Predicting Airbnb Prices in NYC

Introduction

In recent years Airbnb has become the most popular online marketplace for short-term property rentals in cities across the globe. The company has become an affordable alternative to hotels that provide a more home-like feel for travelers. It is important that customers feel like they are paying a fair price for their Airbnb rental. We aim to create a model that can predict the price of an Airbnb using various classification methods. This model can be useful for customers to make sure they are paying a fair price or look for the best deals on the market. Using data from New York City’s Airbnb market, we will be comparing different classification and prediction models to find the most accurate one at predicting the price of an Airbnb. The models we are using are Naïve Bayes, R-part decision trees, Linear SVM, kNN and random forest models. We will also be running a linear regression model to see if it is more accurate than our classification models.

The Dataset

The dataset used to create these models consists of over 48,000 listed rentals spanning all five boroughs of New York city. The rentals each contain specific information ranging from the listers name to how many days the listing has been available. For the classification models, we decided the following variables would be included along with price:

|  |  |
| --- | --- |
| Neighborhood\_group | The borough the listing is in |
| Neighborhood | The exact neighborhood the listing is in |
| Room\_type | The type of room (Entire home/apt, private room, shared room) |
| Minimum\_nights | The minimum # of nights the listing must be rented. |
| Number\_of\_reviews | The number of reviews for the listing. |
| Calculated\_host\_listings\_count | How total many listings the specific host has. |
| Availability\_365 | The number of days in the year the listing is available. |

We then used these variables to create a summary of the data and discover some initial trends.

We first found the mean price for each borough. The results show that Manhattan has the highest mean rent price of about $197, Brooklyn is next at $124, followed by Staten Island, Queens and The Bronx at $114, $99 and $88 respectively.

Next, we looked at how the mean price changed for each type of apartment. Entire homes/apartments are the highest price, followed by Private Rooms and then Shared Rooms. This makes logical sense as a customer would expect to pay a higher price for more space. Overall the effects that our non-numeric variables had on price makes sense. Manhattan is the most expensive borough while The Bronx is the cheapest while customers pay more for bigger spaces.

Visualizations

The following charts provide a visualized look at how the data is related. This will provide us with information on what variables could be most important in our classification models.

Most importantly, what is the distribution of the prices of the listings. By summarizing the data, we know that there are several outliers that are far more expensive than the mean price. What does a distribution look like?

**Histogram of price**

A screenshot of a cell phone

Description automatically generated

The historgram shows that the data is largely skewed to the right. This is the result of several outliers in price that are greater than $1000 per night. This histogram does not give us a good representation of the bulk of the data. To fix that, we transformed the data using a logarithmic scale.

**Log-Transformed Histogram of Price**

A close up of a map

Description automatically generated

This histogram shows a more normal distribution of the price. It is still slightly skewed, but the data centers more around the mean of $152.70

The following is how the distribution of price looks for each of the five boroughs.

**Histogram of Price by Borough**

**A screenshot of a cell phone

Description automatically generated**

Next where are the listings located. Is there a pattern we can identify about likely listing spots, or are the listings mostly spread out?

A sign lit up at night

Description automatically generated

A screenshot of a cell phone

Description automatically generated

The bar-chart shows that a large majority of the listings in the dataset are located in Manhattan and Brooklyn. The density map on the right confirms this. It also shows that the drive of location and likely price is Manhattan. As you move away from Manhattan, the density of listings decreases, and by looking at the following bar-chart of where higher priced listings are located you begin to see the impact the borough has. This makes sense as most tourist locations and business centers are located in Manhattan with some in parts of Brooklyn. Travelers to the city would want to stay as close to Manhattan as possible and would be willing to pay a higher price to do so. This is important to note for our prediction models, as borough is probably an important factor in determining price.

**Barchart with Above Average Listings by Borough**

**A screenshot of a cell phone

Description automatically generated**

Manhattan and Brooklyn unsurprisingly have the highest number of above-average price listings. Also note that most of the above-average listings are entire homes or apartments. Room type will likely be an important factor in the prediction models.

Next we wanted to look at how price changed relative to the numeric variables.

**Price vs Availability**

**A screenshot of a cell phone

Description automatically generated**

The scatter plot shows that there is no identifiable linear relationship between the price of a listing and how many days during the year it is available. Availability will likely not be important in our models.

**Price vs # of Reviews**

A screenshot of a cell phone

Description automatically generated

There is however a clear negative relationship between price and number of reviews. The higher the number of reviews, the lower the price. This makes sense as most people are trying to save money by using Airbnb so more customers would be staying at lower priced listings. In turn more customers would lead to more reviews.

Classification Models

Because price is a continuous integer variable in the dataset, we needed to transform it into a factor that could be predicted with our classification model. The simplest transformation was to create a two-level factor separated into prices that are above average and prices that are below average. Our models would use this new price variable to predict if a listing’s price was above average or below average. The variables listed above were used to create the different models. Our training data was compiled with 70% of the original dataset, selected at random. The test data was composed of the remaining data. Each model was run with 3-fold cross validation to check for the highest accuracy of the model. After the models were created, each one was run on the test data to create predictions and a confusion matrix. Below are the results for each model:

**R-Part Decision Tree**

The R-Part Decision Tree model predicted 78.57% of the test data correctly. It was more accurate at predicting listings that were below average, having a higher precision and recall for those listings. The final cp for the model was .024

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Overall Accuracy | Precision <Avg | Recall <Avg | Precision >Avg | Recall >Avg |
| R-Part | 78.57% | 86.9% | 83.1% | 59.7% | 66.5% |

