Group 18: Predicting Airbnb Prices in Italy

Benjamin Look	Jeremy Gonsalves	John Turnbull	Matthew Morano	Jared Simpson
20151787-School of Computing	*20158418-Queens Smith*	*20235355-Queens Smith*	*20226100-Queens Smith*	*20180803- School of Computing*
Queen's University	*Engineering Dept.*	*Engineering Dept.*	*Engineering Dept.*	*Queen's University*
Kingston, Canada	Kingston, Canada	Kingston, Canada	Kingston, Canada	Kingston, Canada
ben.look@queensu.ca	*18jag10@queensu.ca*	*20jjt@queensu.ca*	*19mjm23@queensu.ca*	*19jms3@queensu.ca*

Abstract—This project develops a predictive model for Airbnb rental prices across various cities in Italy. The significance in this project lies in using past research to delve deeper on factors which affect Airbnb prices and quality with insufficient research. Our main areas of focus included the impact of host-related features, geographical features, and understanding which features make the greatest effect on user ratings. Through our analysis, it is clear host-related features do have a moderate impact and contain useful information on predicting Airbnb prices. Hypothesis tests stated there is a statistically significant relationship between geographical features and price, and geographical features alone were able to relatively accurately predict price in a linear regression model. Average review scores were also able to be predicted to a substantial extent, but not quite as well as prices.

1 Introduction

Our team aimed to develop a predictive model for Airbnb rental prices in Italy, focusing on integrating a wide range of factors, including customer reviews, ratings, minimum stay, reviews and other requirements alongside traditional listing attributes. Additionally, we aimed to predict customer review ratings, both numerical and categorical, by considering host attributes such as amenities, property type and house price. These tasks sought to refine pricing strategies for hosts, offer accurate price estimations for guests, and enhance the overall comprehension of market dynamics in a region renowned for its tourist appeal. Motivated by the intricate dynamics of the short-term rental market and the interest in Italy's unique blend of cultural and historical offerings, our project addresses the need for sophisticated, adaptive pricing models considering quantitative and qualitative listing details. Given the disparity in population, tourist attractions and climate zones in Italy, this is undoubtedly an interesting region to analyze as it provides a more detailed overview of the driving prices of Airbnb rentals.

The main beneficiaries of this model include Airbnb hosts in Italy, potential guests looking for fair pricing, and market analysts interested in the short-term lodging sector.

Our project aims to create an adaptive, dynamic pricing model incorporating real-time market trends and feedback, a novel contribution not extensively covered in the existing Airbnb data. Additionally, it tailors the attributes to ensure only the relevant features are used when performing this type of analysis in the future. This approach promises to offer optimized pricing strategies for hosts and transparent pricing for guests, contributing valuable market intelligence and methodological advancements to the field.

2 Related Work

"An ensemble machine learning framework for Airbnb rental price modelling without using amenity-driven features" [1] is a paper which presents an innovative approach for predicting Airbnb prices. The researchers created a model without inputting any amenity-driven features by utilizing data from 75,000 listings across five American cities and performing feature screening, stacking with ensemble algorithms, particle swarm optimization, and explainable AI. This model focuses primarily on Bedrooms, room type, Reviews, and accommodations. While it may not serve as a direct baseline for our project, the paper is relevant as it provides us with an alternate approach to more traditional amenity-driven models and results, which could be used as a comparative metric to our own results.

"Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings": This article focuses on which factors influence Airbnb pricing using regression models [2]. Its novel feature includes analyzing sharing economy-related factors or the elements unique to Airbnb. One of the main topics is how the quality/professionalism of the host impacts prices. They found that location, size, privacy, capacity, and potential amenities are large indicators. However, it struggled to pinpoint the effect of a highly rated/professional host. This is caused by the charging lower prices to fill their beds, while casual listers carefully consider the risks of hosting others and charge a premium. It explains this context requires further research, which we will include in our project. This article is also relevant to our project in many other ways, including narrowing down the large number of columns in our dataset (75), identifying areas requiring further research to help guide our questions, and helping identify and/or avoid limitations or pitfalls similar projects face.

Peilu Liu's "Airbnb Price Prediction with Sentiment Classification," a 2021 Master's project from San Jose State University [3], innovatively combines machine learning models with sentiment analysis to predict Airbnb rental prices in San Francisco. Utilizing models such as

Regression Tree, Gradient Boosting, and SVR, alongside customer review sentiment analysis via Textblob, Liu's research accurately predicts rental prices, underscoring the significant impact of customer sentiments. This study, which is directly relevant to our project on analyzing Airbnb listings in Italy, highlights the importance of integrating qualitative review insights with quantitative data for enhanced prediction accuracy. As a baseline, Liu's work demonstrates a pioneering approach in the field, suggesting that sentiment analysis significantly contributes to understanding and predicting Airbnb pricing dynamics. This methodology offers a comprehensive framework for our investigation into predictive modelling for Airbnb prices, emphasizing the potential of sentiment analysis to enrich price-prediction models.

