DATA1030 | Airbnb Price Prediction

Matthew Dall’Asen, Data Science Institute, Brown Univeristy
Date: 11/13/2024
Github Repository: mdallasen/Airbnb-LA-Pricing

Introduction | Problem & Solution Review

To determine what an “appropriate” price is for an Airbnb listing, this report will predict price per night against several property, review and market characteristics.

Data Source & Description:

Web Scrape of LA Airbnb Listings

Columns

- **10** categorical
- **31** continuous
- **5** qualitative
- **24** non-essential

Rows

- **44,684** total
- **45%** of rows have missing values
- Varying distributions and scales

EDA Insights:

Property Type

Property type significantly influences Airbnb pricing, with larger or more luxurious properties commanding higher prices and greater variability, while shared accommodations have lower and more consistent pricing.

Neighborhood

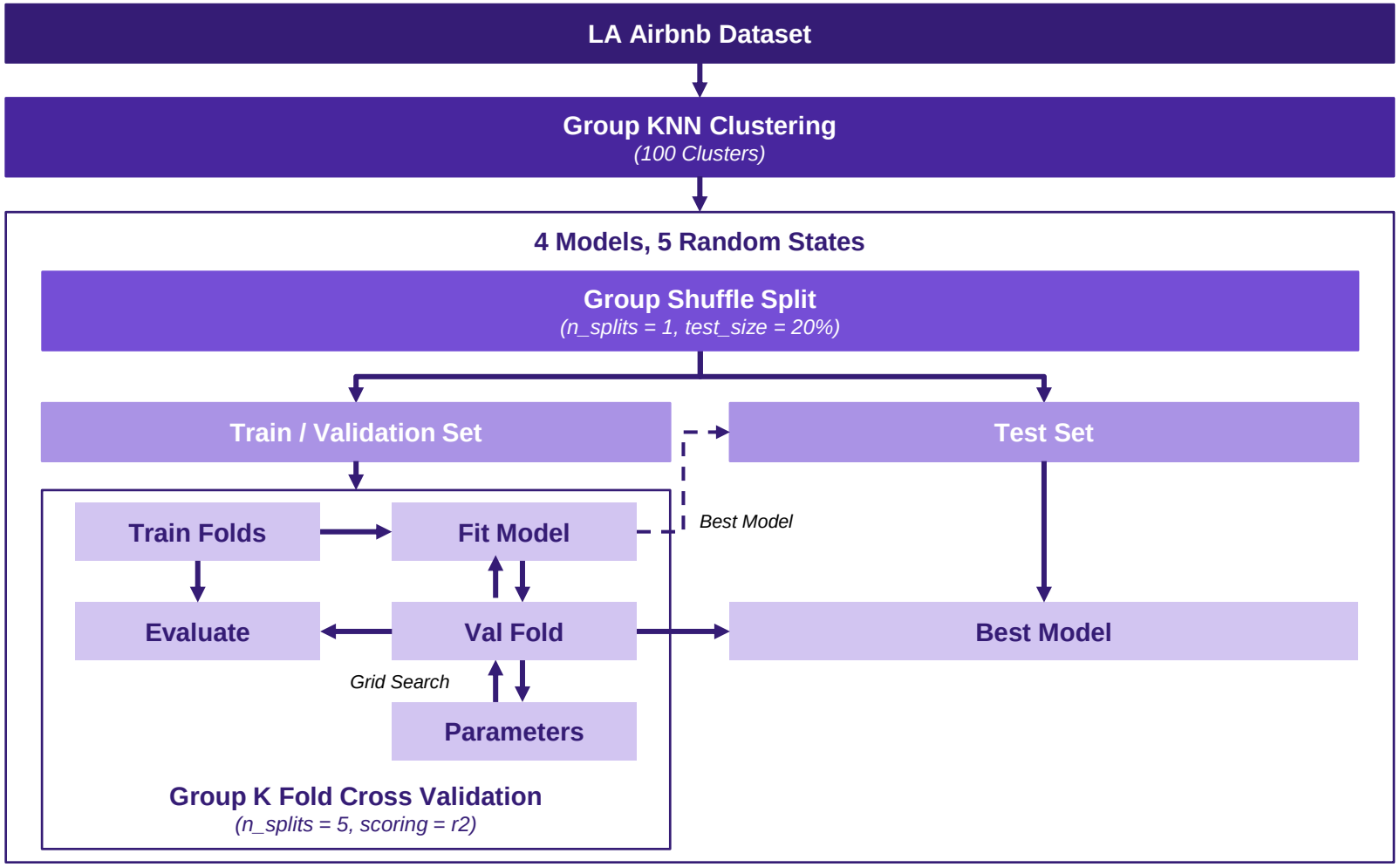
Airbnb prices in LA are highly influenced by neighborhood, with higher prices concentrated in affluent and tourist-heavy areas like Beverly Hills and Santa Monica. Consequently, this is a non IID dataset

