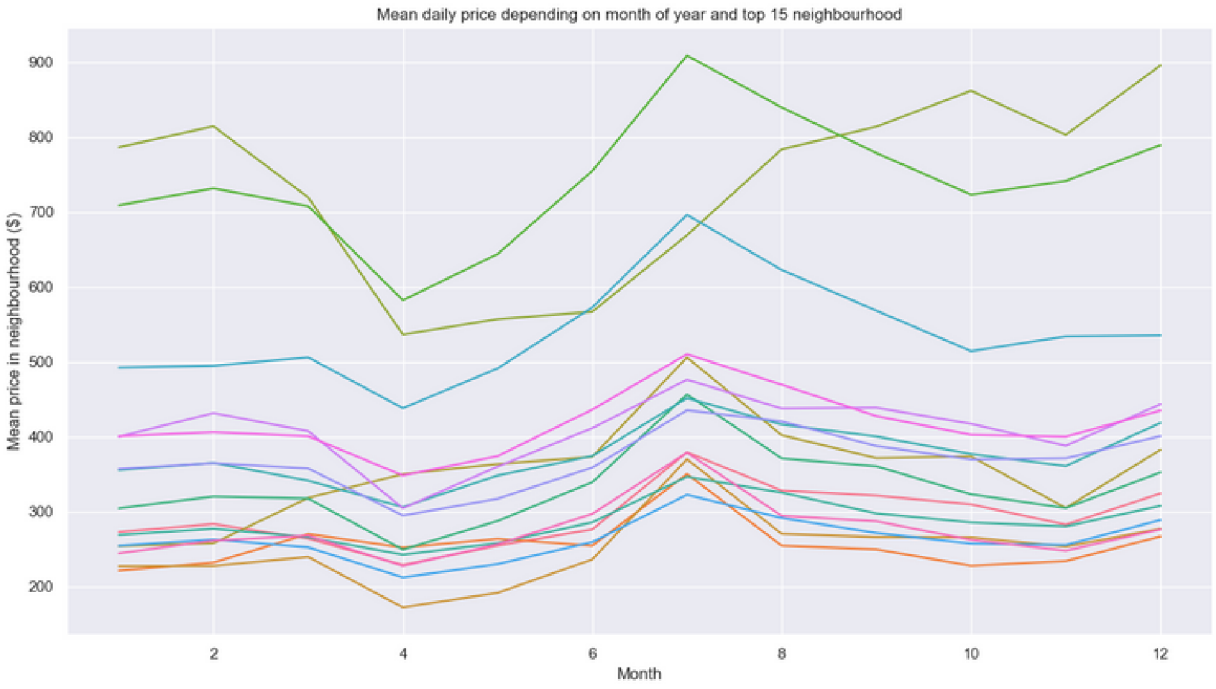


Figure 3: Mean Daily Price Depending on Month of Year and Neighborhood

Figure 3 shows the highest average prices charged for listings in the top 15 neighborhoods in San Diego. The neighborhoods with the highest rental prices are Gaslamp Quarter (~$850), La Jolla (~$850), Mission Bay (~$690), Old Town (~$650), LittleItaly ($600 - $780), and Pacific Beach ($500 - $600). This conclusion is supported by the fact that these neighborhoods also have a high number of listings. Prices rose sharply in July or between month 6 and month 8.

Exploratory Data Analysis - Part 2

In the EDA report, I was able to:

- Identify which months and neighborhoods had the highest prices.
- Identify the factors that greatly influence the listing prices