The paper "The Hedonic Price Model of Online Short-term Rental Market Based on Machine Learning" by Jin Xin and Lei Xue takes a classification prediction approach, using the AutoGluon model to predict different Airbnb price ranges in Beijing [4]. The paper looks to solve the problem of uneven development of short-term rental models in the regions of Beijing and the lack of reasonable pricing in these markets. In addition to creating the model, the paper explores the importance of the features used so that hosts and guests can better understand price drivers in the market. This article is useful to the proposed project as it introduces the methodology of transforming numeric data into categorical data to allow for categorical prediction models given a regression problem.

"Reasonable Price Recommendation on Airbnb Using Multi-Scale Clustering" aims to evaluate clustering methods in predicting Airbnb house prices based on the proximity of landmarks and facilities, incorporating reviews as a crucial factor in the process [5]. The Multi-Scale Affinity Propagation (MSAP) method was introduced to aggregate houses effectively. Within each cluster, the Linear Regression model with Normal Noise (LRNN) was utilized for price prediction and validated through increasing rental reviews. Experimental results significantly improved price prediction precision using the MSAP method, revealing a diverse pricing distribution within each city. The study emphasized the relevance of leveraging clustering methods, particularly MSAP, in predicting reasonable prices and highlighted the applicability of the LRNN model. The paper's relevance to the group project lies in providing valuable insights into advanced nuances of Airbnb data and offering guidance on clustering methods that can benefit similar data exploration, emphasizing the potential relevance and efficiency of the MSAP method in predicting reasonable house prices.

3 Research Questions and Methodology

3.1 RQ1: What is the impact of key features associated with Airbnb listings, such as property type, amenities, and house prices, on user ratings?

Motivation: Understanding how key features influence Airbnb user ratings is essential for improving the platform, helping hosts optimize listings, informing travellers about better choices, and providing insights for researchers. Overall, it enhances the Airbnb system for hosts, guests, and the platform.

Methodology in Data Exploration and Feature Selection: The impact of various listing features on user ratings was studied for this research question. The initial step involved selecting relevant columns directly related to the research question. These columns included listing age, availability, instant bookability, recency of listing, review scores variance, city encoding, room type, accommodations, beds, price, monthly reviews, number of bedrooms and bathrooms, and average review scores. The focus on these specific features was crucial as they directly link to understanding factors that may influence guest satisfaction and rating behaviour.

First, the dataset was refined by removing all empty values and replacing them with the average of that column. Then, a correlation matrix was used to remove multicollinear features, as shown below in Figure 1.

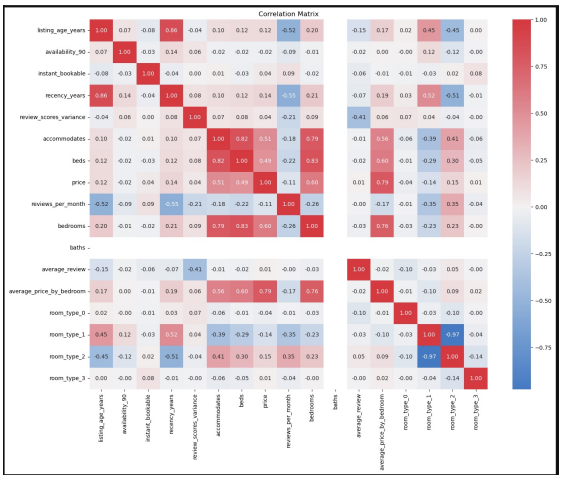

Fig. 1: Correlation Matrix Used for RQ1 in Determining the Pertinent Features in Predicting the Average Review

From this, the data was refined to pertinent columns, removing multicollinear variables such as listing age, number of beds, baths, and bedrooms. This data-cleaning process was essential to ensure the integrity and relevance of the data for the analysis. An initial inspection of the cleaned data was conducted to ensure the correct columns were retained and verified the absence of missing values across these columns. With this, the average review was plotted against the number of bedrooms to understand better the variability in data, which can be seen below in Figure 2.

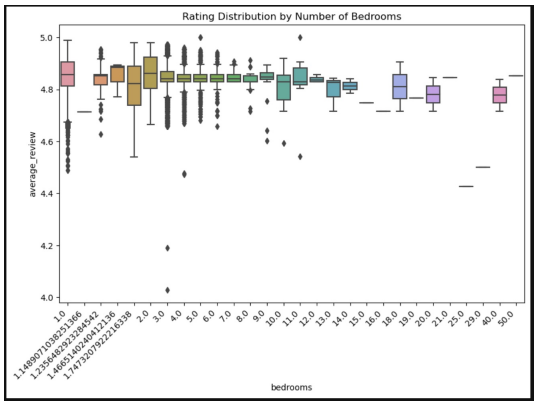

Fig. 2: Rating Distribution by Number of Bedrooms for Airbnb Listings

Next, one-hot encoding was implemented on the room type and city variables to facilitate using categorical data in modelling, making a new column defined by a one or zero, denoting the room type and City it was in. This conversion of categorical variables into a series of binary columns allowed for the inclusion of these attributes in the modelling process. Box plots were generated to visually assess the distribution of average reviews across different numbers of bedrooms, offering insights into potential trends or anomalies for each bedroom type. Additionally, a bar plot displaying the listings count by room type helped understand the dataset's composition, clearly depicting the variety and prevalence of different accommodation types available on Airbnb.

