The University of Texas at Austin

Academic Year 2023/24 Spring Semester

ECO 395M: Data Mining and Statistical Learning

The Dollars and Sense of Staying in NYC: Predicting Airbnb Prices

Jinming Li, Fan Ye, Xiangmeng Qin

Word Count: 1822 **Page Count:** 15

i. Abstract:

Our report investigates the factors influencing Airbnb rental prices and neighborhood classification in New York City. Utilizing a dataset from 2019, we applied various statistical and machine learning methods, including Ordinary Least Squares (OLS) regression, backward selection, LASSO, and Classification and Regression Trees (CART), to identify the relationships between rental attributes and pricing. Our findings indicated that lower-priced listings are predicted with greater accuracy, while predictions for higher-priced listings were less precise. The models also revealed key attributes such as room type, location, accommodation size, and availability as significant in determining price.

ii. Introduction:

Background

Airbnb has changed the way people stay in New York City, offering a variety of places to rent. In 2019, a dataset was shared that lists many details about these rentals, like how much they cost, where they are, and what kind of place they offer. Understanding these details is important for people who rent out their homes to make better decisions and for city officials who make rules about housing.

Purpose of Analysis

This study has two main goals. The first is to figure out what makes some Airbnb places cost more or less than others. We think that where the place is located and what type of place it is (like an entire house or just a room) are important factors. The second goal is to see if we can guess which neighborhood an Airbnb is in based on things like its price and what it offers. This information can help hosts set fair prices and help travelers choose where to stay.

Method of Analysis

We picked different ways to study the Airbnb data to make sense of it more easily. These methods are like different tools in a toolbox—each one does something special. We use things like OLS regression, backward selection, LASSO, and CART models to spot trends and see how the features of an Airbnb, like how many rooms it has or where it is, affect its price. We make sure each tool works right by trying them out a lot. By looking at how different methods work and what they tell us, we figure out the key things that change an Airbnb's price and its spot in the city.

iii. Methods:

Data Set Description

Our data comes from a comprehensive collection of Airbnb listings in New York City for the year 2019. The dataset includes a variety of information on each listing, such as price, location, type of room offered, and several other details that guests might consider when choosing a place to stay. Specifically, the dataset covers factors like how many people a place can accommodate, the number of bedrooms and bathrooms, the kind of amenities provided, and user ratings. These details will help us explore and understand the trends and factors that influence Airbnb pricing and neighborhood classification in New York City.

Analytical Methods

To analyze this data, we will use several statistical and machine learning methods:

Descriptive Statistics and Visualization: We will start by summarizing the data using descriptive statistics and visualizations to understand the central tendencies, dispersion, and distribution of our main variables.

Ordinary Least Squares (OLS) Regression: This method will help us understand the relationship between the price of listings and other variables by estimating the extent to which different factors like location or room type affect the price.

Backward Selection: We'll use backward selection in our regression model to identify which variables are most important. This process starts with all variables and removes the least significant ones step by step to improve the model's performance.

LASSO (Least Absolute Shrinkage and Selection Operator): This technique is particularly useful when we have many variables. It helps to both improve the prediction accuracy and interpretability of the statistical model we develop by selecting only a subset of the provided features.

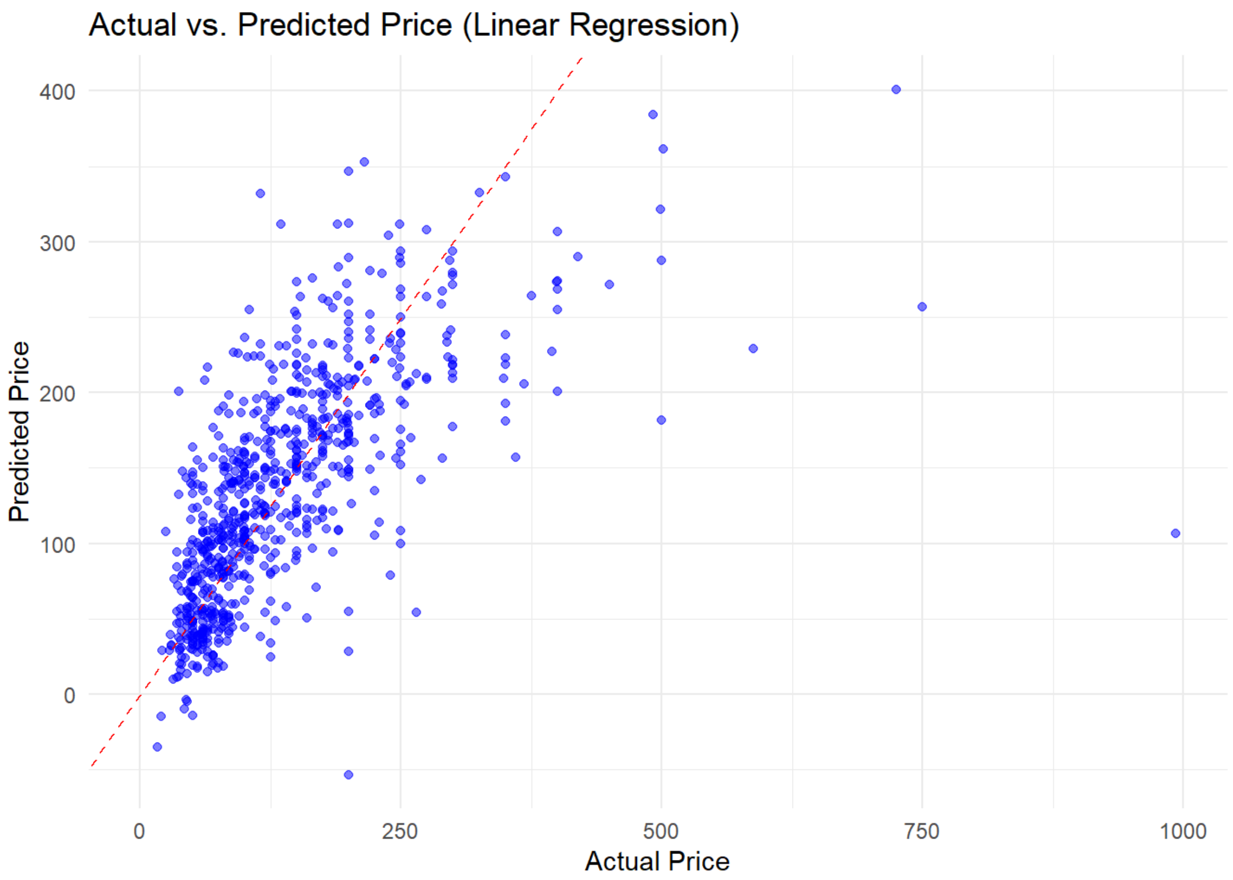
Classification and Regression Trees (CART): For the classification task of predicting neighborhoods, CART will be used. It's a decision tree algorithm that will allow us to classify listings into neighborhoods based on their characteristics.