Figure 4: Correlation Matrix of Numerical Features

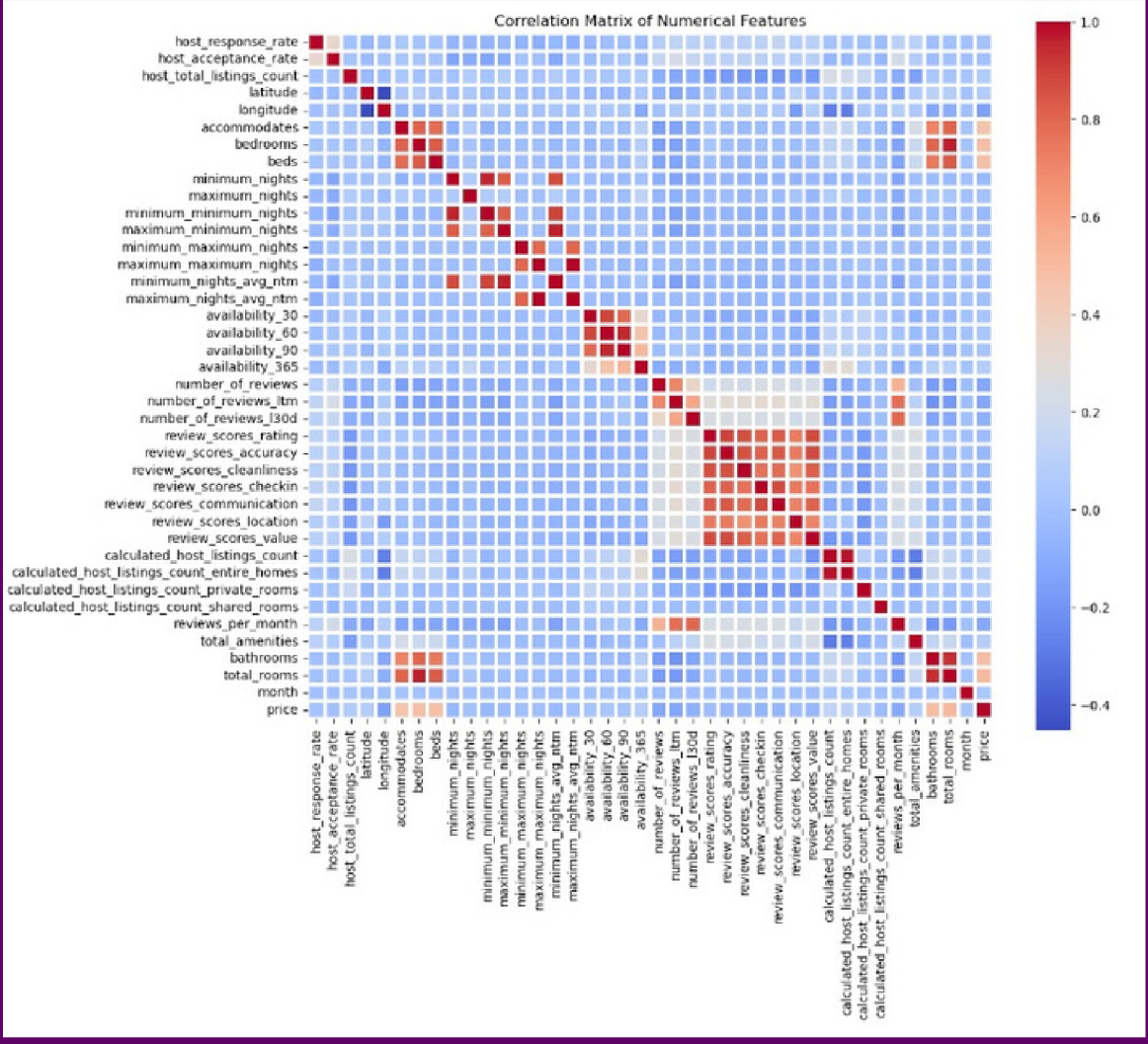


Figure 4 shows that the target variable(price) is highly correlated with the following features: accommodates, bedrooms, beds, bathrooms, total_rooms (sum of bathrooms and bedrooms).