Finally, outlier detection and removal were executed using the interquartile range method, specifically focusing on numerical features grouped by room type and the number of bedrooms. This approach helped mitigate the effects of extreme values that could skew the analysis results in the room type and the number of bedroom groupings. Following this features such as listing age, availability, and prices were normalized using Min-Max scaling to ensure all numerical data were on a common scale, enhancing the fairness and accuracy of the predictive models that are later applied.

For this research question, it is important to highlight that the study faced a notable challenge when analyzing the amenities feature due to parsing issues within the Airbnb database. The inability to accurately extract amenities data in the initial Airbnb database, resulting from web scraping limitations, highlighted a critical area for improvement outside the current project scope.

3.2 RQ2: How do host-related features affect pricing?

Motivation This research question is most relevant to the host as it could give hosts insights on what factors related to them affect pricing. Host-related features such as Superhost status, response time, number of listings and response rate impact pricing give customers the impression that the host is trustworthy and the listing is of high quality, allowing the host to ask for a higher price potentially. Examining this relation will allow hosts to improve the features within their control to raise their revenue and improve guest experiences.

Methodology in Data Exploration and Feature Selection: Research was done to find which host-related features were most relevant to the pricing. To do this, the data set was reduced to only include host-related features. The data was then cleaned, and categorical features were included using one-hot or ordinal encoding, depending on the data type. Histograms for each relevant column were constructed to understand the distribution of relevant features better. A correlation matrix was then made to visualize the linear relationship between features more easily, especially concerning the 'price' column. Features that displayed multicollinearity were removed such as 'host since', 'host response rate' and 'host acceptance rate'. However, the results of these methods still did not show a significant correlation to price. Two models were developed to understand better which host-related features were linked with price. The current dataset was split into training and testing data. An XGBoosted (Extreme Gradient Boosted) and RandomForest models were designed and then fit into the data set. Each model produced a graph displaying feature importance when used to predict pricing. These can be seen below in Figure 3 and 4

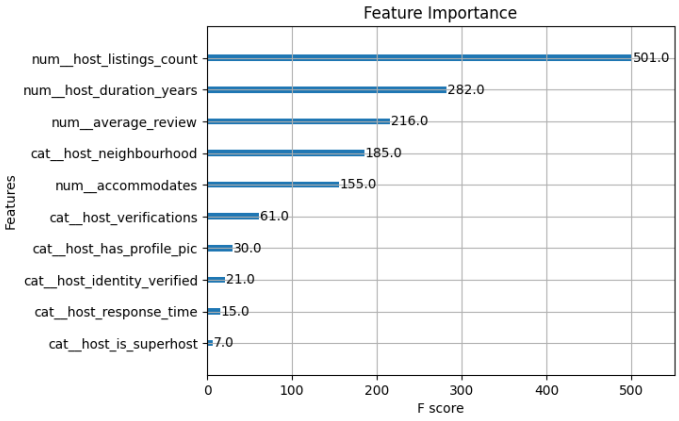

Fig. 3: Feature Importance from XGBoosted Model.

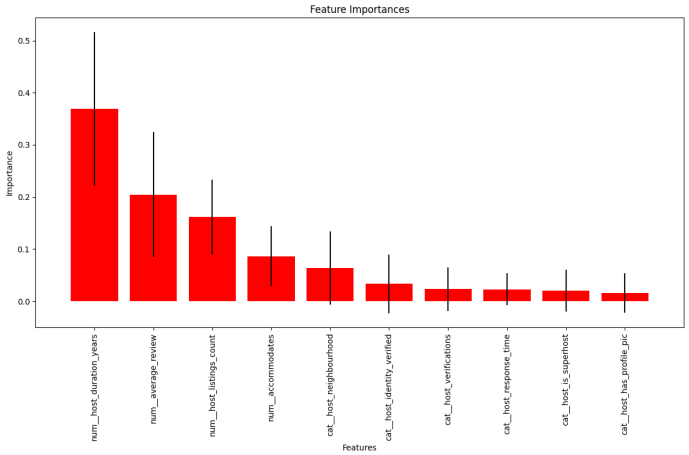

Fig. 4: Feature Importance from Random Forest Model.

The two models were used to account for how different models use information. Cross-referencing results show

that certain features are constant as the most heavily weighted. These features were 'host duration in years', 'average review' and 'host listings count'. While 'host neighbourhood' was also important in both models, it will be too closely related to the neighbourhood feature, so it will not be included. These selected features will be used alongside additional features in baseline models to determine how much they affect pricing. Before testing, outlier detection and removal were done using the interquartile range method. This was done to help reduce the effects of extreme values on the model's analysis. Normalization was applied to the entire data set (except the price) with a Min-Max Scaler to ensure consistent ranges for all features. Then, two data sets were created; one dataset contained host-related features, while the other did not. In preparation for testing, hyperparameter tuning was completed for an XGBoosted and RandomForest model.

