Airbnb Cleaned Europe Dataset - Business Intelligence

AMAR KARABEGOVIC (B)* and NILS KOPALI (A)*

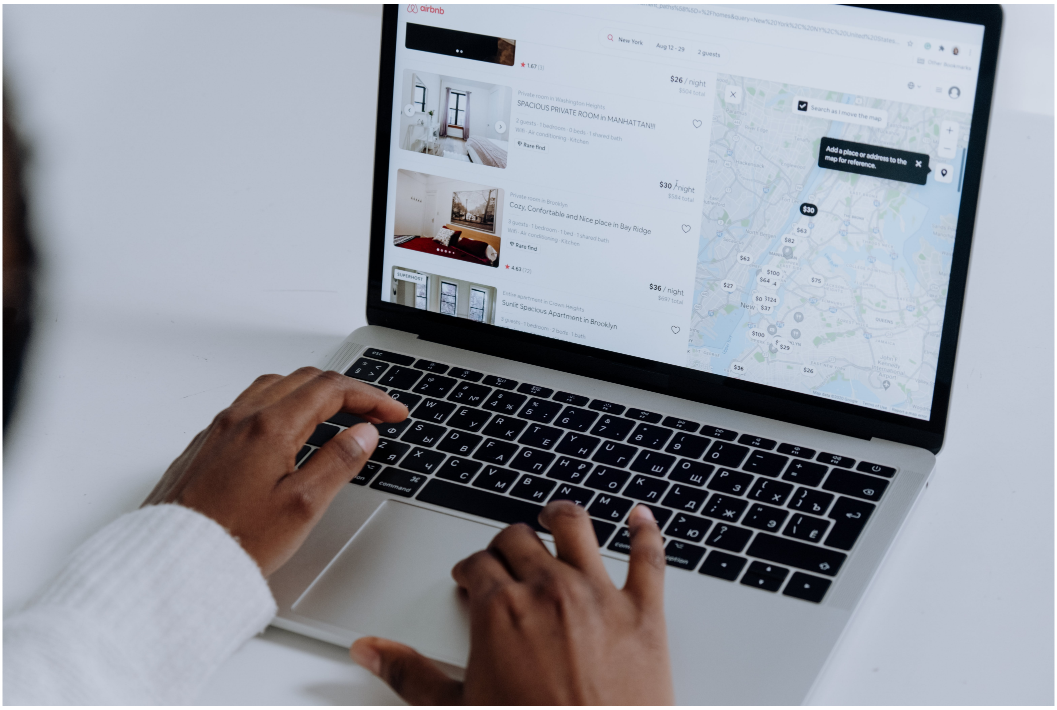

Fig. 1. A person using Airbnb. Source: Pexels

This report details a targeted approach to enhance Airbnb dataset analysis through effective preprocessing techniques. The methodology involves outlier removal, one-hot encoding for categorical variables, and strategic column selection. The attempt aims to streamline the dataset, ensuring it is well-prepared for subsequent analyses. The comprehensive strategy employed in preprocessing endeavors to uncover nuanced insights into the dynamics of the Airbnb market.

Additional Key Words and Phrases: Airbnb, Dataset, Price, Renting, Rental, Room, House, Apartment, ML, Machine Learning, Preprocessing, Modelling

1 INTRODUCTION

The Airbnb dataset offers a detailed examination of Airbnb listings across various cities. It encompasses a range of attributes including the city of the listing, nightly price, whether the price is for a weekday or weekend, and the type of room (private or shared). Additional details include the accommodation capacity, whether the host is an Airbnb Superhost, and the number of rooms available. The dataset also covers aspects relevant for business travelers, cleanliness ratings, overall guest satisfaction, and the number of bedrooms. A significant part of the dataset focuses on the location characteristics of the listings, such as the distance from the city center and the nearest

*Both authors contributed equally to this research.

Authors' address: Amar Karabegovic (B), e1642179@student.tuwien.ac.at; Nils Kopali (A), e1627943@student.tuwien.ac.at.

metro station, along with an attraction index and its normalized form, and a restaurant index with its normalized score. This dataset is a comprehensive resource for analyzing pricing strategies, location attractiveness, and guest satisfaction in Airbnb listings within urban settings.

2 BUSINESS UNDERSTANDING

The first crucial step in a successful data analysis venture involves gaining a thorough understanding of the business context and objectives. This section marks the initiation of our exploration into business understanding, forming an integral part of the report.

