Birgit Preuss, Luke Jarvis, Mengsrun Nit

CSC-272 Final Project Report

April 28, 2025

Dr. Treu

# Airbnb Price Predictions

## Introduction

As short-term rental platforms like Airbnb have grown in popularity, questions about what drives the cost of a stay have become increasingly important for both hosts and guests. Property owners seek to optimize their listings for maximum profitability, while travelers aim to find accommodation that balances amenities and affordability. Understanding the underlying factors that influence rental price can offer valuable insights into both economic and consumer behavior in the short-term rental market.

Through exploration and analysis of real-world Airbnb listing data from Asheville, North Carolina, we sought to uncover patterns between property characteristics, availability, amenities, and nightly rental rates. By applying data preprocessing techniques and variability of machine learning models, we aim to identify which features most strongly predict price and evaluate the effectiveness of different predictive modeling approaches. Insights from this research could help inform pricing strategies for hosts, boosting user experience for travelers, and contribute to a broader understanding of trends within the sharing economy.

## Dataset Description

The Airbnb listing dataset is based on real-world data collected from thousands of rental properties across Asheville, North Carolina. It includes detailed information gathered from public Airbnb listings, sourced from Inside Airbnb (InsideAirbnb.com). The Airbnb listing dataset

focuses on a wide range of attributes such as property characteristics, host details, listing description, and pricing information. Attributes available in the dataset include host demographics, property type, room type, number of accommodations, availability, and review-based rating related to cleanliness, communication, and values.

In addition to the core listing attribute, the dataset contains information about the various amenities such as Wi-Fi, kitchen facilities, televisions, and outdoor space. Amenities were originally recorded as text strings and later processed into binary attributes to facilitate predictive modeling. Prices for each listening are provided as a continuous numeric value representing the cost per night, with additional variables capturing review frequency and host activity levels.

## Data Preprocessing

To prepare our dataset for analysis, we began by selecting only the most relevant attributes. The original dataset was quite large and contained many fields that were not useful for predictive modeling, such as URLs, names, and other descriptive text fields. We manually reviewed the attributes and retained those we believed would be most predictive for our goals. This step helped streamline our dataset and focus on more meaningful variables.

During the initial exploration, we noticed that the dataset contained special characters, including symbols like "$", which could interfere with modeling. To address this, we used R to clean the data and remove unwanted characters. This preprocessing step ensured that the dataset contained only clean, standardized text and numerical values suitable for further analysis. Another issue we identified was missing data. Approximately 10% of the instances in the dataset had missing values across important attributes. Rather than attempting imputation, we decided to remove these instances entirely. While this reduced the total number of available examples, it helped preserve the overall quality and reliability of the remaining data.

Given the large size of the original dataset—around 5,600 instances—we further reduced the number to 2,500 instances to make the dataset more manageable for processing in Weka. This decision balanced computational efficiency with maintaining enough data for meaningful analysis and model training. In addition, one of the attributes, "property type," originally had too many unique categories, making it difficult to use effectively in modeling. To simplify this attribute, we consolidated the numerous categories into five broader groups. This process helped reduce noise and improve the interpretability of the resulting models. Initially, we considered combining the main dataset with two other datasets: a reviews dataset and a calendar dataset. However, after exploration, we found that merging the datasets would be too complicated due to multiple instances with the same IDs and the overall size of the combined data, which would have made it challenging to manage in Weka. That being said, we decided to keep the main dataset separate.

Finally, after splitting and loading the data into Weka, we noticed that the "amenities" column was treated as a string attribute. To better utilize this information, we decided to apply Weka's "String to Word Vector" filter to convert the amenities text into a set of binary attributes representing the presence or absence of specific amenities. We then selected the most predictive amenity attributes for modeling.