3.3 RQ3: What is the impact of geographical-related features, like city, region, distance to the city center, and others, on the pricing of Airbnb in Italy?

Motivation: Understanding the impact of geographical location on price will provide beneficial insight for both hosts and visitors. Hosts will benefit as they better understand the market price in their location, allowing them to better compete on price for their given location. Visitors will gain a more diverse range of benefits as understanding how different regions and cities impact prices will provide insight into where to travel to save money. Additionally, factors like distance from the city center may highlight other ways to save money while travelling. It is hypothesized that larger, more populated cities will have higher rental costs. Additionally, houses closer to the city center will, no matter the city or town, will have higher rental costs.

Proposed Methodology: The following methodology will assess the different impacts of geographical features. Geographical-related features include longitude, latitude, and distance to the center, city, and neighbourhood. First, the data was preprocessed. In this step, the data was examined for missing values, outliers, and unnecessary features, increasing the data set's complexity. Next, descriptive statistics and visuals were explored to understand the distributions and variations of the different columns. Correlations were examined for potential causes of multicollinearity within the attributes. Multicollinearity is likely to occur, given the relationships between geographical features. Thus, it was essential to address this. Once the data is better understood, feature engineering will be pursued. Feature engineering will likely involve appending a population column from a related table, calculating the distances from the city center given the longitude and latitude data, and other necessary methods to improve the predictability of price given the features. Due to complications with obtaining population data, it was scrapped from the analysis.

To start, key visualizations were made to try and identify any visual trends, such as the average price of cities, and scatter plots with price as the y-axis to highlight any price trends for given geographical variables. After identifying potential visual trends, hypothesis testing was completed. The hypothesis tested was for each city, do changes in the distance from the city center impact pricing? To test this hypothesis, each city's price is checked for a normal distribution, and then a T-test or Wilcoxon signed-rank test is completed depending on the distribution result. After the hypothesis is checked for each city, the cities where the hypothesis is true have the same relationship price to distance from the center, which was tested using an OLS linear regression. This regression is used to identify if the relationship is linear.

The final stage of the analysis was then to test a couple of different ML models on both the data sets, combined with the cities and then for each city individually. The models used were linear regression and a random forest. For both models, it was decided to remove outliers using 1.5 times IQR, as there are a few very large outliers in each city.

Findings from Data Exploration:
The first step in data exploration for RQ3 was to identify the geographical-related features. The geographical-related features include city, neighbourhood, neighbourhood cleansed, neighbourhood group cleansed, latitude, longitude, and distance to the centre. Exploration analysis was then completed. First, missing values were identified in the neighbourhood, neighbourhood group cleansed, and price. Therefore, moving forward in the analysis, neighbourhood cleansed will be used for neighbourhood values as it has no missing data, and missing price data will need to be dealt with (method TBD will depend on analysis type). A correlation matrix was created, which identified a weak correlation between price and distance to the city center. Given this, future statistical tests will be completed, but it is unlikely this feature linearly relates to price in any way. Additionally, outliers are present in all the cities, as they were highlighted in the completion of boxplots by city and price. These outliers must be considered and appropriately dealt with for future analysis. Lastly, the average price by city was plotted, and larger cities appear to have a larger average price. Unfortunately, challenges were faced regarding obtaining and appending population data, and thus, it was decided not to proceed with a population analysis. As a result, cities were examined individually for hypothesis testing and ML analysis.

4 Dataset

We collected and combined datasets from the Inside Airbnb website, which relies on the contribution of various collaborators and partners. The dataset being combined will be all of the quarterly listing data over the last year for all 10 cities in Italy that Inside Airbnb encompasses. The dataset includes 75 columns and nearly 200,000 records. Two main prepossessing steps were taken after merging the datasets.

Firstly, the data was prepared for proper data exploration and understanding. This included accounting for missing values, which dropped columns such as description, bathrooms, bedrooms, calendar_updated, and name. Few other columns which were irrelevant to our objective were also dropped. Afterwards, columns in unusable types/formats were converted. This included converting columns with type String because of "%" or "$" symbols to float, encoding columns where its values are lists of items, and converting features to type dateTime where relevant. Lastly, features which are key elements to our objective were added or transformed. One of the key added features is "distance_from_town_hall", which was created from the raw data latitude and longitude columns.

Secondly, the data was preprocessed for final analysis and model development. The overall goal of this step was to greatly reduce the number of features and encode columns for effective analysis and prediction. Any remaining columns with many missing values were dropped if there were many, or imputed using cluster median if there were few. Any categorical column which is a binary indicator but of type float were converted to 1s and 0s. Feature reduction then took place in three steps; removing features based on our intuition of its relevance to our objective, through a correlation matrix analysis, and using a random forest to determine feature importance. With our final set of features, categorical variables were encoded using one-hot-encoding or ordinal encoding as deemed appropriate.