2.1 Data Source Description and Scenario for Business Analytics Task

The data source is an extensive collection of Airbnb listings, detailing various attributes such as pricing, room types, location, and guest satisfaction metrics. This dataset could be from an aggregated source of Airbnb listings across multiple cities, potentially gathered for market analysis purposes. A scenario for a business analytics task using this dataset might involve a real estate investment firm looking to enter the short-term rental market. The firm aims to identify profitable locations for purchasing properties to list on Airbnb. The analysis would involve examining trends in pricing, guest satisfaction, and location attractiveness to determine the most lucrative areas for investment.

2.2 Business Objectives

The primary business objective is to maximize return on investment in the short-term rental market. This includes identifying high-demand areas, understanding pricing strategies that maximize revenue while maintaining high occupancy rates, and ensuring guest satisfaction for repeat business and positive reviews.

2.3 Business Success Criteria

Success would be measured by the ability to identify properties and locations that yield high occupancy rates, competitive pricing, and excellent guest reviews. Long-term success would be reflected in sustained revenue growth, repeat customers, and an expanding portfolio of profitable Airbnb listings.

2.4 Data Mining Goals

The goal of data mining is to uncover patterns and insights from the dataset that inform investment decisions. This includes identifying key factors that influence rental prices, guest satisfaction, and demand. Analysis might focus on correlations between location features (like proximity to city centers or tourist attractions), property characteristics (like room type and capacity), and their impact on profitability.

2.5 Data Mining Success Criteria

Success in data mining would be marked by the development of a predictive model or a set of actionable insights that accurately forecast rental demand, pricing strategies, and guest satisfaction levels. The criteria might include high accuracy in predicting price points that balance occupancy and revenue, and insights that lead to above-average guest satisfaction ratings in chosen investment locations.

2.6 AI Risk Aspects for Consideration

When utilizing AI and data mining techniques, it is crucial to consider several risk factors. Firstly, data privacy and ethical use are very important, especially when dealing with individual host or guest data, ensuring compliance with privacy regulations. Secondly, it is essential to address bias and fairness in the model to prevent any unfair advantages or disadvantages for specific locations or types of properties. Thirdly, accuracy and reliability are key, as the predictive models must be precise and dependable to ensure that investment decisions are made based on solid analysis. Additionally, the model needs to adapt to dynamic market conditions, such as changing tourist patterns or shifts in short-term rentals. Finally, the reliance on historical data is a significant consideration, as the model's predictions are only as good as the data it is trained on. Significant changes in market dynamics may involve a retraining of the model.

3 DATA UNDERSTANDING

3.1 Attribute Types and Their Semantics

- **City:** Categorical. Represents the city where the Airbnb property is located.
- **Price:** Numerical (Continuous). The price of the Airbnb per night.
- **Day:** Categorical. Indicates whether the day is a weekday or weekend.
- **Room Type:** Categorical. Type of room offered (e.g., Private room).
- **Shared Room, Private Room:** Boolean. Indicate the type of room.
- **Person Capacity:** Numerical (Discrete). The maximum number of people the property can accommodate.
- **Superhost:** Boolean. Indicates if the host is classified as a Superhost.
- **Multiple Rooms:** Numerical (Discrete). Number of rooms offered.
- **Business:** Boolean. Indicates if the property is suitable for business trips.
- **Cleanliness Rating:** Numerical (Continuous). Rating for cleanliness.
- **Guest Satisfaction:** Numerical (Continuous). Overall guest satisfaction rating.
- **Bedrooms:** Numerical (Discrete). Number of bedrooms.
- **City Center (km):** Numerical (Continuous). Distance from the city center in kilometers.
- **Metro Distance (km):** Numerical (Continuous). Distance from the nearest metro station in kilometers.
- **Attraction Index:** Numerical (Continuous). An index indicating the attractiveness of the property's location.
- **Normalised Attraction Index:** Numerical (Continuous). Normalized version of the Attraction Index.
- **Restaurant Index:** Numerical (Continuous). Index indicating the availability of restaurants nearby.
- **Normalised Restaurant Index:** Numerical (Continuous). Normalized version of the Restaurant Index.