## Data Attributes

Once we preprocessed the data, we decided to treat the amenities attribute with the class attribute, price, as its own dataset, and kept our original 19 attributes plus class attribute, price, as the original dataset. Below is a table of all our attributes for both datasets and their values.
Table with the original dataset without the amenities attribute:

| Attributes | Values |
|---|---|
| host_location (nominal) | Contains a list of 100 cities in which the hosts who own the rental properties in Asheville live. |
| host_is_superhost (boolean) | Count of how many of the Airbnb's are superhosts with 0 as no and 1 as yes. |
| host_acceptance_rate (numeric) | Percentage rate with values ranging from 0 to 100 |
| host_listings_count (numeric) | Number depicting how many listings a host may have. Numbers range from 1 to 1642 (an outlier), where the majority is 1. |
| host_total_listings_count (numeric) | Number depicting how many listings a host has had past and present. Numbers range from 1 to 3224 (an outlier), where the majority range from 1 to 3. |
| neighbourhood_cleaned (numeric)<br><br>Note: should have been treated as nominal. | Neighborhoods in Asheville grouped into 8 categories by zip code.<br>- 28704<br>- 28715<br>- 28732<br>- 28801<br>- 28803<br>- 28804<br>- 28805<br>- 28806 |
| room_type (nominal) | The type of room the Airbnb is, divided in 4 categories.<br>- Private room<br>- Entire home/apt<br>- Shared room<br>- Hotel room |
| availability_365 (numeric) | Number of days the property was available for booking throughout a 1 year period.<br>Numeric values ranged from 0 to 365 with a mean value of 199. |
| accommodates (numeric) | Number of people the property accommodates. Values ranged from 1 to 16 with a mean value of 4.5. |
| bathrooms (numeric) | Number of bathrooms a property has. Values range from 0 to 9 where the mean value was 1.5. |

| | |
|---|---|
| number_of_reviews (numeric) | Number of reviews a single property has. Values range from 1 to 1418 where the mean value 173. The majority of the values lie between 1 and 709. |
| review_scores_rating (numeric) | Scores range from 1 to 5 where the mean lies at 4.9. |
| review_scores_accuracy (numeric) | Accuracy of the reviews signify how well the listing and photos match the property. These range from 1 to 5 where the mean lies at 4.9. |
| review_scores_cleanliness (numeric) | This attribute signifies the cleanliness of the property. These range from 1 to 5 where the mean lies at 4.9. |
| review_scores_checkin (numeric) | This attribute signifies how the check-in process was. These range from 1 to 5 where the mean lies at 4.9. |
| review_scores_communication (numeric) | This attribute signifies how well the host communicated with the guests. These range from 1 to 5 where the mean lies at 4.9 |
| review_scores_location (numeric) | This attribute signifies the rating of the property's location. These range from 1 to 5 where the mean lies at 4.9 |
| review_scores_value (numeric) | This attribute signifies how well the price reflects the property and its amenities. These range from 1 to 5 where the mean lies at 4.8. |
| property_category (nominal) | This attribute shows what type of property it is in 5 categories.<br>- Private_room<br>- Entire_place<br>- Unique_stay<br>- Hotel_like<br>- shared_room |
| price (numeric) | The price attribute signifies price per night and has a minimum of 18 and maximum of 2043 (an outlier) where the mean lies at about 148. |

Table 1: Attributes and values for Airbnb price prediction.

Below is a table of all the amenities we selected that had over 200 instances, not including those which we removed that we believed did not make sense without context. All of them (except for

price, which is defined in the table above) are booleans with 0 being no and 1 being yes the property has that amenity.