For all models, we will employ cross-validation to ensure that our findings are robust and not merely tailored to a specific subset of the data. Additionally, we will compare models based on their predictive accuracy using metrics such as the Mean Absolute Error (MAE) for regression tasks and accuracy rate for classification tasks.

By applying these methods, we expect to uncover actionable insights that can inform Airbnb hosts on pricing strategies and provide a deeper understanding of the local Airbnb market structure.

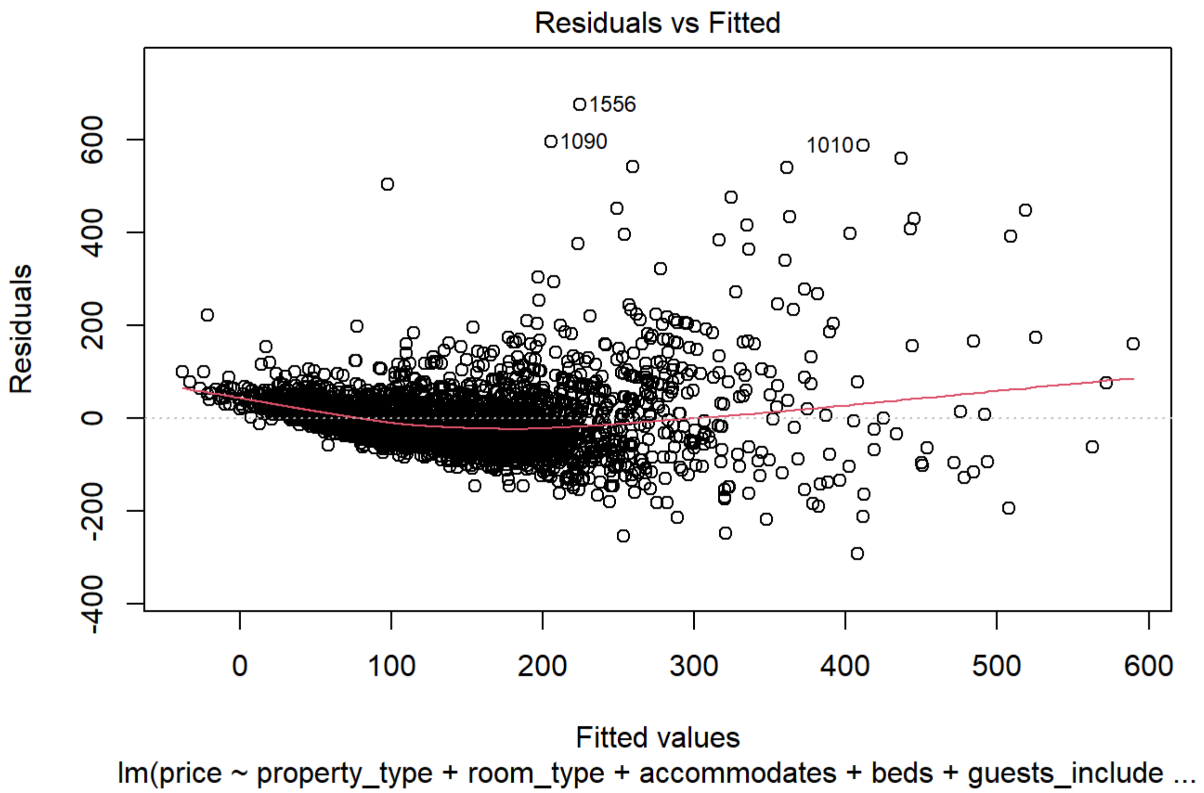
iv. Results:

1. Linear regression model:



Actual vs. Predicted Price Plot

The first plot compares actual prices to those predicted by the model. The plot shows a dispersion of points around the line, especially as the actual price increases, suggesting that the model predicts lower prices more accurately than higher ones. The points spread out as they move away from the line, indicating the model's predictions are less precise at higher price points.