3.2 Statistical Properties and Correlations

Statistical Properties:

- **Price:** Ranges from approximately 35 to 18,545 with a mean of 260.09.
- **Person Capacity:** Mostly between 2 to 6, with a mean of 3.24.
- **Bedrooms:** Range from 0 to 10, with most properties having 1 bedroom.
- **Cleanliness Rating & Guest Satisfaction:** High average ratings (around 9.44 and 93.10, respectively).
- **Distances (City Center & Metro):** Vary significantly, indicating diverse property locations.
- **Attraction and Restaurant Indexes:** Wide range, suggesting varied proximity to attractions and restaurants.

Correlations:

- Price shows some correlation with features like Bedrooms, Attraction Index, Restaurant Index.
- Attraction Index and Restaurant Index are highly correlated.
- There's a notable negative correlation between City Center distance and Attraction/Restaurant Indexes. This exists also between Private Room and Person Capacity and Business with Multiple Rooms.

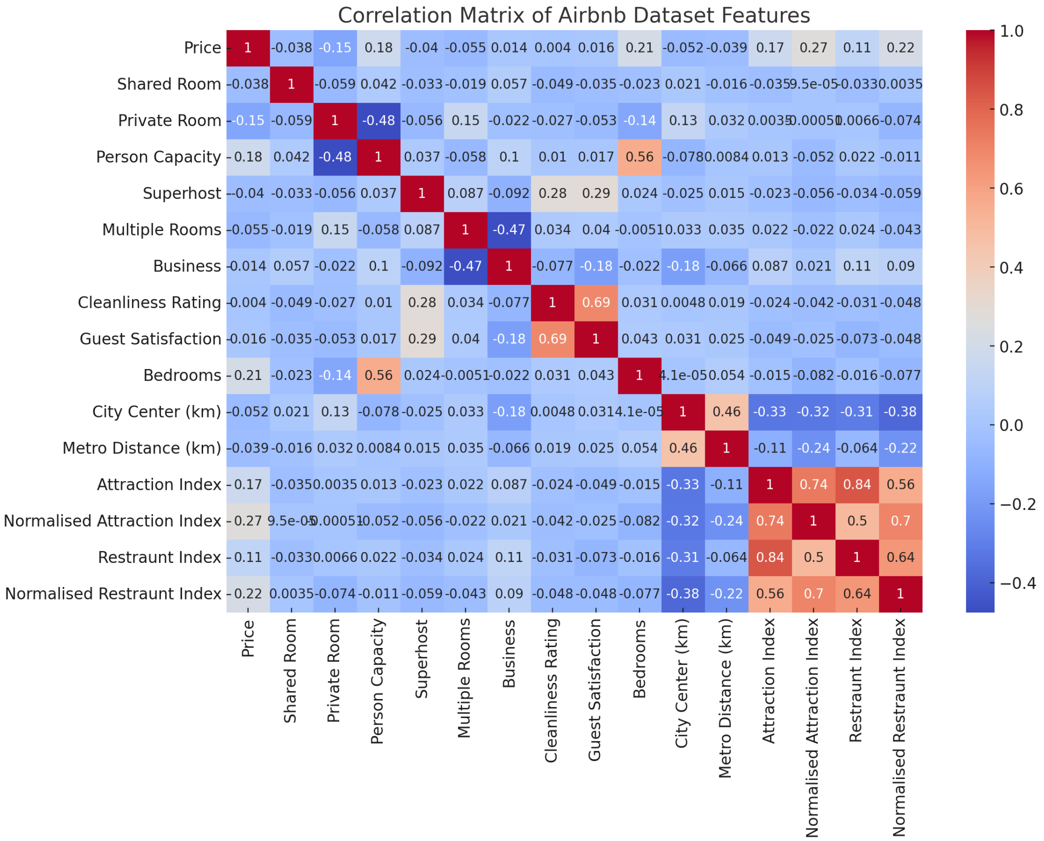

Fig. 2. Correlation Matrix

3.3 Data Quality Aspects

- **Missing Values:** There are no missing values in the dataset.
- **Outliers:** We utilized the z-score method for outlier detection in our data analysis. In total we found 4,068 outliers in different features. In Table 1 we show each of these features.
- **Uneven Distributions:** There's a wide range in prices, with a maximum of over 18,000 euros and a minimum of around 35 euros. This indicates a diverse range of accommodations but could also suggest a skewed distribution towards more affordable options.

The source of this dataset comes from https://zenodo.org/records/4446043.