Table of amenities:

| | **Attributes** |
|---|---|
| | Amazon |
| | BBQ |
| | Backyard |
| | Bathtub |
| | Bed |
| | Blender |
| | Books |
| | Carbon |
| | Ceiling |
| | Cleaning |
| | Coffee |
| | Crib |
| | Dining |
| | Dishes |
| | Dishwasher |
| | Dryer |
| | Ethernet |
| | Exterior |
| | Fire |
| | Freezer |
| | Garden |

| | |
|---|---|
| | HDTV |
| | Hangers |
| | Hulu |
| | Iron |
| | Keurig |
| | Keypad |
| | Kitchen |
| | Laundromat |
| | Lockbox |
| | Microwave |
| | Netflix |
| | Outdoor |
| | Oven |
| | Patio |
| | Refrigerator |
| | Roku |
| | Room-darkening |
| | Shampoo |
| | Shower |
| | Stove |
| | TV |
| | Toaster |
| | Washer |
| | Wifi |
| | air |

| | |
|---|---|
| | alarm |
| | balcony |
| | blankets |
| | cable |
| | cameras |
| | closet |
| | dresser |
| | driveway |
| | extinguisher |
| | games |
| | grill |
| | kettle |
| | parking |
| | patio |
| | pillows |
| | play/travel |
| | security |
| | storage |
| | tub |
| | view |
| | price |

Table 2: Amenities attributes for Airbnb price prediction.

The further analysis of key attributes within our data can be seen below. These graphs curated within Tableau give us insight into how certain attributes like Property Type, Price, Reviews and Availability interact, therefore, effecting listing prices of properties on Airbnb within Asheville, NC.

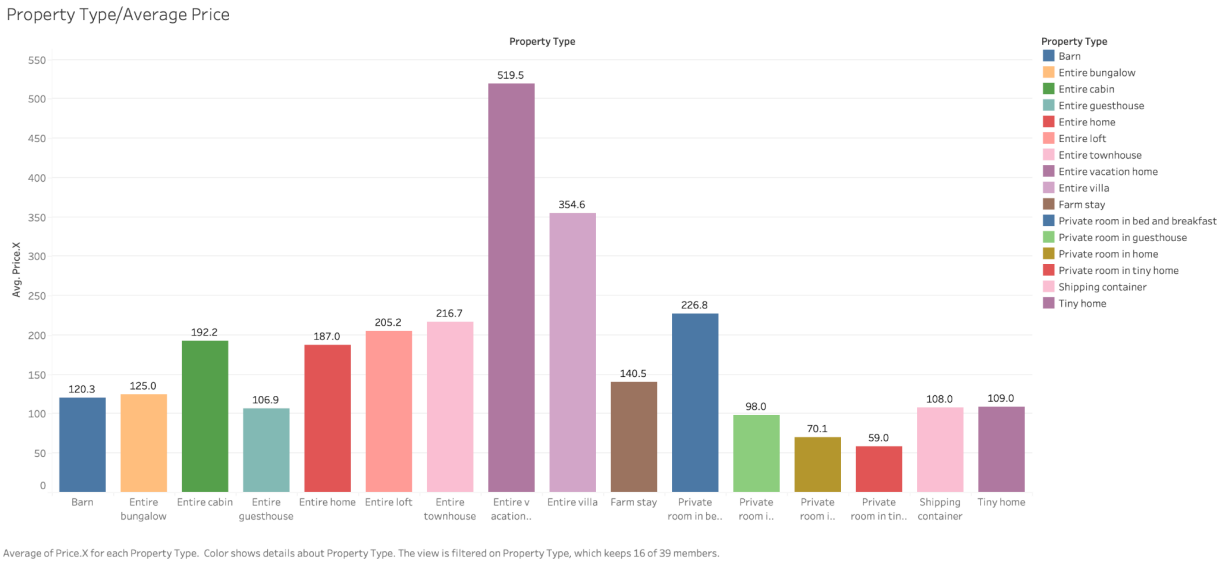


Chart 1: Depicts Average Price across the main Property Types. A key aspect of this analysis is to recognize that entire properties tend to be more expensive on average compared to single rooms/apartment-style properties.