5 Experiments and Results

5.1 Experimentation and Results for RQ1:

The analysis aimed at quantifying the impact of Airbnb listing features on the average review scores across multiple models. The Support Vector Machine (SVM) model demonstrated superior predictive performance with an accuracy of 78.80% within the defined threshold of $\pm$ 0.12 from the actual average review score, surpassing Neural Network (70.08%), Linear Regression (74.74%), and XGBoost (76.85%). Specifically, the SVM achieved a Mean Squared Error (MSE) of 0.0103, a Root Mean Squared Error (RMSE) of 0.1014, and an R-squared (R^2) value of 0.0486. These metrics underline its effectiveness in predicting user ratings amid the data's complex and possibly nonlinear relationships.

Linear Regression also exhibited commendable performance, though it did not reach the accuracy level of SVM. It recorded an R^2 value of 0.0453, an RMSE of 0.1016, and an MSE of 0.0103, indicating a reasonable fit for the variability in user ratings. These results suggest moderate predictive power, capturing the general trends in the data but potentially missing finer nuances influenced by less obvious listing features. The Neural Network model, tailored for handling complex data relationships, resulted in an RMSE of 0.1040 and an MSE of 0.0108. Despite its sophisticated architecture, the model faced challenges in fully capturing the subtleties of the data, as reflected in its modest R^2 and accuracy metrics. This underscores the intricate nature of user ratings and the difficulty in modelling such data with high accuracy. XGBoost, known for its efficiency with structured data, achieved an RMSE of 0.0997 and an MSE of 0.0099, marking predictive solid performance. However, its accuracy slightly trailed behind SVM. The model's ability to generalize well is evident from these results, suggesting it could be a strong contender with further tuning.

A detailed residual plot for the SVM model is shown in Figure 5, highlighting the heteroscedasticity of residuals as the predicted values increased. This variance suggests that the model's errors are inconsistent across all levels of user ratings, indicating potential areas for model refinement, such as advanced preprocessing or feature engineering.

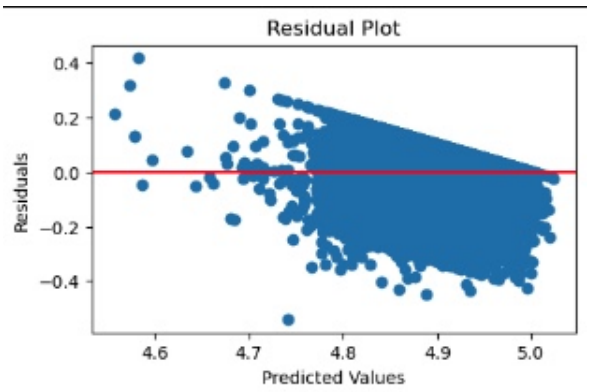

Fig. 5: Residual plot from the SVM model, highlighting the disparities between actual and predicted user ratings.

This plot provides crucial insights into the SVM's performance by illustrating discrepancies between actual and predicted user ratings. Ideally, residuals would cluster around the zero line; however, the plot shows a diverging trend as predicted values increase, indicating heteroscedasticity where prediction errors vary across different values. This fanning out of residuals suggests increasing errors with higher user ratings and highlights potential areas for SVM model refinement, such as advanced data preprocessing or feature engineering to stabilize variance across predictions. Despite these issues, the model's current accuracy demonstrates its potential for making reliable predictions.

Although no model achieved perfect accuracy, the SVM model's performance was particularly notable, predicting user reviews within 97.6% (0.12 threshold) at 80% accuracy. The presence of heteroscedasticity suggests that prediction errors vary with user rating levels, highlighting opportunities for model improvement. It's also important to note that trends might differ across various cities compared to the Italian data used in this study. Additionally, the substantial data volume in the training set posed a significant risk of overfitting. Nevertheless, the predictors showed considerable capability in estimating average review scores, albeit not flawlessly.

5.2 Experimentation and Results for RQ2:

The testing was designed to determine the influence of Airbnb host-related features on a pricing model. Employing a variety of predictive algorithms accounted for potential variance in results and increased the confidence of the conclusion. The three models used for testing were an XGBoosted Regressor, a Random Forest Regressor and a Stacked Regressor. The results are as follows:

Models	RMSE	R^2	MAE
XGBoost (without host)	42.028	0.309	31.8
XGBoost (with host)	41.437	0.328	31.3
Random Forest (without host)	42.348	0.298	32.2
Random Forest (with host)	41.708	0.320	31.7
Stacked (without host)	41.792	0.316	31.5
Stacked (with host)	41.081	0.320	30.9

TABLE 1: Model Performance Metrics for Pricing Prediction

Feature importance was again collected to verify whether the model utilized the host features.

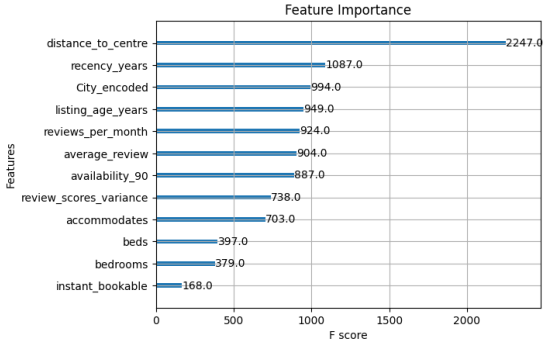

Fig. 6: Feature Importance from XGBoosted Model without Host Related Features.