Machine Learning Models

- We used 3 different models (Random Forest Regression, XGBoost Regression Model, and Light Gradient Boosting Machine) to determine the most important features.
- We applied cross-validation with a split in 5 folds to avoid overfitting. Additionally, we removed the least important features and some highly correlated features to avoid overfitting.
- We encoded the categorical variables using one-hot-encoding to improve performance and fitted and predicted using each of the three models.

Model Evaluation Results

Model	Cross-validation Score	R^2-score (Coefficient of Determination)
Random Forest	-8%	52%
LGBM	20%	49%
XGBoost	-15%	61%

Analyzing Model Evaluation Results

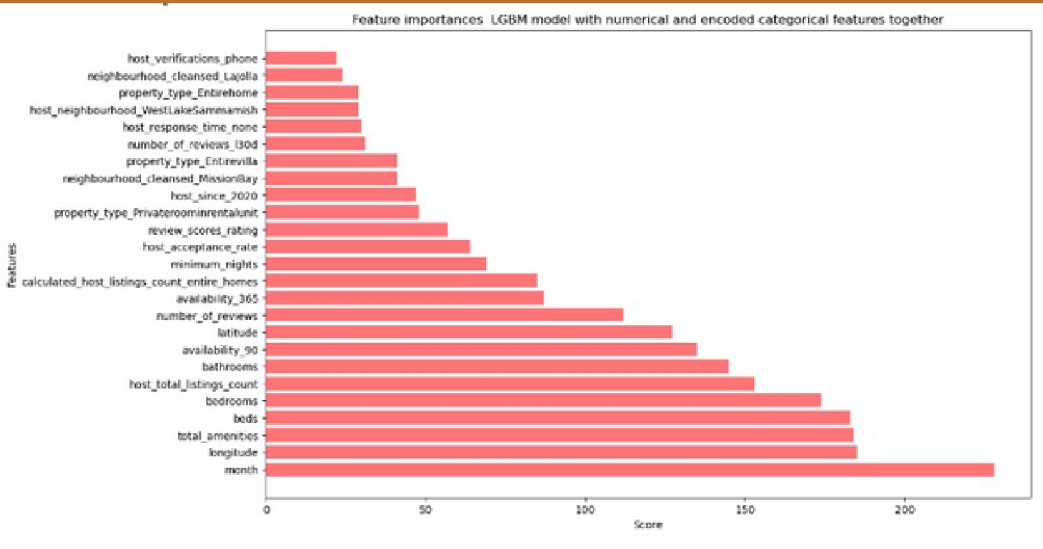
Model	Cross-validation Score	R^2 Score
Random Forest	-8% suggests poor model performance because it's not capturing the underlying patterns in the training data effectively and may be underfitting	52% suggests that the model explains about 52% of the variance in the training data
Light Gradient Boosting Machine	20% indicates that the model explains approximately 20% of the variance in the training data during cross-validation	The model captures about 49% of the variance in the training data. However, it is not exceptionally strong and there may be some unexplained variance
XGBoost Model	-15% suggests that the model maybe underfitting and not capturing the underlying patterns in the training data	61% implies that the model explains around 61% of the variance in the training data.

- **The cross-validation score -**assesses how well the model is likely to perform on unseen data.
- **The R^2 score** measures how well the model explains the variability in your target variable.
- High values for both **cross-validation** scores and ***R^2*** scores generally indicate a model that is performing well but be cautious about overfitting.