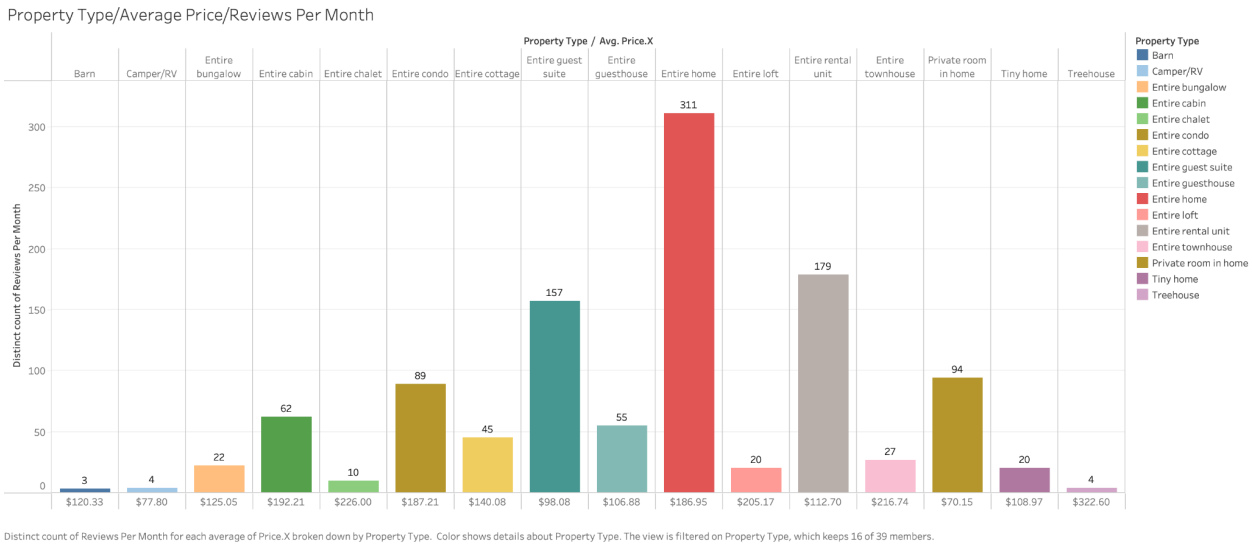


Chart 2: Depicts Average Price across Property Types, factoring in Reviews Per Month of those properties. This takes into account the reliability that reviews give future customers. When shopping for anything online, reviews are typically a strong indicator of trust and reliability of the product. In this case, the product being an Airbnb property. As we can see, the properties with more reviews on average tend to be more expensive on average.