Table 1. Outliers detected in different features

Feature	Outliers
Attraction Index	559
Bedrooms	86
City Center (km)	615
Cleanliness Rating	578
Guest Satisfaction	661
Metro Distance (km)	865
Normalised Attraction Index	515
Normalised Restaurant Index	244
Price	376
Restaurant Index	633

3.4 Visual Exploration of Data Properties

The histograms on Fig.3 provide insights into the distribution of various attributes in the dataset:

- **Price:** The distribution is right-skewed with a long tail, indicating most listings are in the lower price range, but there are some extremely high-priced listings.
- **Person Capacity:** Most properties accommodate 2 to 4 people, with a peak at 2.
- **Cleanliness Rating:** High ratings are common, with a peak at 10, suggesting overall high standards of cleanliness.
- **Guest Satisfaction:** Similar to cleanliness ratings, there's a concentration of high satisfaction scores.
- **City Center (km):** A wide range of distances to city centers, with a concentration of properties closer to the city center.
- **Metro Distance (km):** Many properties are located close to a metro station, as indicated in the graph.
- **Attraction Index & Restaurant Index:** Both show a wide range of values with a concentration in the lower end, indicating that while some properties are very close to attractions and restaurants, many are not.

3.5 Ethically Sensitive Attributes

The dataset does not contain indicators of race, gender, or religion.

3.6 Potential Risks and Bias Questions

Consideration of potential biases in the dataset raises several questions:

- **Geographic Bias:** Is the dataset representative of all areas within each city?
- **Economic Bias:** Does the dataset reflect a range of economic backgrounds?
- **Cultural Bias:** Are certain cultural or neighborhood preferences overrepresented?

3.7 Actions Required on Data Preparation

- **Outlier Handling:** Handling outliers in numerical columns with the z-score method.
- **Removing Redundant Features:** Eliminating redundant features to simplify the model and reduce complexity such as "Attraction Index", "Restaurant Index", "Normalised Restaurant Index".

- **Encoding Categorical Data:** Converting categorical variables (e.g., 'City', 'Room Type', 'Day') into a machine-readable format.

