



# Predicting Airbnb Prices in San Diego

Welcome to the world of predicting Airbnb prices in San Diego! In this presentation, we will explore the importance of accurate price predictions, the data collected, and the methodology used.

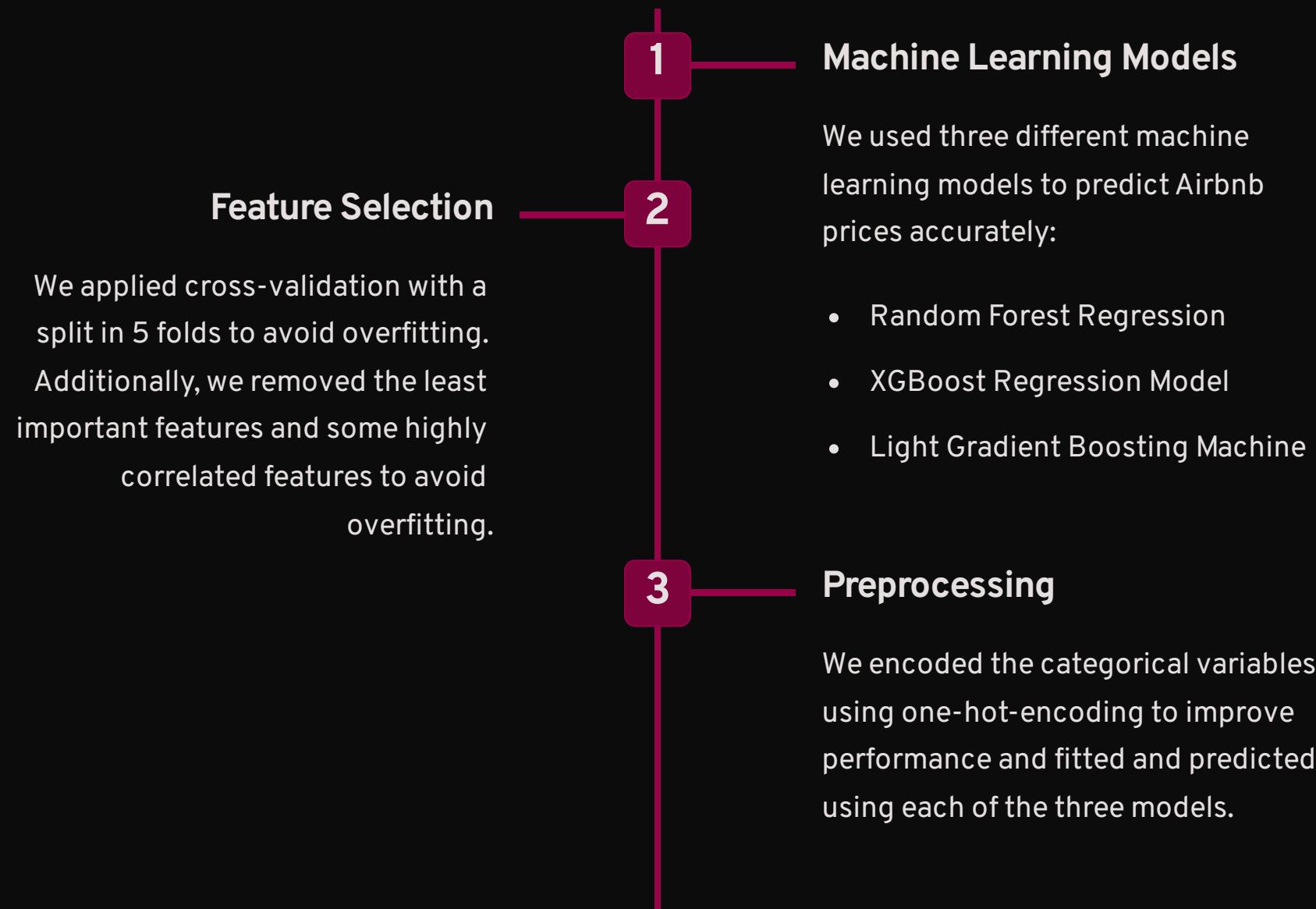
 by Gary Waiyaki



# Introduction

- **Objective:** Identify factors that influence Airbnb prices in San Diego
- **Importance:** Enable hosts to competitively set prices to attract customers
- **Target Variable:** Price
- **Independent Variables:** Amenities, Total Rooms, Property Type, and many others

# Methodology



# Results and Analysis

## Evaluation Metrics

We employed various evaluation metrics, including R-squared Cross-Validation score, and R-squared, to assess the accuracy of our models.

## Key Findings

The Random Forest model accuracy score is: R-squared cross-validation score of -8% and an R2-score of ~52%.

The Light Gradient Boosting (LGBM) model accuracy score is: R-squared cross-validation score of 20% and an R2-score of ~49%.

The XGBoost model accuracy score is: R-squared cross-validation score of -15% and an R2-score of ~61%.

## Implications

Given that the Random Forest and XGBoost models were not able to capture the underlying patterns in the training data, more tuning of these models is required. This means revisiting feature selection and selecting only the important 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, which shows that it too would benefit from more tuning.

A photograph of a tropical beach scene. In the foreground, a large palm tree with a thick trunk and dense green fronds stands prominently. Behind it, another tall palm tree reaches towards the top of the frame. The background shows a sandy beach where a few people are walking or sitting. The ocean is visible in the distance under a bright, clear blue sky.

# Future Improvements

1

## 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.

2

## 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.

- 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 are by the ocean.
  - Old Town and LittleItaly, which have attractions such as boutiques, restaurants, hotels, Museums, etc.
- 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.