Github

Introduction

In this study, we embarked on a mission to develop a predictive model for apartment rental prices, focusing on units accommodating between 2 to 6 guests. Our data source was an extensive collection of Airbnb listings. Through our analysis, we aimed to not only forecast rental prices for apartments yet to enter the market but also to uncover the determinants of these prices, thereby gaining deeper insights into the real estate sector.

Initial Setup

The groundwork for our analysis involved setting up our computational environment. This setup process entailed importing essential libraries that would enable data manipulation, visualization, statistical modeling, and more. Given the large nature of our dataset, it was partitioned into six segments for manageable processing and subsequently merged into a single DataFrame named listings. The loading phase of this composite dataset took approximately one minute.

Data Cleaning

Our cleaning process focused on refining the dataset to retain only the most relevant columns—specifically, those with less than 30% missing data. We further refined our dataset by adjusting data types for certain columns and eliminating rows with missing price information.

Data Transformation

To enhance our dataset's compatibility with analytical models, we transformed various columns to appropriate data types:

- The host_response_rate, host_acceptance_rate, and price columns were converted to float.
- Boolean data types were assigned to columns such as host_is_superhost, host_identity_verified, and instant_bookable.
- We extracted numerical values for the number of bathrooms and implemented additional transformations to align with our analytical needs.
- We converted essential features such as host response rate, acceptance rate, and price to float types and assigned boolean types to key indicators like host_is_superhost and instant_bookable. Recognizing the importance of complete data, we employed median values to impute missing information across numerical columns, including bedrooms, beds, and bathrooms, ensuring our dataset's integrity.
- Additionally, we introduced binary indicators for missing values in critical host-related features, further refining our analytical foundation. By filling missing values with the most frequent or median values, we significantly enhanced the dataset's compatibility with our analytical models, thereby facilitating more accurate and reliable pricing recommendations.

Data Preprocessing and Analysis

Following an initial filter to include only properties accommodating 2 to 6 guests, we cleaned and prepared the dataset for modeling. This preparation involved:

- Dropping irrelevant columns.
- Converting data types for consistency.

- Handling missing values strategically.
- Encoding categorical variables for model readiness.
- Model Fitting and Evaluation
- Each model underwent a rigorous fitting process using the transformed and cleaned dataset. The Random Forest model, in particular, was optimized through a GridSearchCV process to identify the best combination of hyperparameters.

Model Development

Our analytical endeavor involved constructing three sophisticated models:

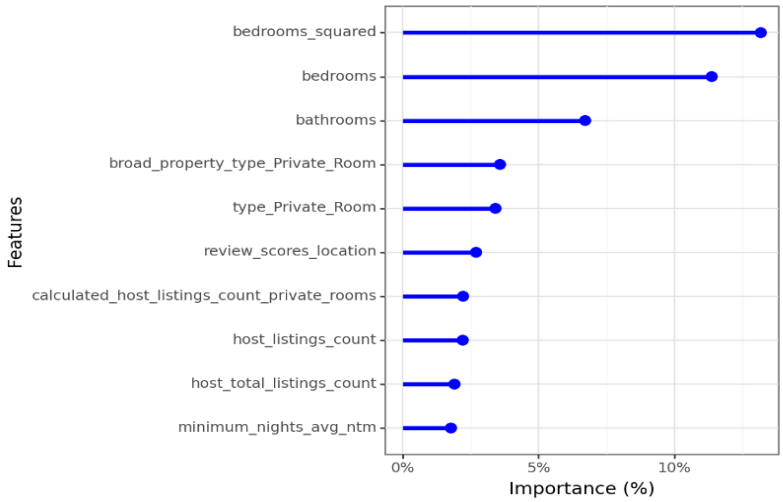
- OLS Linear Regression: Served as our baseline model.
- OLS with Lasso Regularization: Aimed at refining the baseline model by incorporating regularization.
- Random Forest: Explored non-linear relationships and interactions among variables.

The primary objective was to accurately predict the rental prices and, through this predictive endeavor, to identify the variables most influential in determining those prices.

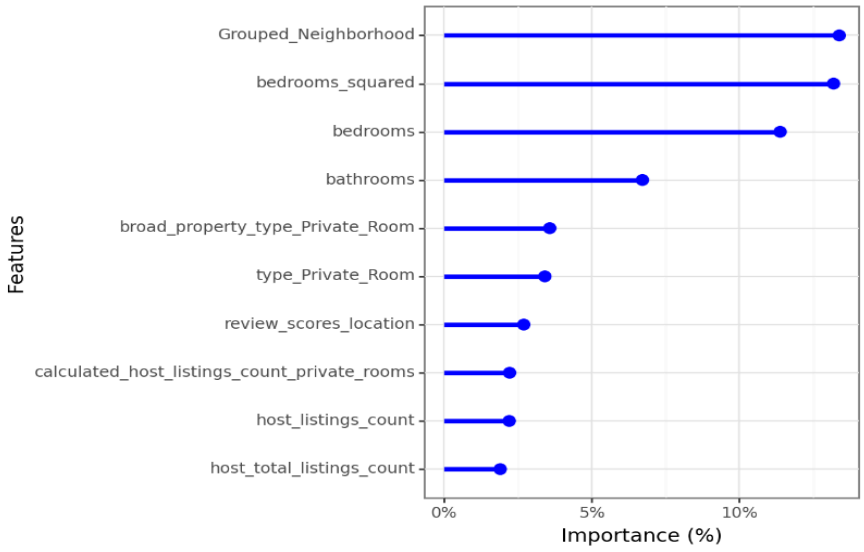
Conclusion

Our comprehensive analysis, underpinned by robust statistical models, has yielded a predictive model that not only estimates apartment rental prices with accuracy but also sheds light on the key factors influencing these prices. This dual outcome facilitates a granular understanding of the real estate market dynamics, particularly in the small to mid-size apartment rental segment. RMSE for Random Forest was better than for OLS and OLS with Lasso. The order is RMSE, OLS with Lasso, OLS. The latter two are actually pretty close. When we calculate RMSE/y coefficient we get ….