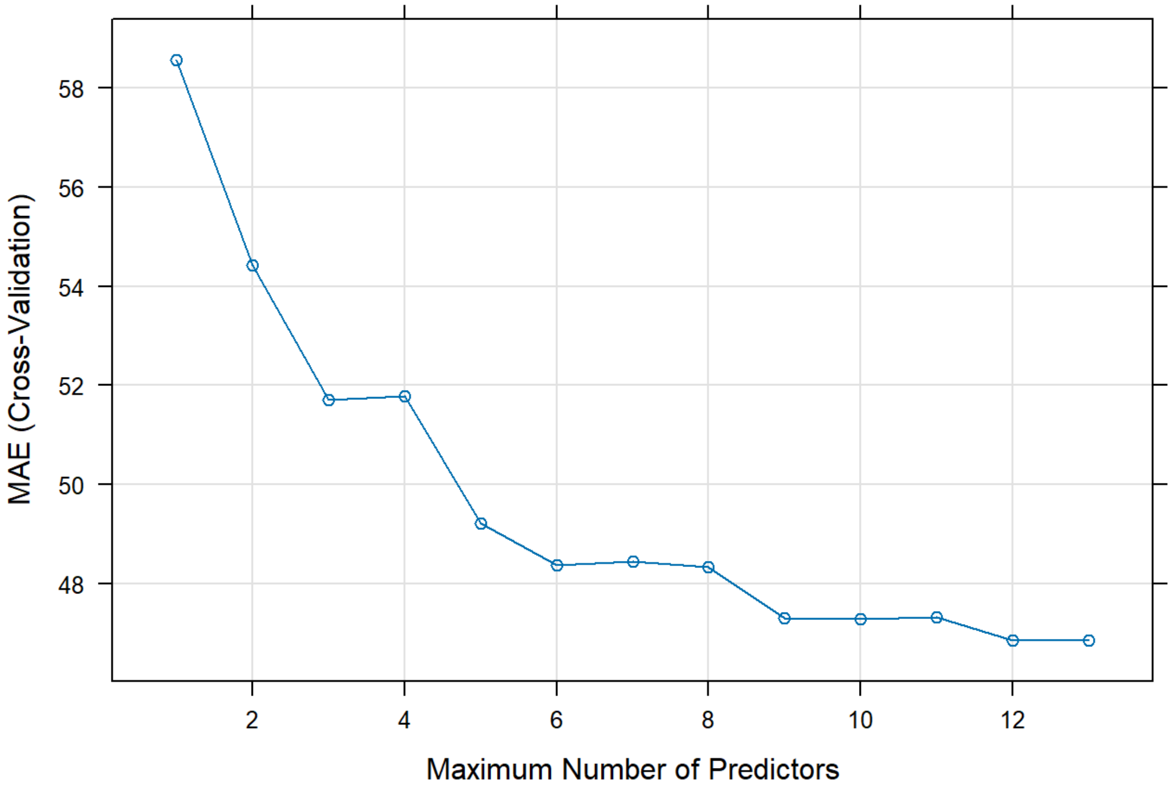


Residuals vs. Fitted Plot

The residuals plot against fitted values aims to detect non-linearity, unequal error variances, and outliers. Our plot shows a slight curve, suggesting potential non-linearity in the relationship between predictors and price. The residuals also fan out as the fitted values increase, implying an increase in the variability of predictions (heteroscedasticity).

RMSE: 71.2741

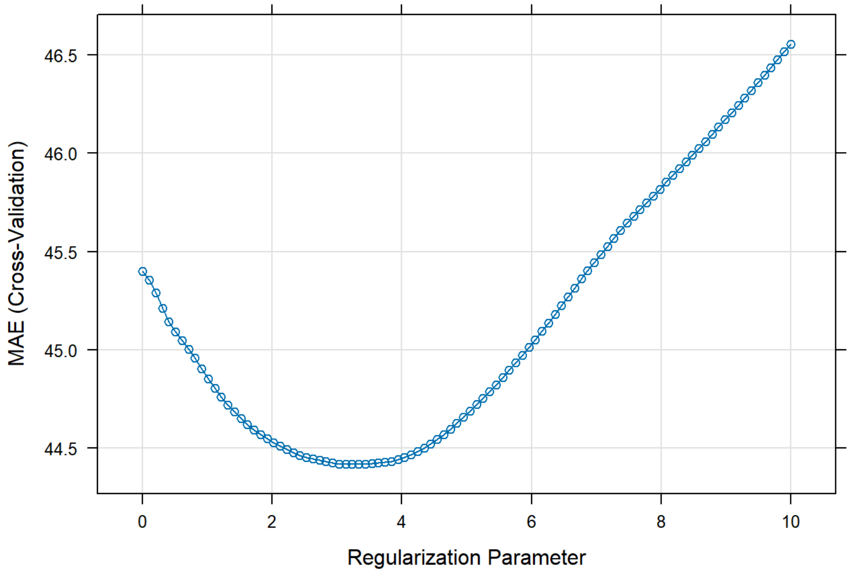
2. LeapBackward model:

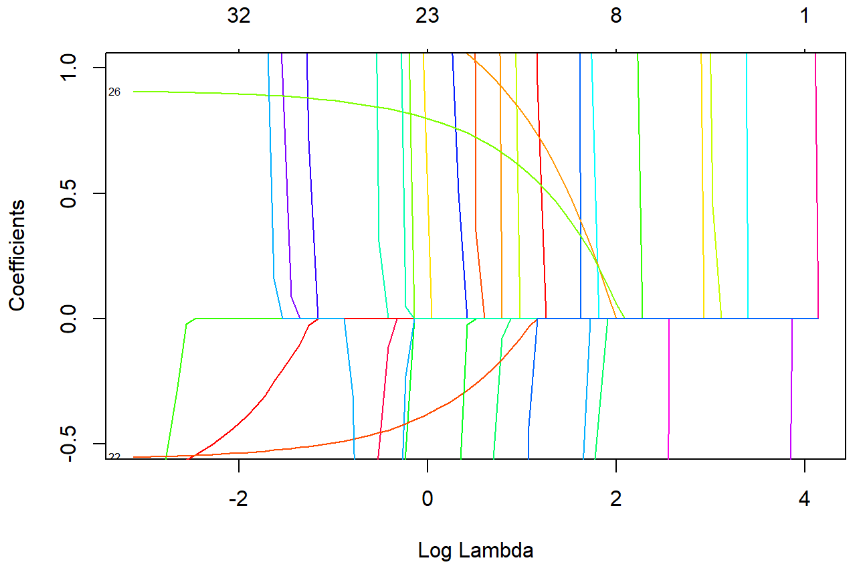


The line plot shows a sharp decrease in MAE as more predictors are introduced, from just one predictor to around five. This decrease suggests that the initial variables added significantly improve the model's ability to accurately predict Airbnb prices. As the number of predictors increases beyond five, the MAE reduction becomes more gradual, indicating diminishing returns on adding more variables to the model.