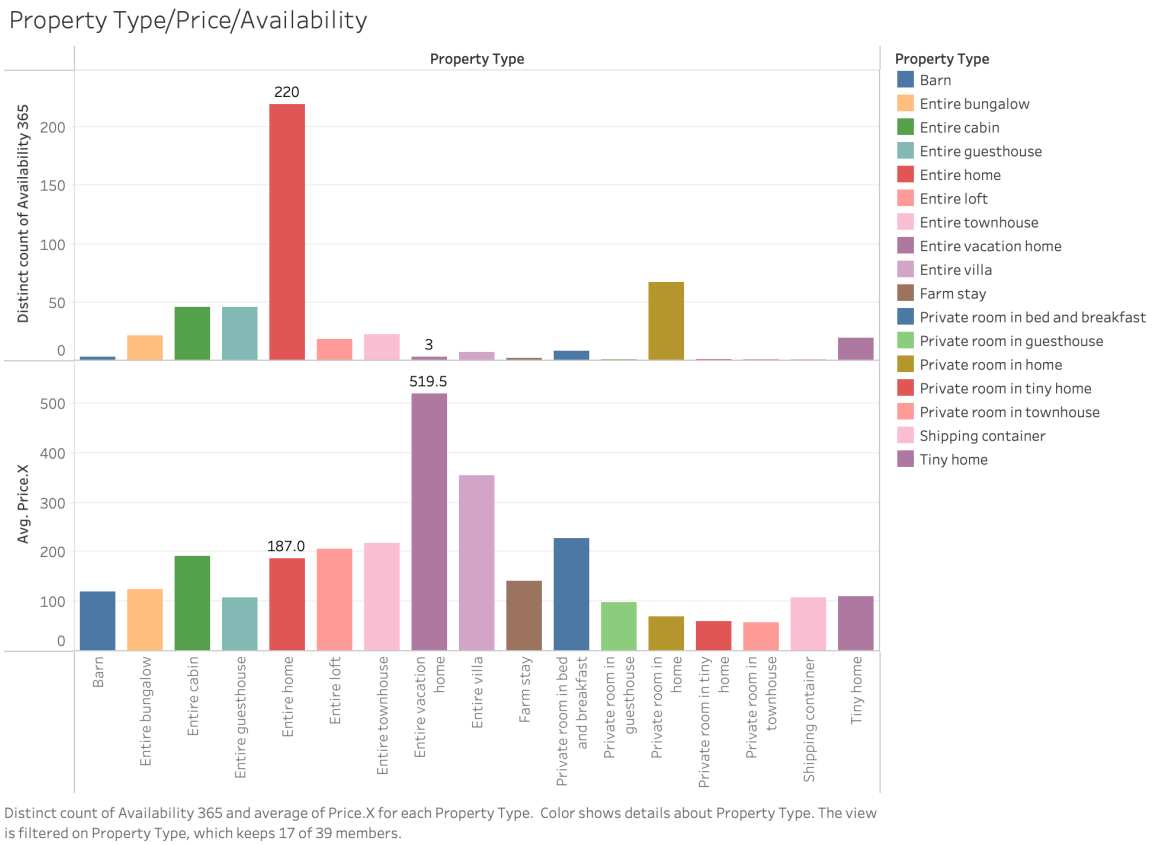


Chart 3: Depicts Price across Property Types while considering the availability of the properties. When considering simple supply and demand, the less available a product is, the more expensive said product tends to be on average. In this case, Airbnb has a much higher volume of entire homes listed online compared to rarer vacation homes in locations that are harder to come by. Therefore, vacation homes tend to cost more on average per night compared to the readily available entire homes. This analysis further leads us to analyze how location affects the overall price of Airbnb properties.

## Data Analysis

Given that our class attribute is price which is numeric we did numeric estimation on both of our datasets. In our data analysis, we experimented with several algorithms, including Gaussian Processes, Linear Regression, Support Vector Regression, Nearest Neighbor, M5 Rules, and M5P. First, we split both datasets into testing and training sets, of which the testing set had one-third of the datasets, with 833 instances, leaving our training datasets with 1667 instances.

### Gaussian Processes

The Gaussian Processes algorithm is a supervised learning algorithm that can be used for regression and classification. The model is defined by the mean and covariance functions. These functions are then used to make a prediction on new data. It returns the distribution over several possible output values. The Gaussian Processes algorithm is also able to provide uncertainty estimates for the predictions.

### Linear Regression

Linear Regression works by finding a line of best fit that best represents the data. Then the minimum distance between every datapoint and the line is calculated; this distance is the squared error between the predicted and actual values. It is a relatively simple algorithm and one that is primarily chosen when doing numeric estimation, and therefore was perfect for our experimentation. We also used the Simple Linear Regression classifier in Weka for the amenities dataset, which picks the most predictive attribute and does linear regression only on that attribute's values.

### Support Vector Regression

The Support Vector Regression algorithm (SMOreg) is very similar to linear regression, however, it calculates a maximum margin from the hyperplane and all the data points that fall within the maximum margin are not considered errors as they are in linear regression. Any datapoints outside that space are support vectors. Once again, this algorithm is very good for numeric estimation and could be considered better than linear regression.