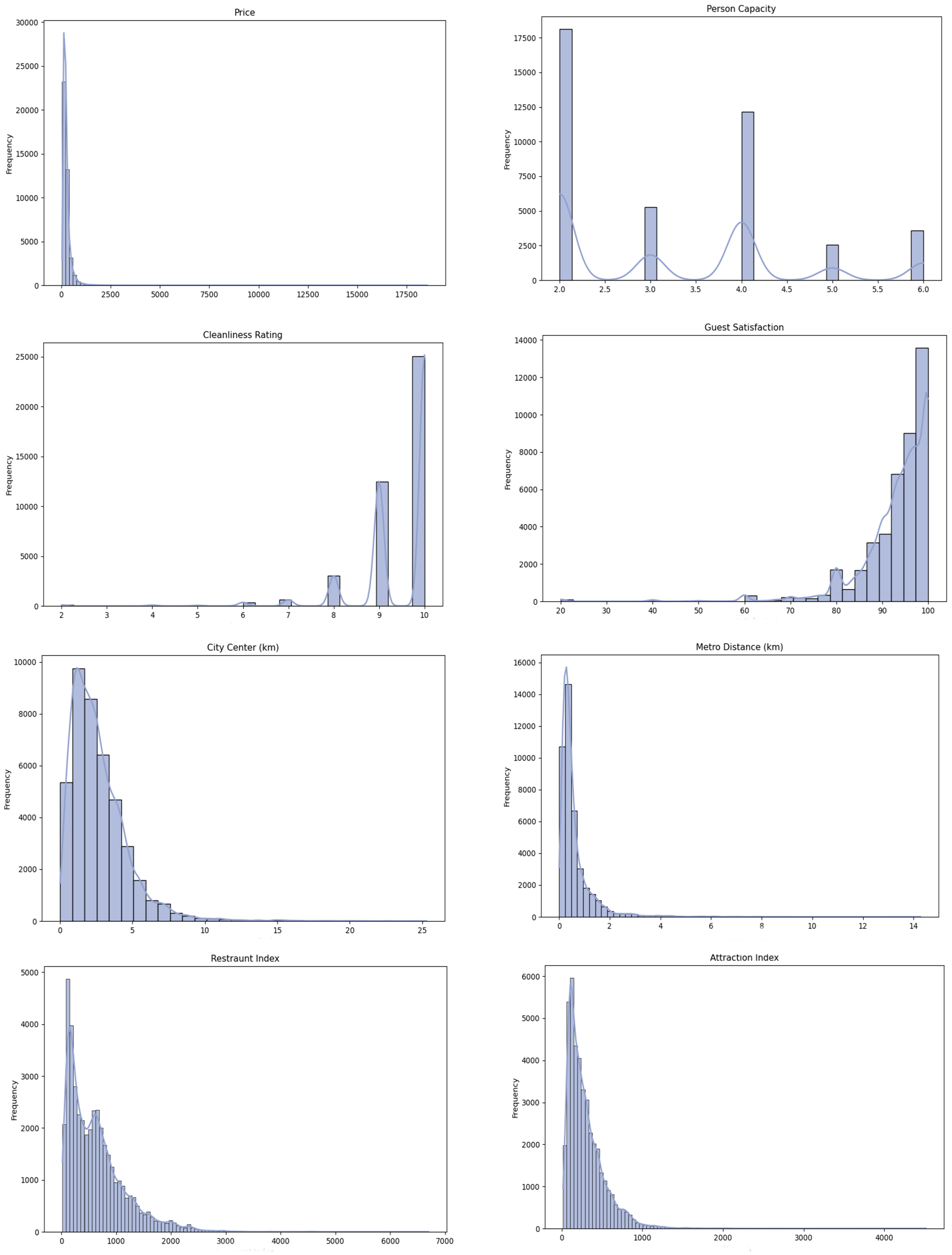

Fig. 3. Bar Charts of the Features

4 DATA PREPARATION

Building upon the recommendations outlined in the preceding section, the data preprocessing strategy strategically incorporates outlier removal, encoding, and column selection to refine the Airbnb dataset. The identification and removal of outliers, highlighted through Z-score analysis, contribute significantly to improving the dataset's overall quality and reliability. Another proposed approach was to use the interquartile range for outlier detection but it was far more aggressive than Z-score as it was removing far more rows, which would most likely cause poor performance of models down the road. Total number of rows removed after the outlier detection was 4068.

In adherence to best practices, the deliberate removal of specific columns, namely 'Restraunt Index' and 'Normalised Restraunt Index', was guided by the identification of high correlation and redundancy within the dataset. This intentional choice is rooted in a nuanced grasp of the dataset's context and is in sync with the broader goals of our analysis. The removal of these correlated columns during preprocessing ensures our dataset is more focused and streamlined. This sets the stage for future investigations with less multicollinearity, making the data more straightforward and easier to interpret.

Moreover, a deliberate strategy is used in managing categorical variables—specifically, 'City,' 'Day,' and 'Room Type'—by employing a one-hot encoding strategy. This method not only improves the representation of categorical data but also aligns with the overarching goal of preparing these variables for subsequent analyses involving machine learning models or statistical methods. To shed light on the unique values within each categorical column, here's a quick rundown:

- **City:** In this column, we find a diverse set of locations, including 'Amsterdam,' 'Athens,' 'Barcelona,' 'Berlin,' 'Budapest,' 'Lisbon,' 'Paris,' 'Rome,' and 'Vienna.'
- **Day:** Moving on to the 'Day' column, we encounter two distinct values: 'Weekday' and 'Weekend,' providing insight into the temporal aspect of the dataset. This column was under high consideration because it might be understood as an ordinal column.
- **Room Type:** Lastly, this column showcases the variety in accommodation offerings, featuring 'Private room,' 'Entire home/apt,' and 'Shared room' as unique values.

4.1 Exploring Opportunities for External Data Integration

In the context of our Airbnb dataset analysis, we have considered potential avenues for incorporating additional external data sources. While it is essential to note that the integration of such data is presented here in a hypothetical manner, the exploration of these options reflects a proactive approach to better addressing our business objectives and data mining goals. Firstly, delving into **Economic Indicators** such as local employment rates, GDP figures, or housing market trends holds the potential to illuminate the broader economic context of the regions in focus. These indicators could shed light on the economic health of specific areas, providing valuable context for variations in Airbnb rental demand. Secondly, the inclusion of **Tourism Statistics** presents an intriguing avenue. Imagining the integration of data on visitor numbers, seasonal patterns, and popular tourist destinations could significantly enhance our grasp of the broader tourism landscape. This context may prove invaluable in understanding patterns of demand for Airbnb accommodations.

5 MODELING

In the modeling section, we aim to predict prices, making regression algorithms suitable. Options include Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, and Gradient Boosting Regression. The following sections will cover tuning, data splitting, training, and evaluating the model's performance using metrics like Mean Squared Error and visualizations.