Ratings

While most Airbnb listings in LA maintain high review scores (around 5.0) across all price ranges, there is little evidence to suggest a strong linear relationship between review scores and price, indicating that high ratings are common regardless of listing price

ML Pipeline | Splitting & Cross Validaiton

The pipeline addresses non-IID data by leveraging group-aware splitting, cross-validation, and hyperparameter tuning to ensure robust, generalizable performance evaluation.



Description

- Target variable, price, was log transformed
- Given dataset is non-IID, KNN clustering was used to determine groups (neighborhood feature was not deemed sufficient)
- Random State reduces uncertainty and evaluates robustness across splits. Five were used to reduce computational requirements
- Group Shuffle Split prevents data leakage by keeping groups distinct between train and test
- Group K-Fold Cross Validation evaluates generalizability with neighborhood-based folds
- Grid Search tunes hyperparameters to optimize model performance
- R2 was used as the evaluation metric to ensure interpretability of the model, whilst also selecting with the highest ability to explain the target variable

ML Pipeline | Model Deployemnt & Hyperparameter Tuning

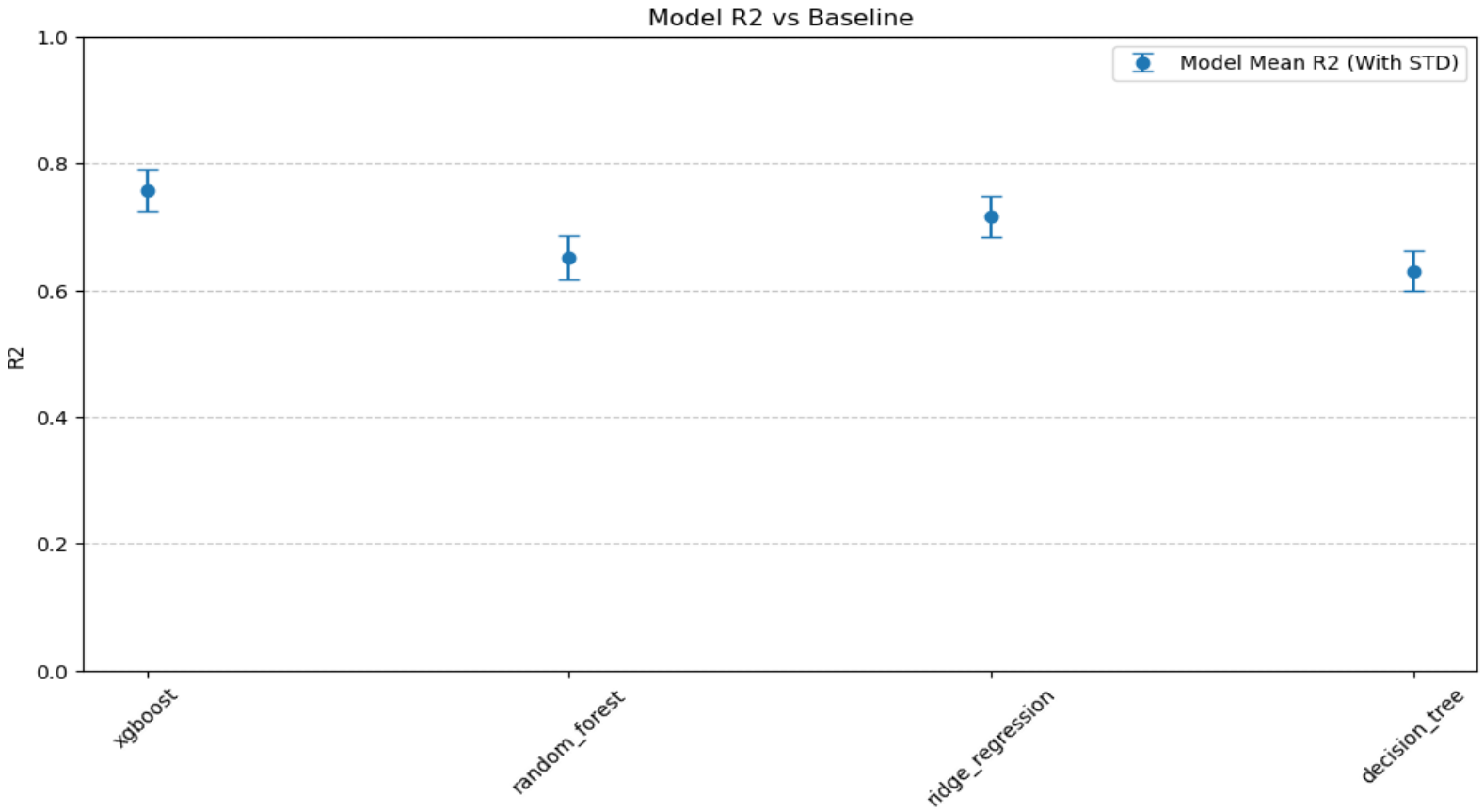
Several linear and non linear models were tuned to determine the optimal testing score. Decision tree like models were preferred given their interpretability and lower computational requriements. All models were fixed at a random state for deterministic purposes.