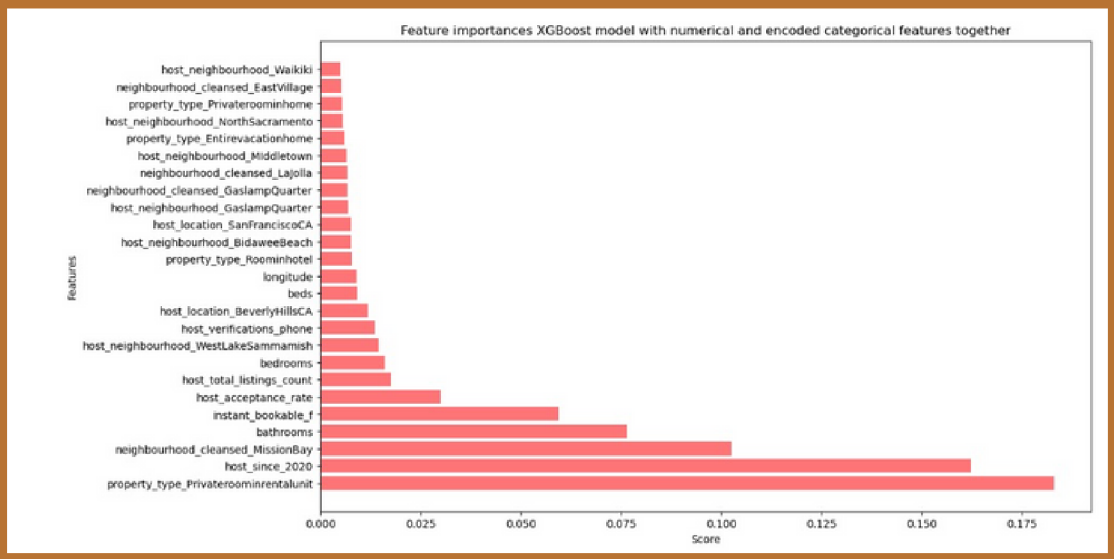
Predictions

Given that a higher R2 score is better for predictive performance, the XGBoost model would be the best-suited model. However, it should be noted that more tuning is required to address its low R-squared cross-validation score which suggests that it may not generalize well on unseen data. Solutions to this issue will be addressed on the next slide.

What we can see however, is that the LGBM model selects a broad set of features that are important: **month, longitude, total_amenities, beds, bedrooms, host_total_listings_count, bathrooms, availability_90 and number_of_reviews.**



In contrast, XGBoost model, the target variable(price) is dependent on a small set of dominant features: **property_type_Privateroominrentalunit, host_since_2020, neighborhood_cleansed_MissionBay, bathrooms, and instant_bookable_f.**



Next Steps

Given that the Random Forest and XGBoost models were not able to capture the underlying patterns in training data, more tuning of these models is required. This means revisiting feature selection and selecting only the import features to help reduce model complexity.

Note: the Light Gradient Boost Machine Model only explained 20% of the variance in the training data during validation, this shows that it too would benefit from more tuning.

Future Improvements:

01 | Feature Selection and Hyperparameter Tuning

- Use feature selection techniques to identify and retain the most relevant features while discarding irrelevant or redundant ones. e.g. PCA (Principal Component Analysis)
- Experiment with different hyperparameters for your model, such as learning rate, regularization strength, or tree depth, and use techniques like grid search or random search to find optimal settings.

02 | Feature Engineering and Cross-Validation

- Add more relevant features to your dataset that might help the model learn the underlying patterns.
- Collect more training data if possible. A larger dataset can help the model learn better representations and reduce underfitting.
- Use techniques like k-fold cross-validation to assess your model's performance on different subsets of the data. This can help you identify if underfitting is a consistent issue.

Conclusion

In this report, we looked at which features greatly influenced Airbnb price listings in San Diego in order to help hosts set competitive prices to attract customers.

Highlight 1

- Airbnb prices are highly dependent on the time of year and neighborhood. The highest prices are charged in the month of August.
- The most expensive neighborhoods are:
  - Gaslamp Quarter which has a prominent nightlife i.e. clubs, restaurants, theaters, etc.
  - La Jolla, Pacific Beach, and Mission Bay which are by the ocean side.
  - Old Town and Little Italy which have attractions such as boutiques, restaurants, hotels, Museums, etc.

Highlight 2

- The XGBoost model was the most robust predictive model but it was not able to capture the underlying patterns in the training data.
- This suggests that more tuning of these models is required. For Example: revisiting feature selection and hyperparameter tuning.