5.1 Data Mining Algorithms

Given that the focus is on predicting price, regression algorithms would be suitable for this task. Some suitable algorithms include: Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, Gradient Boosting Regression. Considering the nature of the task (predicting a numerical value - price), **Random Forest Regression** is selected. Random Forests are robust, handle non-linearity well, and are less prone to overfitting. They also automatically handle feature importance, which could be insightful for understanding the factors influencing the price.

5.2 Hyper-parameters and Selection

For Random Forest Regression, some key hyperparameters include the number of trees, maximum depth of the trees, and the minimum number of samples required to split a node. In this context, **the number of trees** (n estimators) is selected for tuning. The justification is that adjusting the number of trees can impact the model's generalization and performance.

5.3 Train/Validation/Test Split

Given that the task involves predicting prices, a common practice is to use a train/validation/test split. A typical split might be 60% for training, 20% for validation, and 20% for testing. Stratification might be considered if there is a class imbalance in the target variable. Additionally, since there is a potential for dependencies between data instances (e.g., time series data), the split should be done carefully, ensuring that the temporal order of data is preserved. To ensure that each subset (training, validation, test) had a representative range of prices, we used quantile-based binning. This process involved dividing the range of the "Price" variable into a fixed number of intervals, or bins, with each bin containing an approximately equal number of data points (based on quantiles). By using these price bins as a basis for stratification in the train-test split, we ensured that the distribution of prices in each subset was similar.

5.4 Training and Hyperparameter Tuning

Random Forest Regression will be trained on the training set, and the number of trees (n estimators) will be tuned on the validation set. The performance of different hyperparameter settings will be compared, and the best setting will be selected based on the validation set performance.

5.5 Performance Metrics and Visualization

Performance metrics for regression tasks include Mean Squared Error (MSE) and R-squared. These metrics will be used to evaluate the model's performance. The tunning process is visualized in Fig.4. To determine the optimal "n_estimators" for our model, we utilized GridSearchCV, exploring values from 100 to 350. The best value for us was 350.

6 EVALUATION

In this section, we evaluate the performance of the final model applied to the test data, re-trained with identical hyperparameters on the full train and validation data, and compare the outcomes with benchmark performances and baseline expectations. Additionally, we examine whether the model exhibits any bias towards a protected attribute.

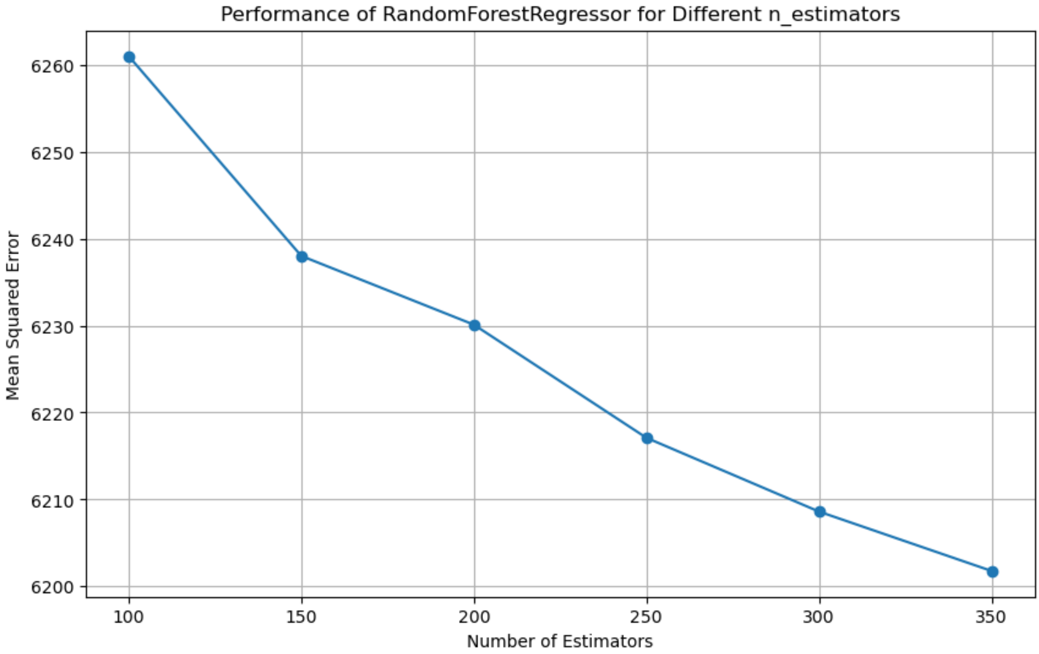

Fig. 4. Hyper-Parameter Tunning

6.1 Application of Final Model on Test Data

The quest for achieving optimal model performance began with the initial findings revealing an RMSE range of 6000 to 7000. This considerable variability indicated a notable inconsistency, particularly given the dataset's price range. In response, an exhaustive reassessment of the preprocessing pipeline was initiated. The improvement involved refining outlier detection through the utilization of the Interquartile Range (IQR) and implementing log transformations on attributes.