RMSE: 126.3288

3. LASSO model

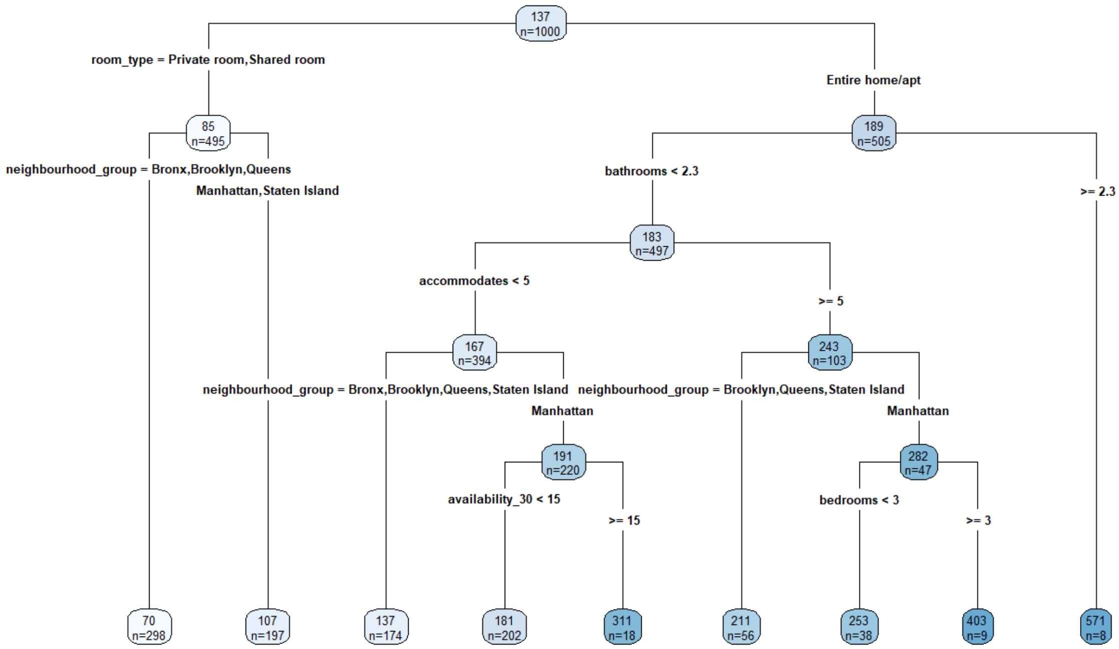




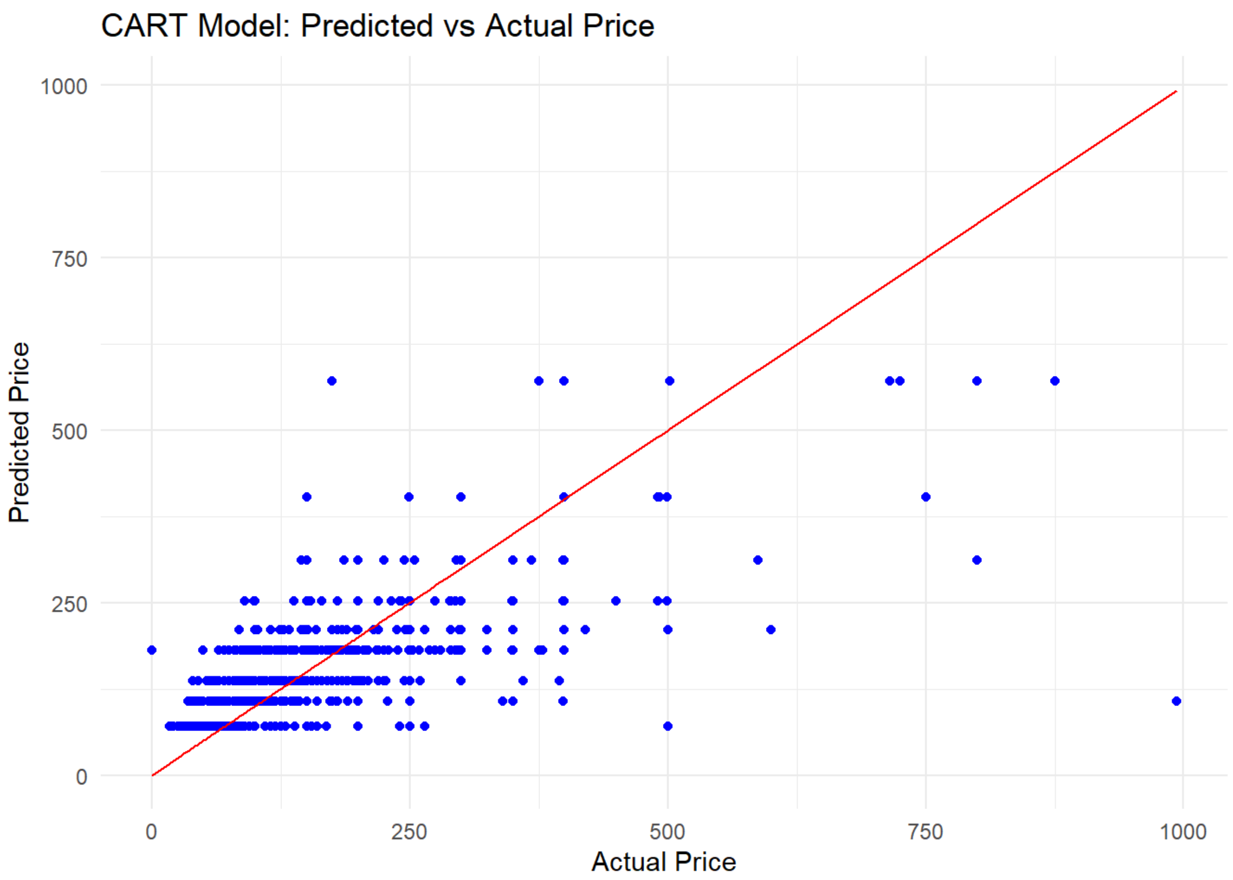
From the graph, we can observe that the MAE decreases as the regularization parameter increases from 0, reaching a minimum around a parameter value of 2, and then starts to increase again. The lowest point represents the optimal trade-off between bias and variance, indicating the most regularized model that still keeps predictive power.

RMSE: 118.9282

4. The CART model:



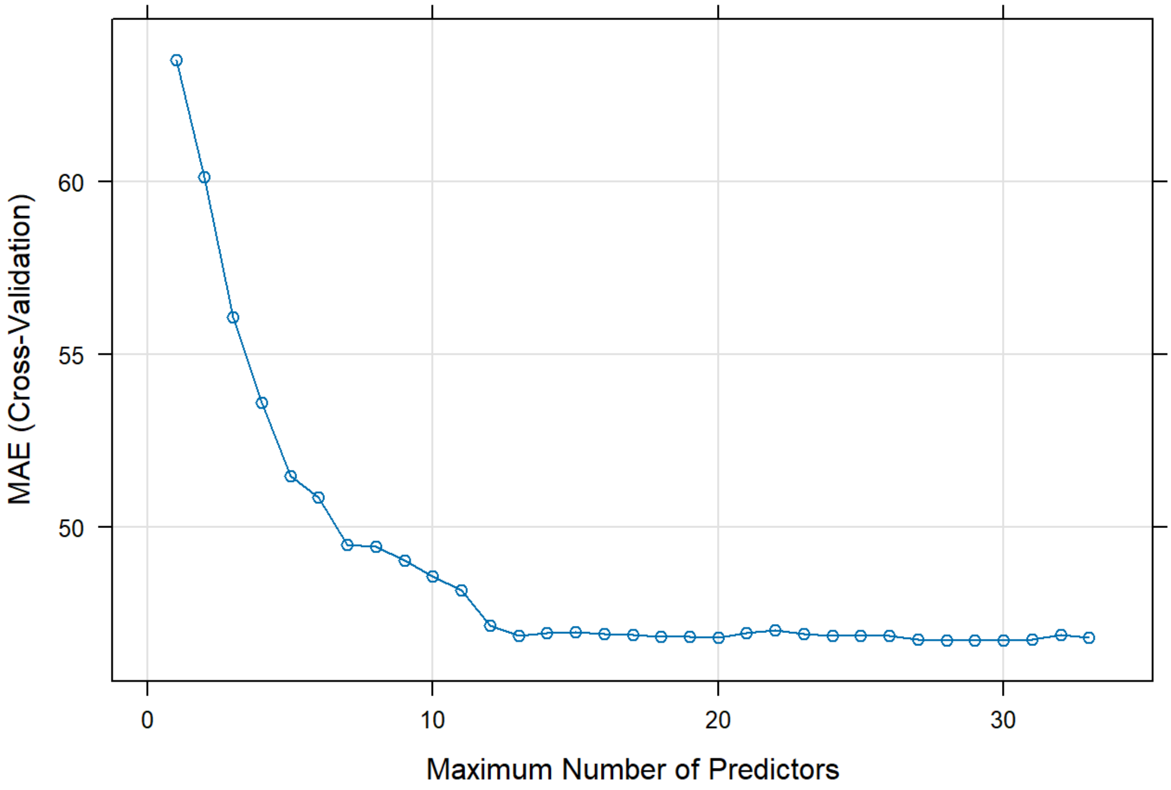
The decision tree diagram displays the hierarchy of features that the CART model used to split the data. It starts with the type of room and then branches out based on other characteristics like the neighborhood group, number of bathrooms, accommodations, and availability over 30 days. The numbers at the top of the nodes indicate the count of listings, and the nodes at the bottom represent the final decision leaves. This tree structure helps us understand the rules the model has learned to predict prices based on listing features.