### Nearest Neighbor

The nearest neighbor algorithm (IBK), when doing numeric estimation, finds the closest instance to the one one is trying to predict by calculating the means of the instances. In our project, we experimented with K=1, K=3, and K=5. These different neighborhood sizes determine how many nearest neighbors are considered when classifying that new data point. In our analysi,s we found K = 1 to be the best size.

### M5Rules and M5P

The M5Rules algorithm generates a set of rules by building a model tree using regression formulas on the attributes. It then selects the best nodes and makes them into a set of rules. It does this by using multivariate linear models instead of constant values, giving the rules and the tree more complexity.

### Results

Once we ran our datasets through all the algorithms, we got correlation coefficients and the mean absolute error, showing us which algorithms did the best job of classifying our dataset. Below is a table of each correlation coefficient and mean absolute error for each algorithm for our original dataset without amenities.

| Algorithm | Correlation Coefficient | MAE |
|---|---|---|
| Gaussian Processes | 0.68 | 48.2 |
| Linear Regression | 0.687 | 48.1 |
| Support Vector Regression | 0.682 | 44.61 |
| Nearest Neighbor K=1 | 0.764 | 22.57 |
| Nearest Neighbor K=2 | 0.6726 | 42.05 |

| | | |
|---|---|---|
| Nearest Neighbor K=3 | 0.6629 | 45.15 |
| M5Rules | 0.6432 | 48.14 |
| M5P | 0.6314 | 48.27 |

Table 3: Data analysis results for Airbnb price predictions.

When assessing the correlation and mean absolute erro, it is clear that the nearest neighbor with a K=1 has both the highest correlation coefficient and lowest MAE. The second-best model appears to be Support Vector Regression with a correlation coefficient of .682 and a low MAE of 44.61. This goes in line with what we expected after analyzing how SVR works.

The linear regression formula weka output included only 15 out of our 19 total attributes to predict price. It did not include host_acceptance_rate, host_total_listings_count, neighborhood_cleansed, and availability_365. This signifies that these attributes were not predictive of price. We expected availability_365 to be very predictive, so we were surprised to see that. On the other hand, the M5Rules algorithm output 4 rules, of which every attribute was in at least one rule.

Below is the tree that the M5P algorithm outputs. This algorithm only chose bathrooms and review_scores_accuracy to be predictive of price.

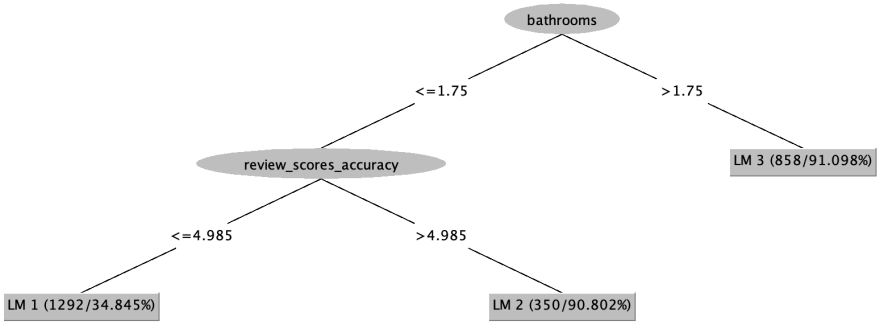


Figure 1: M5P model tree.

We then ran the same algorithms, plus simple linear regression on our second dataset, which is made up of the amenities in the Airbnbs. Just as in Table 1, we analyzed the correlation coefficient.

| **Algorithm** | **Correlation Coefficient** |
|---|---|
| Gaussian Processes | 0.5675 |
| Linear Regression | 0.5713 |
| Simple Linear Regression | 0.4514 |
| Support Vector Regression | 0.5613 |
| Nearest Neighbor K=1 | 0.7388 |
| Nearest Neighbor K=2 | 0.6226 |
| Nearest Neighbor K=3 | 0.5781 |
| M5Rules | 0.512 |

| | |
|---|---|
| M5P | 0.5116 |

Table 4: Data analysis results for amenities in Airbnb price prediction.