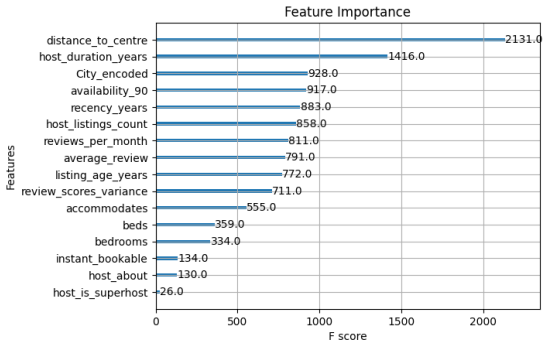

Fig. 7: Feature Importance from XGBoosted Model with Host Related Features.

The Feature Importance plots further display the significance of host-related attributes, with 'host duration years' and 'host listings count' showcasing significant F-scores, as predicted. This confirms that they impact pricing decisions. This would suggest that the host's credibility, experience, and activity matter.

The inclusion of host-related features seems to improve model performance modestly. This suggests that these features contain useful information for predicting prices not captured by the other features in the model. However, using these features better explains variance between similar listings, as we can see by improved R^2 while including them. Host-related features benefit the model's generalization ability and improve predictive performance.

Partial dependence graphs were used to display fully how these features individually affected pricing.

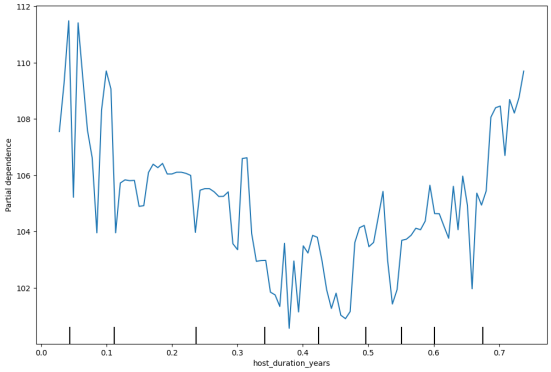

Fig. 8: Host Duration Compared to Price (Scaled)

The plot suggests a nonlinear relationship with price nonlinear duration. It shows that newer hosts tend to have higher prices, and then as the host has been using Airbnb for longer, their prices tend to drop. After a longer host duration, though, the price sharply increases again. While there is some obvious fluctuation, there is a mild parabolic trend, which may indicate that very new hosts and very experienced hosts tend to get higher listing prices.

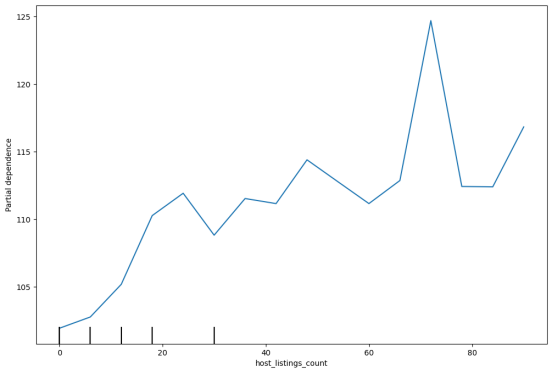

Fig. 9: Host Listings Compared to Price (Unscaled)

This host listing plot displays a different pattern. It is a much more linear pattern, with prices generally increasing as the number of listings a host manages goes up. This pattern continues up to a certain point, after which the relationship becomes more volatile. Thus, this implies that managing too many listings creates a complex effect on pricing. This could be due to hosts having to spread out their attention among too many properties at once.

5.3 Experimentation and Results for RQ3:

The first experiments completed were in the preprocessing stage. The key experiments completed to understand the data were a correlation matrix and box plotting price for the different cities. In the correlation matrix, aside from expected high correlations between longitude, latitude, and distance to the city center, no other high correlations are noted, signalling no presence of multicollinearity except for between the specific location features. Additionally, this signals that there is likely no strong linear relationship between price and location attributes.

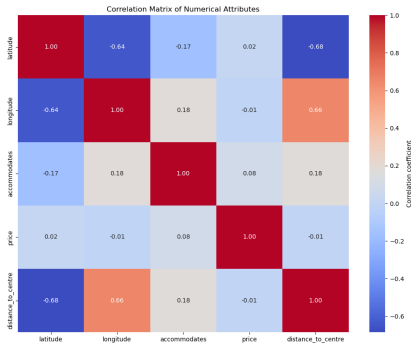

Fig. 10: Geographical Feature Correlation Matrix.

The box-plot and descriptive statistics for price by city highlighted price outliers in every city, some 600 times the third quartile value.