The scatter plot below the tree diagram illustrates the relationship between the actual prices and those predicted by the CART model. The points are spread around this line, indicating the variance in the model's predictions. The model seems to perform well for lower-priced listings but less so for higher-priced ones, where it tends to underpredict the price, as seen by many points above the line in the higher price range.

RMSE: 74.3971

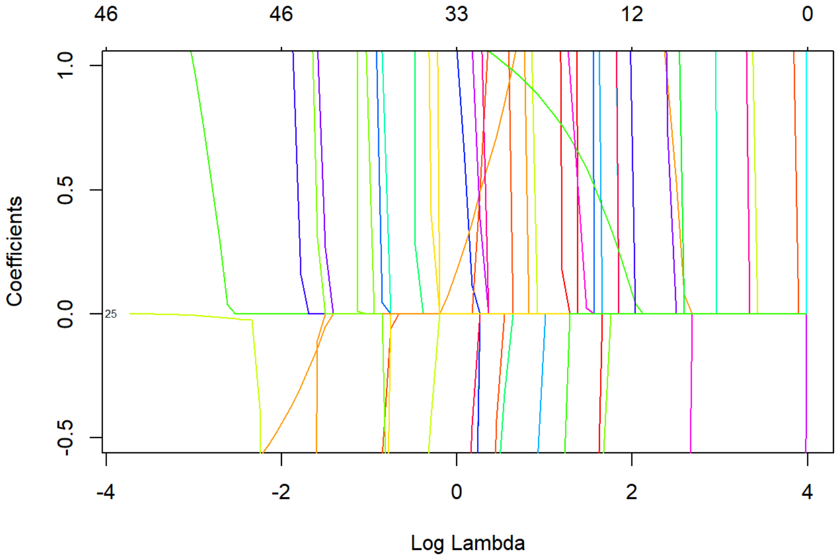
5. Backstep model(non-linearity)



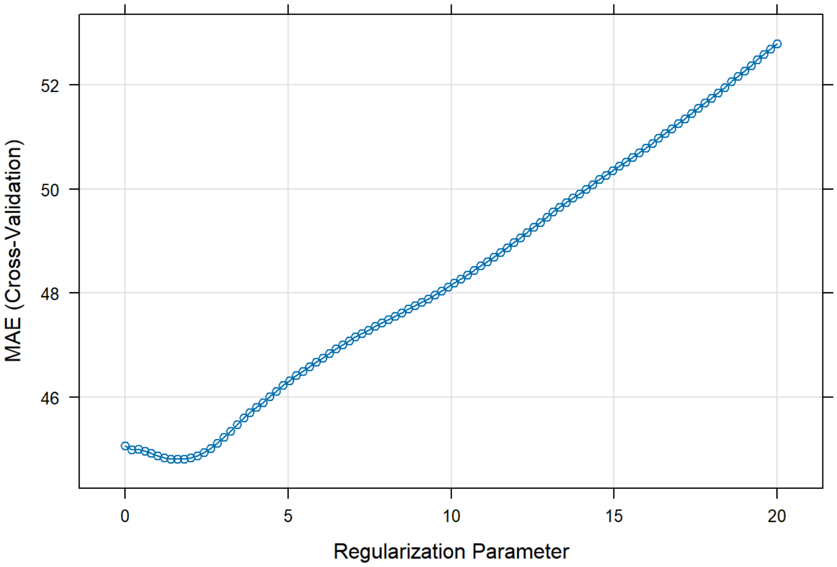
From the curve, we observe a rapid decrease in MAE as the number of predictors increases initially. This suggests that incorporating additional non-linear transformations of predictors into the model significantly improves predictive accuracy. The MAE drops sharply up to around 10 predictors and then levels off, maintaining a relatively constant MAE despite the addition of more predictors.

RMSE: 127.351

6. LASSO model(non-linearity)



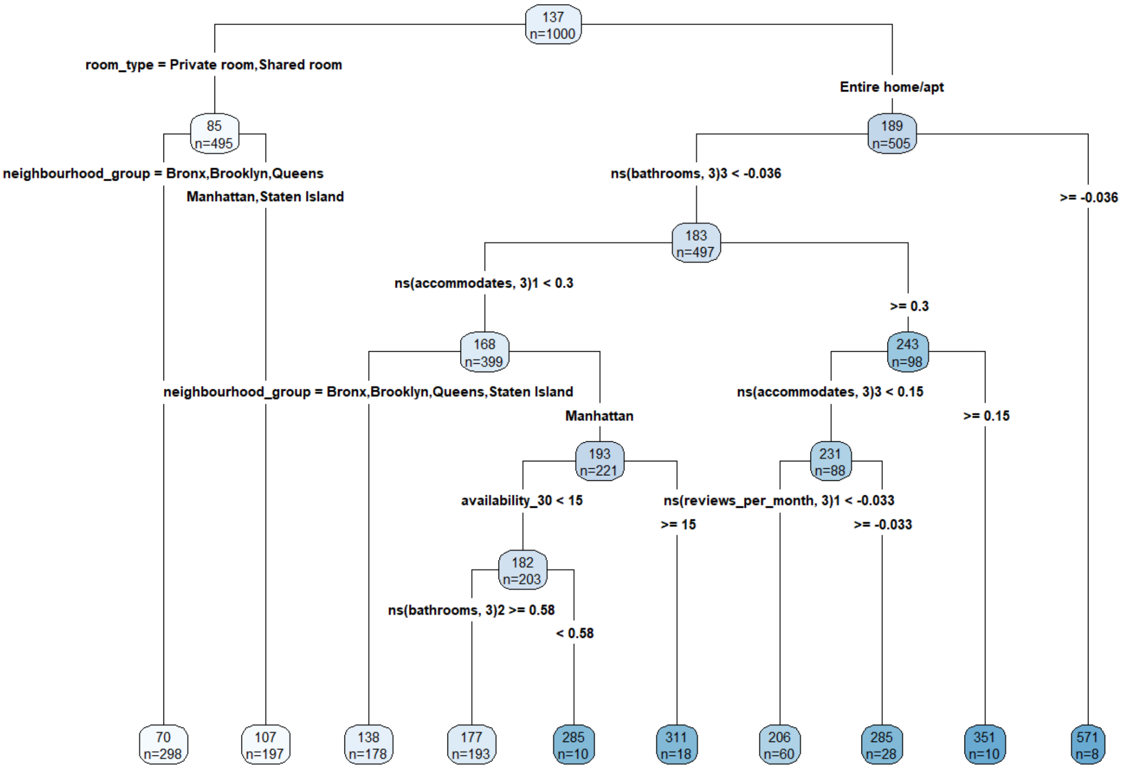
The graph shows us the best amount of 'shrinkage' to apply to our model. If we don't shrink enough, we keep too much unnecessary information. If we shrink too much, we lose important details. There's a sweet spot where the error is the lowest, and that's where our model works best.