	Ridge Regression	Decision Tree	Random Forest	XGBoost
Linear / Non Linear	Linear	Non-Linear	Non-Linear	Non-Linear
Description	A linear regression model with L2 regularization that penalizes large coefficients	A tree-based model that splits data into subsets based on feature conditions to minimize prediction error	An ensemble learning method that builds multiple decision trees and averages their outputs for regression tasks	An ensemble of gradient boosted decision trees that uses a collection of weak learner trees to create a strong learner
Parameters Tuned *(Based on Predictive Influence)*	**L2 Regularization:** Adjusts the level of L2 norm regularization present within the MSE loss function	**Max Depth:** Limits the depth of the tree to prevent overfitting	**Max Depth:** Determines the maximum depth of the tree (e.g. splits)	**Max Depth:** Determines the maximum depth of the tree (e.g. splits)
Rationale For Selection	Provides a solid baseline for modeling linear relationships between features (e.g., number of bedrooms, bathrooms, location metrics) and prices.	Captures non-linear relationships effectively and provides interpretable decision paths	Captures interactions between features (e.g., host features and neighborhood amenities) without requiring extensive feature engineering	Well-suited for high-dimensional structured datasets like Airbnb listings, where interactions and complex patterns (e.g., seasonal price fluctuations or feature interactions) are crucial

Results | Testing Scores

From the models tested, XGBoost achieves the highest score when compared to Decision Trees, Random Forests and Ridge Regression.

Mean & STD Test R2 Score:



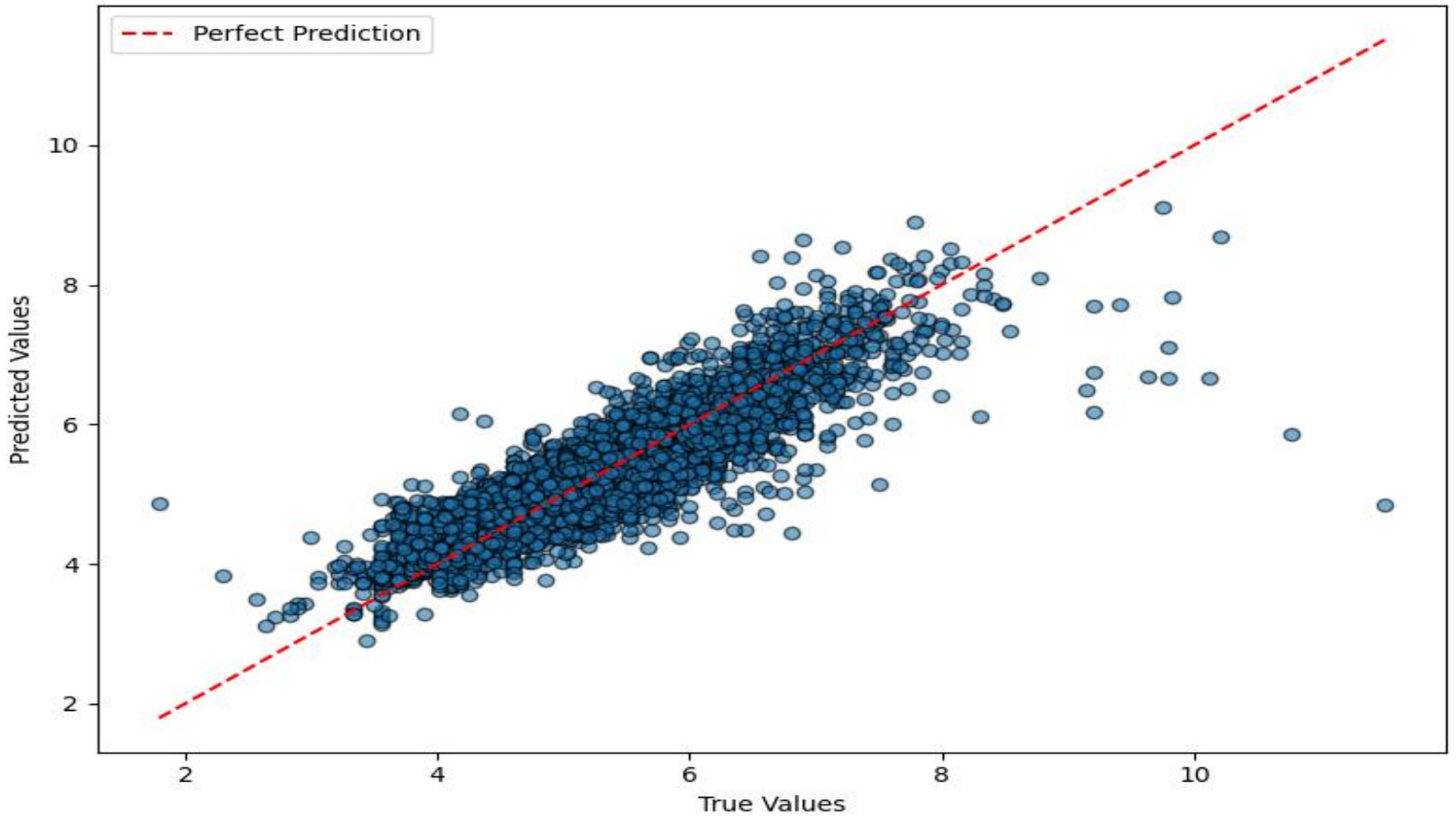
Analysis

- Differences arise from model complexity, regularization, and ensembling, which influence each model's ability to handle non-linear relationships and prevent overfitting.
- XGBoost performs best due to gradient boosting, which optimizes errors sequentially and includes regularization to prevent overfitting
- Random Forest performs well by averaging multiple decision trees, reducing variance, but lacks the boosting advantage
- Ridge_regression struggles with highly non-linear relationships, as it's a linear model that performs best with simpler data patterns
- Decision Tree underperforms because it’s prone to overfitting and lacks ensembling, leading to lower generalization ability. Also sensitive to slight changes in the training data

Results | True Vs Predicted Values

Given that XGBoost is the best-performing model, it exhibits a significant correlation between the predicted and actual values. The model's predictive power is notably high, making it well-suited for addressing pricing problems

Predicted Vs True Value Scatter Plot (XGBoost):