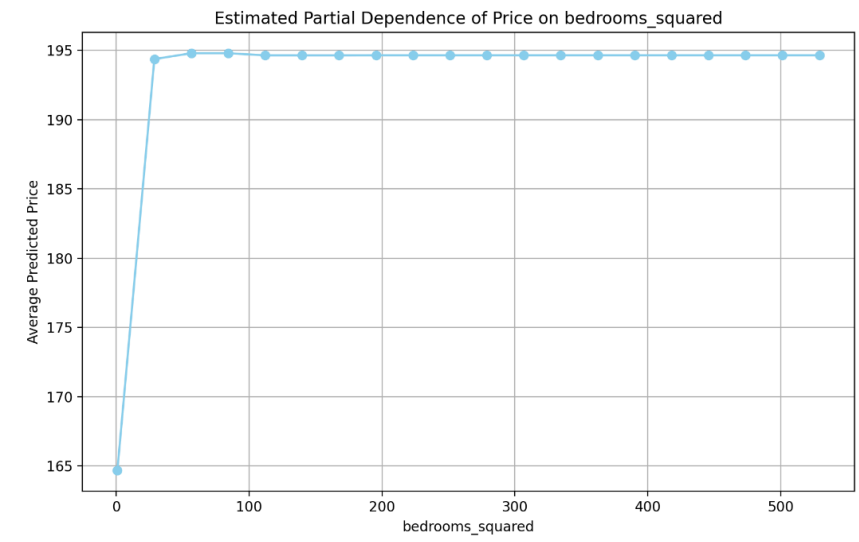
From the initial feature importance analysis, it is evident that the most influential variable is "bedrooms squared," accounting for approximately 13% of the model's predictive power, closely followed by "bedrooms" at about 11.5%. Subsequent to these, "bathrooms" emerges as another significant predictor.



Upon consolidating neighborhoods into a single variable, the analysis yields a slightly altered outcome. The updated feature importance visualization reveals that the aggregated "neighborhood" variable now stands as the most critical predictor, contributing approximately 14% to the model's efficacy, with "bedrooms squared" closely behind at around 13%. "Bedrooms" and "bathrooms" continue to be important, following these leading features.



Partial dependence plot reveals a positive correlation between price and the square of the number of bedrooms up to a certain point. Specifically, this upward trend plateaus around the value of 30 for bedrooms squared. Converting this back to the original scale of bedrooms, this threshold corresponds to approximately 6 bedrooms. This suggests that beyond 6 bedrooms, the increase in the number of bedrooms does not significantly affect the price, indicating that the price largely stabilizes for properties with more than 6 bedrooms.



Insights and Interpretations

- OLS: The average RMSE from cross-validation is approximately 421.61, indicating the model's predictions deviate from the actual log-transformed prices by this amount on average. The average actual price of listings, when converted back from the log scale, is about 216.01 units, serving as a benchmark for evaluating the model's predictive accuracy. The ratio of the average RMSE to the average actual price, standing at 1.95, suggests that the model's error is nearly twice the average price of the listings, highlighting a

significant discrepancy between the predicted and actual values, which may indicate areas for model improvement or the presence of outliers influencing the model's performance.

- OLS with LASSO: The Lasso model's average predicted price, when converted back from the log scale, is approximately 214.23 units. This value represents the mean price across the dataset according to the model's predictions. The ratio of the model's Root Mean Square Error (RMSE) to the average actual price stands at 1.943. This ratio indicates that the average error in the model's price predictions is nearly twice the average actual price of listings. Such a high ratio suggests that while the model provides a general estimation of price trends, its predictions deviate significantly from actual prices, pointing towards potential areas for model refinement or the need for more nuanced feature selection and regularization to improve prediction accuracy.
- For the Random Forest model, the Root Mean Square Error (RMSE) is approximately 263.81, indicating the average magnitude of the errors between the model's price predictions and the actual prices on the original scale. The average actual price across the dataset is about 184.09 units, serving as a baseline to assess the model's predictive performance. The ratio of the model's RMSE to the average actual price is 1.433, reflecting that the model's prediction error is roughly 1.4 times the average price of listings. This ratio suggests that while the Random Forest model exhibits a deviation in its predictions, the discrepancy is somewhat lower compared to the earlier Lasso model example. This indicates a better, yet still improvable, alignment of the model's predictions with the actual prices, highlighting the Random Forest's effectiveness in capturing the dataset's underlying patterns with room for further optimization.

For a company targeting small to mid-size apartments hosting 2-6 guests in Los Angeles, our analysis through the Random Forest model provides a solid foundation for pricing strategy. Given the model's average predicted price of approximately $184.09, with a notably lower RMSE to average price ratio of 1.43 compared to other models, it suggests an optimal pricing range that could enhance market competitiveness while ensuring profitability. Implementing this pricing strategy, especially in a diverse market like Los Angeles, could help in accurately pricing new apartments not yet on the market, catering to a wide array of customer preferences and market demands. This data-driven approach enables the company to strategically position its offerings, potentially leading to increased occupancy rates and maximized revenue.

	Model	RMSE	Average_y	RMSE/Average_y
0	OLS	421.611477	216.014739	1.951772
1	Lasso	416.332300	214.225692	1.943428
2	Random Forest	263.810623	184.086091	1.433083