These enhancements to the preprocessing steps had a significant impact on the model's performance. The initially high RMSE saw a substantial reduction, reaching approximately 51 when evaluated on the validation data after training. This noteworthy progress highlighted the crucial role of robust preprocessing in elevating the model's predictive capabilities and aligning it more closely with the underlying dynamics of the data.

6.2 Re-training and Test Performance

To amplify the model's predictive capabilities, a re-training process ensued, utilizing the complete training and validation datasets with the previously determined optimal hyperparameters (350 n estimators). The outcome of this re-training was promising, as the model achieved an improved RMSE of around 51 on the test set. The enhancement in performance indicates the efficacy of the refined preprocessing strategies.

6.3 Benchmark and Baseline Performance

Our model has demonstrated superior performance compared to the baseline model, achieving an RMSE of approximately 51. Despite the absence of information on state-of-the-art performance in peer-reviewed papers, we identified a benchmark from a Kaggle notebook by Evan Jones. In this

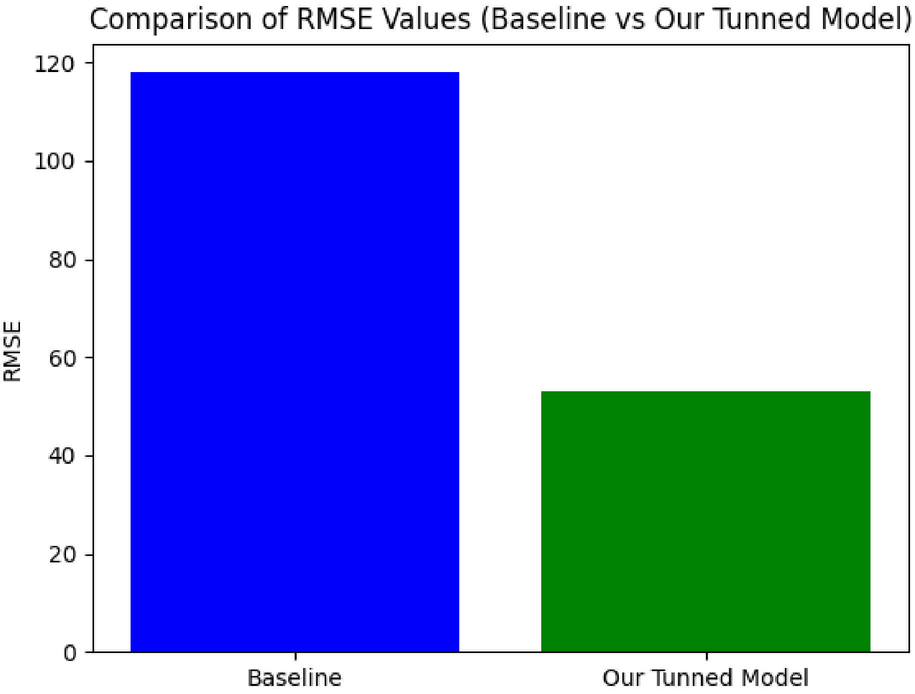

Fig. 5. RMSE Comparison

benchmark, the reported RMSE was around 118 in multiple runs. The substantial disparity between our model's initial performance and this benchmark prompted a reevaluation of our modeling approach. The realization that the initial RMSE ranged from 6000 to 7000 underscored the need for urgent improvement. Consequently, we implemented a robust preprocessing strategy, which included outlier detection using the Interquartile Range (IQR) and log transformation of attributes. This transformative preprocessing significantly reduced the RMSE, resulting in a final model with an RMSE of approximately 51 on the test set.