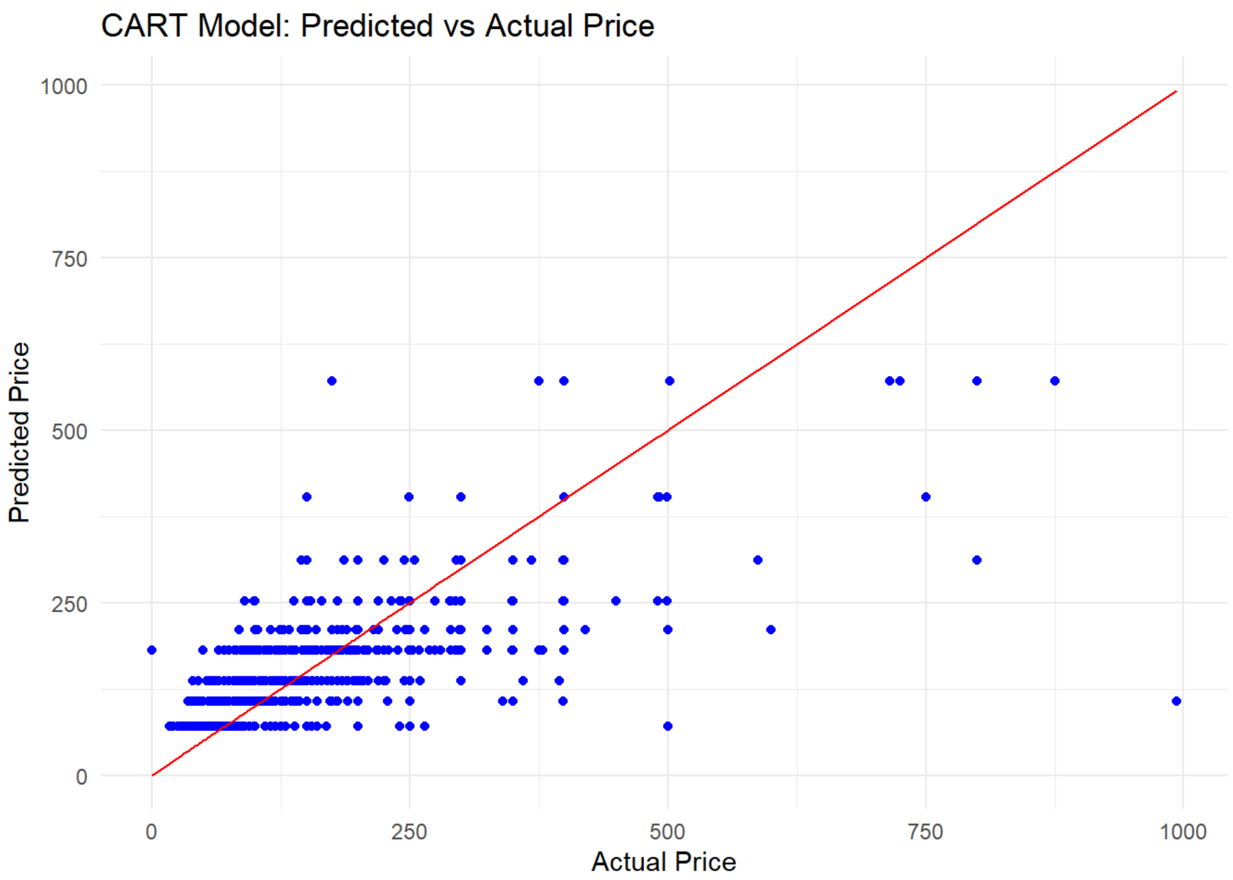
The graph shows how much each detail about the Airbnb (like type of room, location, etc.) influences the price. As we move to the right, we're applying more 'shrinkage' to simplify the model. Some lines drop to the bottom, meaning those details aren't really important for predicting the price. The lines that stay higher up for longer are the details that matter more.