After the preprocessing experiments, the relationship between price and distance to the city center was examined. First, these price plots by distance to the city center were created to examine any immediate visual relationships. While not precisely clear and concise in multiple cities like Venice, Trentino, Rome and others, there is a small increase in the lowest price as the distance increases. However, this is minimal, and the plots do not show any major visual relationships.

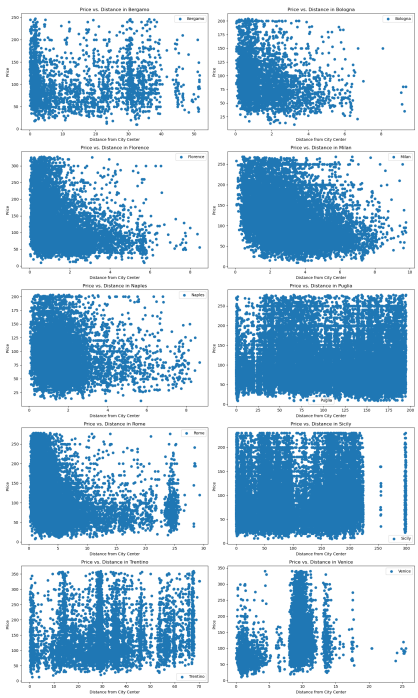

Fig. 11: Italian Cities Listing Price by Distance from City Center.

Next, the hypothesis "Does a change in distance from the city center impact price?" was tested for each city. This test was completed by first using the Kolmogorov-Smirnov normality test to determine whether the data for each city is normally distributed, then using a T-test if normal and a Wilcoxon signed rank test if not. The table below shows that every city used a Wilcoxon test, and every city had a P-value less than 0.05, signifying that the distance from the city center impacts price, as the Null hypothesis is rejected.

City	Test Used	Statistic	P-value
Bergamo	Wilcoxon	240,632	1.08×10^{-130}
Bologna	Wilcoxon	997	5.24×10^{-10}
Florence	Wilcoxon	3,247	2.17×10^{-15}
Milan	Wilcoxon	222,424	3.43×10^{-216}
Naples	Wilcoxon	15,515	2.25×10^{-17}
Puglia	Wilcoxon	179,160,582	0.00
Rome	Wilcoxon	1,329,696	0.00
Sicily	Wilcoxon	271,220,436	0.00
Trentino	Wilcoxon	5,895,668	3.29×10^{-284}
Venice	Wilcoxon	3,673,392	0.00

TABLE 2: Summary of Hypothesis Test Results by City

With these results that there is an impact, the linearity of the relationship was examined using an OLS regression. This yielded that the only cities with a statistically significant linear relationship were Bergamo, Naples, Sicily, Trentino, and Venice, with P-values below 0.05. Interestingly, Bergamo, Sicily, and Trentino had positive coefficients, signalling that as the distance from the city center increases, the price also increases. As expected, Naples and Venice have negative coefficients, signalling that increasing the distance from the city center decreases the listing value.

City	P-value	Coefficient
Bergamo	0.00026	2.83
Bologna	0.279	-4.22
Florence	0.107	-20.54
Milan	0.075	-9.48
Naples	0.00437	-7.19
Puglia	0.204	-0.08
Rome	0.104	-1.65
Sicily	0.00049	0.15
Trentino	0.00264	1.51
Venice	1.65×10^{-13}	-23.94

TABLE 3: Summary of OLS Linear Regression Test Results by City

Finally, after finding statistically significant relationships, some linear, regarding price distance, the final analysis uses all the geographically related features, developing a linear regression model and a random forest to see how well the combination of these features can predict price. Models were created for the data set as a whole and for each individual city, with outliers removed prior. Starting with the entire data set. The Linear Regression achieved an RMSE of 48.53 on the training data set and 1,343,085,089 on the test set. The random forest with a max depth of 6 and 500 estimators had a similar train RMSE of 46.84 but performed better on the testing data set with an RMSE of 48.27. For the individual cities, as shown in Table 3 below, the linear regression performs better than the combined dataset for most but still performs poorly for Bergamo, Milan, and Sicily; here, the Random Forest performs well for all of the individual cities similar to the combined data set with RMSE's between 36-48. This RMSE range represents only 13.6%-18.1% of the price range with the outlier removed, $265. Therefore, this can be considered a good error range, given that only geographical features are used. Additionally,

as Random Forest performs better, it can be inferred that the relationship between geographical features and price is primarily non-linear for most cities.

City	Linear Regression		Random Forest	
	RMSE Train	RMSE Test	RMSE Train	RMSE Test
Bergamo	44.52	4.54×10^{11}	43.98	47.66
Bologna	44.71	44.63	40.69	43.40
Florence	51.40	52.05	48.69	51.08
Milan	47.76	1.52×10^{12}	46.50	47.08
Naples	44.80	44.63	42.75	44.12
Puglia	49.48	1.92×10^{8}	48.90	49.29
Rome	48.69	48.38	46.10	46.86
Sicily	47.27	3.04×10^{10}	47.55	47.67
Trentino	54.24	54.17	48.41	50.44
Venice	48.71	48.80	46.84	48.27

TABLE 4: Comparison of RMSE for Linear Regression and Random Forest Models by City

6 Conclusion and Future Work

In conclusion, our project has addressed the need for sophisticated, adaptive pricing models in Italy's short-term rental market, considering both quantitative and qualitative listing details from the Airbnb website. By integrating a wide range of factors and predicting customer review ratings, we have refined pricing strategies for hosts, providing accurate price estimations for guests, and enhanced overall comprehension of market dynamics in this culturally and historically rich region.