Evan Jones' Kaggle notebook served as a compelling benchmark, providing a notable reference point. It is worth highlighting that Evan Jones had the flexibility to tune multiple hyperparameters, a luxury not available in our approach, where only one hyperparameter could be adjusted. This emphasizes the significance of recognizing and considering variations not only in model adaptability but also in the tuning of specific hyperparameters when comparing benchmarks. Moreover, it is crucial to acknowledge that differences in preprocessing methodologies could contribute to the observed distinctions in model performance. Therefore, a comprehensive benchmark analysis requires a nuanced understanding of both the model's design and preprocessing strategies.

6.4 Alignment with Business Understanding

With our model showcasing strong performance, characterized by an RMSE of 51, it emerges as a valuable asset for investment considerations. The success criteria established in the Business Understanding phase underscore the importance of identifying properties and locations leading to

Model	RMSE
Our tunned model	51.129724
Baseline - All features	93.77347332445888
Baseline - Top half features	95.89540288623643
Baseline - Top quarter features	128.1794086534142

Table 2. RMSE Comparison Baseline (All runs) and Our model

high occupancy rates, competitive pricing, and positive guest reviews. Anticipated long-term success involves sustained revenue growth, repeat customers, and an expanding portfolio of profitable Airbnb listings.

Evaluating the model against these success criteria, the achieved RMSE of 51 on the test set indicates accurate predictions of property prices. This level of precision aligns well with the business goals, emphasizing the identification of properties that ensure high occupancy rates and competitive pricing. In summary, the model's outstanding performance underscores its applicability for investment decisions. The achieved level of accuracy aligns effectively with the outlined success criteria, emphasizing its utility in identifying properties that contribute to the desired outcomes in the realm of Airbnb property management and investment.

6.5 Bias Evaluation

Turning attention to the assessment of bias in our model, we directed our focus towards a key attribute, namely the "Superhost" column. By comparing the Mean Predicted Price for Superhost and Not Superhost categories, we sought insights into any potential bias in the model's predictions.

The analysis of the Mean Predicted Price revealed no discernible bias. Specifically, the model predicted a Mean Price of 209.66 for Superhost properties, while it forecasted a slightly higher Mean Price of 220.93 for Not Superhost listings. This lack of substantial disparity suggests that the model's predictions do not favor one category over the other, indicating a fair and unbiased approach in its price predictions for Superhost and Not Superhost properties.

It's important to note that the choice of the "Superhost" attribute for bias evaluation stems from its relevance and interest in investigating any potential differential treatment in predictions based on this distinction. This comprehensive evaluation contributes to ensuring fairness and impartiality in the model's predictions across diverse segments of the dataset.

7 DEPLOYMENT

7.1 Comparison of Performance with Business Objectives

The performance of the Random Forest Regressor, characterized by the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) on test and validation datasets, suggests a reasonable level of predictive accuracy. However, its alignment with business objectives requires a deeper analysis:

(1) **Accuracy vs. Business Needs:** The MSE and RMSE values provide an initial measure of the model's predictive accuracy. However, it is crucial to evaluate how these predictions impact key business variables such as demand, pricing, and guest satisfaction.

(2) **Additional Analyses Needed:** To fully meet the business objectives, analyses encompassing customer sentiment, market trends, and competitive benchmarking are recommended. These should complement the predictive model.

7.2 Recommendations for Deployment

- Implement a hybrid solution where the model aids in decision-making, supplemented by human expertise.
- Engage in continuous model improvement using new data and feedback.

7.3 Ethical Aspects and Impact/Risk Assessment

(1) Address data privacy and security concerns, especially regarding personal information.
(2) Maintain transparency about the model's influence on decision-making and establish accountability mechanisms.

7.4 Monitoring Aspects During Deployment

- Regularly monitor performance metrics such as MSE and RMSE.
- Track the correlation between model predictions and actual market outcomes.
- Implement a feedback loop to monitor user feedback on model recommendations.

7.5 Revisiting Reproducibility Aspects

The use of a Jupyter Notebook and a specific seed for operations in our analysis significantly enhances reproducibility. The interactive and well-documented nature of Jupyter Notebooks, combined with the consistency brought by a fixed seed, ensures reliable replication of results.

8 CONCLUSION

In conclusion, our analysis using a Random Forest Regressor, shows promising results for addressing our business objectives in the short-term rental market. While the model demonstrates reasonable accuracy, further enhancements and analyses are recommended to fully align with the business goals. Ethical considerations, careful deployment, and continuous monitoring are imperative for the model's success. The approach taken enhances reproducibility, but it is crucial to maintain comprehensive documentation throughout the process for complete transparency and replicability of the results.