RMSE: 121.873

7. The CART model(non-linearity):


137
n=1000
room_type = Private room,Shared room
Entire home/apt
85
n=495
189
n=505
neighbourhood_group = Bronx,Brooklyn,Queens
Manhattan,Staten Island
ns(bathrooms, 3)3 < -0.036
>= -0.036
183
n=497
ns(accommodates, 3)1 < 0.3
>= 0.3
168
n=399
243
n=98
neighbourhood_group = Bronx,Brooklyn,Queens,Staten Island
Manhattan
ns(accommodates, 3)3 < 0.15
>= 0.15
193
n=221
231
n=88
availability_30 < 15
>= 15
ns(reviews_per_month, 3)1 < -0.033
>= -0.033
182
n=203
ns(bathrooms, 3)2 >= 0.58
< 0.58
70
n=298
107
n=197
138
n=178
177
n=193
285
n=10
311
n=18
206
n=60
285
n=28
351
n=10
571
n=8

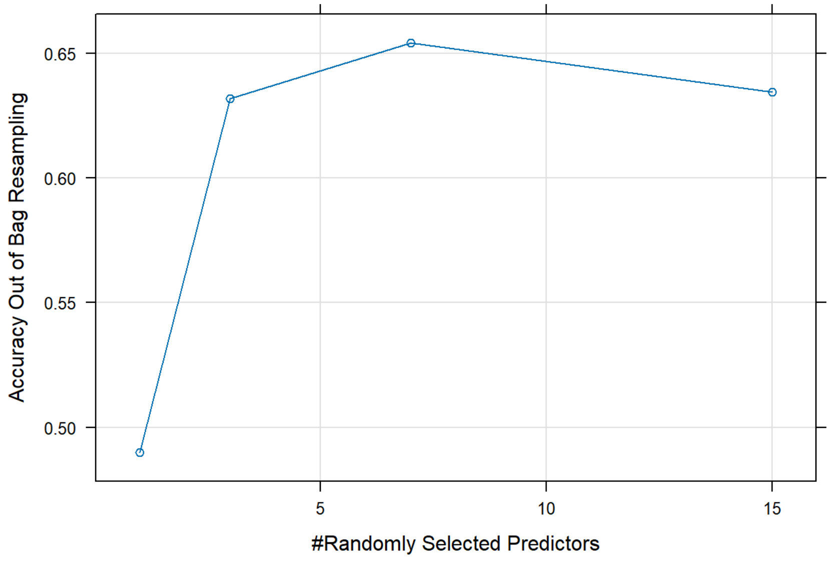
The graph is the map of the model's guess of Airbnb prices. It starts by looking at the type of room and then considers other things like the neighborhood, how many people the place fits, how many bathrooms there are, and how often it's been booked. Each box and branch is like a decision point, leading to what the model thinks the price should be.



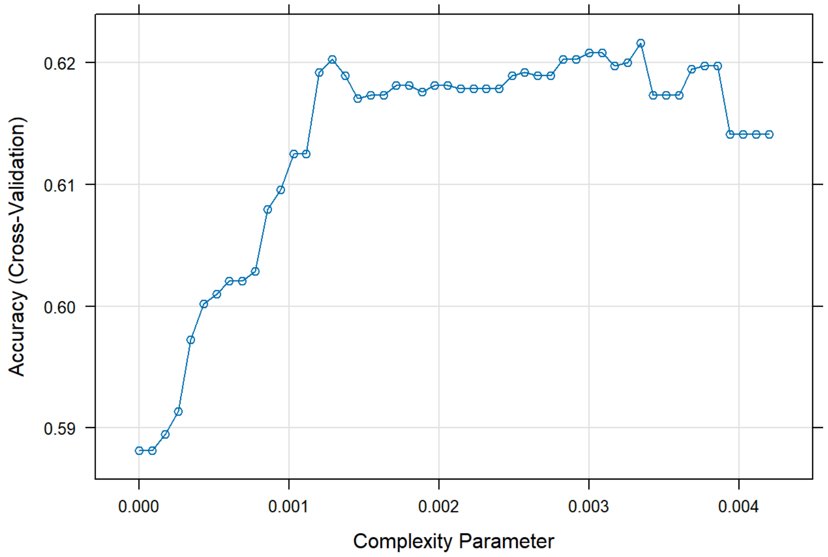
The graph compares what the model thought Airbnb prices would be versus what they actually were. The red line is where the model's guess and the real price are the same. We can see a lot of dots along the line at the lower prices, which means the model did a pretty good job guessing cheaper places. But as the price goes up, the dots spread out, and the model starts to miss the mark, especially for the priciest places.

RMSE: 74.09838

8. Random Forest Model



The graph that tracks how well the model predicts Airbnb neighborhoods based on different numbers of predictors. The accuracy goes up quickly when we first add a few predictors, but after a certain point, adding more doesn't really help; it actually starts to drop off a bit.



After that, we've got a graph that's about finding the right level of complexity for the model. As the complexity increases, so does the accuracy—but only up to a point. After that, even if we keep making the model more complex, the accuracy doesn't really get better.

v. Conclusion:

After comparing all the models, it turned out the random forest model turn out to be the most suitable model. The random forest model looks at the Airbnb data from every angle. It doesn't just stick to one path; it takes a bunch of different routes, checks them out, and then puts all that info together to make a really solid guess about prices.

What really made the random forest stand out was its teamwork approach. It used lots of predictors, but unlike other models that got confused with too much information, the random forest kept its cool. It got smarter as it went, learning which predictors were noisy and which ones actually mattered.

So, when we lined up all the models and checked which one was the sharpest model for predicting Airbnb prices, the random forest model was the standout. It handled the twists and turns better than models that worked alone, and it didn't get tripped up by the tricky parts of the data. That's why it came out on top, making it the MVP of our modeling bunches.