This dataset appears to have much lower correlation coefficients with all models resulting with a correlation coefficient of about 0.5. Once again, a neighborhood of K=1 appears to have the most predictive value, however K=5 is more in line with the rest of the models making that the more perforable neighborhood size.

The linear regression formula output includes 29 out of the 71 attributes to predict price as we can see below in figure 2. Out of those the simple linear regression formula found Dishwasher to be the most predictive for price.

```
price =

     35.3739 * Carbon +
     18.9485 * Crib +
    -40.3094 * Dishes +
     44.9347 * Dishwasher +
     13.0016 * Dryer +
     19.6831 * Exterior +
    -27.1379 * Garden +
     14.3741 * HDTV +
    -15.0915 * Iron +
     13.1962 * Keypad +
     15.6857 * Kitchen +
    -16.6646 * Shampoo +
     26.2105 * TV +
     18.3628 * Toaster +
     16.1549 * Wifi +
    -63.8029 * alarm +
    -24.0461 * backyard +
     25.8273 * cable +
    -21.3656 * closet +
     10.5712 * coffee +
    -29.8719 * crib +
    -26.384  * driveway +
     34.4128 * games +
     18.2927 * grill +
    -52.9151 * parking +
    -16.4444 * patio +
     49.1383 * play.Travel +
     45.3776 * tub +
     39.0947 * view +
    176.7435
```

Figure 2: Linear regression formula for Airbnb price prediction.

The M5Rules model output 2 rules, including 40 of the 71 attributes. As we can see from the M5P model tree (Figure 3), an Airbnb providing shampoo proved to be the most predictive of price, which we found to be very strange and possibly a weird correlation.

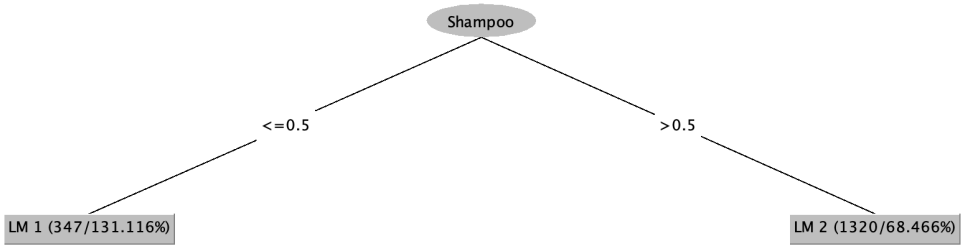


Figure 3: Model tree for Airbnb price prediction using amenities.

## Conclusion

In this project, we found that the Nearest Neighbor model(IBk) with K=1 provided the highest predictive accuracy for both the original dataset and the amenities dataset. For the core property attributes, Nearest Neighbors achieved the strongest correlation coefficient of 0.764 and the lowest mean absolute error of 22.57. This performance clearly surpassed other models such as Linear Regression, Support Vector Regression, and M5rules. These results suggest that Airbnb rental prices are highly influenced by localized property features rather than broader general attributes.

In the amenities dataset, the predictive performance was lower overall. Although Nearest Neighbor with K=1 again achieved the highest correlation coefficient of 0.7388, amenities alone proved to be weaker predictors of price compared to the core structural features. Certain amenities, such as dishwashers and shampoo, showed significance in some models. However, these findings are likely the result of indirect correlation rather than strong direct influence of pricing.

Overall, this study shows that structural property characteristics are more reliable indicators of nightly rental price than amenities alone. Models like Nearest Neighbor and Support Vector Regression are effective tools for identifying key patterns in rental pricing. Future research that incorporates dynamic pricing trends, detailed location information, and guest review analysis could further strengthen predictive models and provide deeper insight into the short-term rental market.