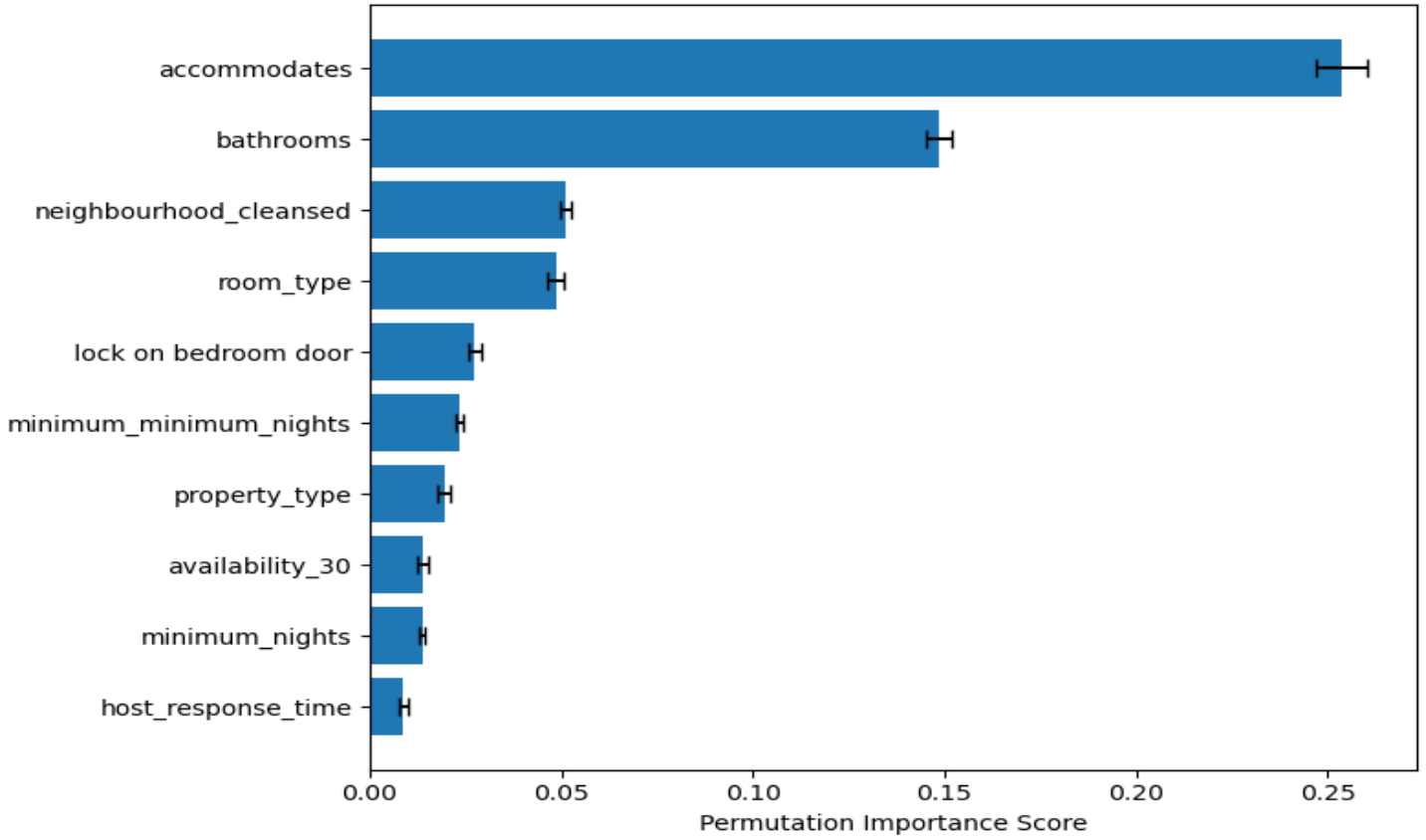
Analysis

- The scatter plot shows a strong alignment between true and predicted values, indicating good model performance
- Points deviate slightly from the red line at higher values, suggesting minor underfitting or noise effects
- XGBoost with max depth 6 balances complexity, capturing trends well while avoiding overfitting
- The spread of points around the line suggests that the model performs well overall but has some deviations, especially for higher values, which might indicate underfitting or noise in the data
- Some points deviate significantly from the red line, especially at the extremes, which could represent outliers or cases where the model struggles to generalize

Results | Feature Importance | Global

To interpret this model, feature importances have been leveraged to understand the key features to are important for LA Airbnb prices. Understandably, core property characteristics are essential to predicting listing prices

Top 10 Permutation Importances



Analysis

- Accommodates, neighbourhood, bathrooms, room_type, and property_type are understandably the most important features for predicting property prices. These core characteristics define the overall “size” and functionality of a property
- Security also emerges as a significant consideration for consumers. Hosts charging more for properties with additional security features highlights its perceived value among renters.
- Availability plays a critical role in pricing as well. Properties available within the next 30 days are seen as more desirable and command higher prices, while those with a higher minimum night requirement cater to longer-term stays, influencing their pricing strategy accordingly

Outlook | Future Work & Improvements

Develop More Complex Models	Leverage additional ensemble methods and neural networks to better capture non-linear patterns in Airbnb pricing, while also expanding the number of hyperparameters to tune within the current models
Capture Additional Pricing Data	Pricing is based on other economic, financial and temporal factors, where additional inclusion of such features would improve model accuracy
Conduct Principal Component Analysis	Reduce feature dimensionality and identify the most significant contributors to price variance
Alternate Between Missing Value Techniques	Compare the model performance against other imputation models, e.g. KNN, or other missing value techniques such as pattern submodel approach