For RQ1, the analysis showed that certain features associated with Airbnb listings, such as property type, amenities, and house prices, significantly impact user ratings. The Support Vector Machine (SVM) model performed the best, with an accuracy of 78.80%. However, residual plots indicated the presence of heteroscedasticity, suggesting varying prediction errors across different user rating levels and room for model refinement. While the models could predict average review scores to some extent, the amenities were not able to be properly analyzed. For next steps, more data collection would be needed to further increase the accuracy and reliability of the model.

For RQ2, the host-related features significantly impact pricing in Airbnb listings. The analysis, conducted using XGBoosted Regressor, Random Forest Regressor, and Stacked Regressor models, showed that including host-related features improved model performance modestly. Attributes such as 'host duration years' and 'host listings count' were found to be significant, indicating that a host's credibility, experience, and activity influence pricing decisions. The relationship between host duration and price appears to be nonlinear, while the relationship between the number of listings a host manages and pricing is generally linear up to a certain point, beyond which it becomes more volatile. For next steps, we recommend further data exploration, particularly focusing on outcome metrics such as booking frequency, guest satisfaction, and host revenue. This will help determine if there are additional data dependencies on rental prices.

Finally for RQ3, geographical-related features, such as city, region, does not have a significant impact on the pricing of Airbnb listings in Italy. However, statistical analysis revealed that the distance from the city center significantly affects the price of listings, with a nonlinear relationship observed in most cities. Linear regression and random forest models incorporating geographical features predicted prices reasonably well, with random forest performing better, indicating a primarily nonlinear relationship between geographical features and price. For our next steps, we aim to incorporate population data and segment it into housing types or occupancy sizes to enable more direct comparisons. This will enable us to calculate the distance to the city centres instead of estimating it, leading to a higher accuracy. In summary, our project not only provides valuable insights into the factors influencing pricing and user ratings in Italy's Airbnb market but also underscores the need for ongoing model refinement to ensure accurate and reliable pricing predictions, thereby contributing to enhanced experiences for both hosts and guests in this vibrant tourism destination.

7 Group Member Contributions

Name	Responsibilities and Contributions
Ben Look	Responsible for research on "The hedonic price model of the online short-term rental market based on machine learning." Managed RQ3 and report formatting. Conducted data exploration for RQ3, including missing value exploration, city-to-price descriptive statistics, correlation analysis for geographic features and price, and geographical feature visualizations. Completed entire RQ3 analysis and conclusion.
Jeremy Gonsalves	Conducted sensitivity analysis and machine learning predictions for Airbnb price analysis, managed Motivation, Problem sections, and report formatting. Performed data exploration for RQ1, including missing value analysis, amenities, pricing, and house features' impact on reviews. Engaged in feature engineering and predictive model development to analyze average review scores. Also handled report formatting.
John Turnbull	Researched reasonable price recommendations on Airbnb using multiscale clustering. Also assisted and performed data exploration, feature engineering, and model development for RQ1.
Matthew Morano	Researched Airbnb Pricing without using amenity-driven features, an alternative approach to modelling pricing. Managed Research Question 2 (RQ2). Also explored host-related features and their relation to pricing, creating code for predictive models to understand the impact of host-related features on pricing models (RQ2 analysis).
Jared Simpson	Focused on research on Airbnb Pricing, emphasizing the impact of social-related features. Also responsible for explaining the dataset and supporting the development of research questions.

TABLE 5: Team Member Responsibilities and Contributions

8 Replication Package

CISC351_Group_Assignment submitted on OnQ

REFERENCES

[1] I. Ghosh, R. K. Jana, and M. Z. Abedin. An ensemble machine learning framework for airbnb rental price modelling without using amenity-driven features. International Journal of Contemporary Hospitality Management, 2023.

[2] C. Gibbs, D. Guttentag, U. Gretzel, J. Morton, and A. Goodwill. Pricing in the sharing economy: a hedonic pricing model applied to airbnb listings. *J. Travel Tour. Mark.*, 35:1–11, Apr. 2017.

[3] P. Liu. Airbnb price prediction with sentiment classification. In *Master of Science, San Jose State University*, 2021.

[4] J. Xin and L. Xue. The hedonic price model of online short-term rental market based on machine learning. In *Proceedings of the 7th International Conference on Cyber Security and Information Engineering*, ICCSIE '22, pages 972–976, New York, NY, USA, Oct. 2022. Association for Computing Machinery.

[5] T. Y. Y. Li, Q. Pan and L. Guo. Reasonable price recommendation on airbnb using multiscale clustering. In *Chinese Control Conference (CCC), 2016 35th Chinese, pp. 7038–7041, IEEE*, 2016.