**J48 Decision Tree**

The J48 Decision Tree model predicted 81.42% of the test data correctly. It was also more accurate at predicting listings that were below average, having a higher precision and recall for those listings. (seed = 1, numFolds = 5, control(U=FALSE, M=1, C=0.5))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Overall Accuracy | Precision <Avg | Recall <Avg | Precision >Avg | Recall >Avg |
| J48 | 81.42% | 86.7% | 86.3% | 69% | 69.8% |

**Naïve Bayes:**

The Naïve Bayes model predicted 79.53% of the test data correctly. Similar to the previous two models, the NB model was more accurate in predicting listings that were below average. (with discretization, kernel estimation, and numFolds = 4, seed = 3)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Overall Accuracy | Precision <Avg | Recall <Avg | Precision >Avg | Recall >Avg |
| Naïve Bayes | 79.53% | 83.4% | 88% | 68.6% | 59.9% |

**K-Nearest Neighbor**

The KNN correctly predicted 79.48% of the test data. It actually returned the same prediction measures as the Naïve Bayes model. (control(K = 2), and numfolds = 4, seed = 3)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Overall Accuracy | Precision <Avg | Recall <Avg | Precision >Avg | Recall >Avg |
| KNN | 79.48% | 83.4% | 88% | 68.6% | 59.9% |

**Random Forest**

The Random Forest model correctly predicted 80.96% of the test data correctly. Like every other model, the RF model is more accurate in predicting below-average prices. The final model was run with 10 trees. (control(I = 19), numfolds = 4, seed = 3)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Overall Accuracy | Precision <Avg | Recall <Avg | Precision >Avg | Recall >Avg |
| Random Forest | 80.96% | 86.7% | 85.8% | 68.2% | 69.9% |

**Model Summary and Comparisons (Green = Highest, Red = Lowest)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Overall Accuracy | Precision <Avg | Recall <Avg | Precision >Avg | Recall >Avg |
| R-Part | 78.57% | 86.9% | 83.1% | 59.7% | 66.5% |
| J48 | 81.29% | 86.7% | 86.3% | 69% | 69.8% |
| Naïve Bayes | 79.48% | 83.4% | 88% | 68.6% | 59.9% |
| KNN | 79.48% | 83.4% | 88% | 68.6% | 59.9% |
| Random Forest | 80.96% | 86.7% | 85.8% | 68.2% | 69.9% |

The summary of the models shows that they are all very similar. They have similar accuracy, precision and recall. When examining the models further, the tope 20 variables for each model are what we predicted when looking at the data. Each feature is either a room type, a borough or a specific neighborhood. It is clear that location and the type of the room that is being offered are the biggest drivers off price.

Regression Models

Linear regression models were run to find a model that could more accurately predict the actual price of a listing. The regression models used only numeric variables. To substitute for the effect of missing exact neighborhoods and identify location, the latitude and longitude of the listings were used in the model.

The first model that was run used the following formula:

Price ~ latitude + longitude + room\_type + minimum\_nights + availability\_365 + neighbourhood\_group

This model ended up being very weak. The Adjusted R2 of the model was .10, which means it can only explain about 10% of the variation in price.

A second regression model was created using a logarithmic transformation of price as it would help smooth out the distribution. Outliers were also removed for this model. The second model was:

log(price) ~ room\_type + neighbourhood\_group + latitude + longitude + number\_of\_reviews + availability\_36 + reviews\_per\_month + calculated\_host\_listings\_count + minimum\_nights

This model was more accurate with an adjusted R2 of 49.9 but contained two variables that were not statistically significant; number\_of\_reviews and reviews\_per\_month.

A third model was created removing these variables. This model had an adjusted R2 of 49.9 as well, but some of the coefficients did not make much sense given the data. For example, listings in The Bronx were used as the base for the neighborhood\_group variable. We would expect every other borough to have a positive effect on price because off this. However, only Manhattan had a positive effect. Our data showed that there was no relationship between availability and price and a strong negative relationship between number of reviews and price. In our model however, number of reviews was not significant while availability was. Overall, the linear regression model is not very strong, accounting for only about half of the variability in price, and does not make sense logically from what we see in the summary of the data.

Conclusions

The classification models we created using the data are fairly accurate at predicting if the price of an Airbnb listing is going to be above or below the average of $152.70. Each one does it roughly 80% of the time, with the J48 decision tree model being the most accurate and the Rpart decision tree model being the least. This can be very useful for a customer looking to find a great deal on a place or making sure they aren’t getting ripped off. If a price is listed above average, but our model shows that it should be below average, there is a good chance that the listing is a deal. So, which model should customers use? If they are looking for the most accurate prediction, then J48 is the best model. However, if the customer is looking for the best deal then Naïve Bayes or K-Nearest Neighbor are better models. They have the highest recall of below average listings, meaning that it correctly predicts the highest percentage of below average listings. To make sure they are not getting ripped off, a customer should use the Random Forest model as it has the highest recall amongst above average listings. Overall, the models are very similar. They have roughly the same accuracy and are all better at predicting listing with below-average prices. They have minor differences, but choosing the best model depends on what the user is looking for.

Ultimately, we understand that our model has several limitations. First, it only makes a difference when looking at price changes around the mean. As prices move away from the mean, the model becomes very accurate and probably will not be useful to customers. Future models should attempt to break-up the price variable into smaller groups and attempt to classify those. The best models to predict actual price of a listing are most likely regression models. The linear regression models we created however, are not very strong. Getting access to new data will help improve these models and likely our classification models. The exact square footage of the listing would be very useful. Although we can breakdown listings by the type of room, the size of these rooms can vary and likely have an impact on price. A second piece of data that would be helpful in predicting price is when the listing was offered. The time of year and even specific dates are very likely to have an impact on price. Demand for places would likely increase in the summer and around the holiday’s so it would not be surprising to see price rise during these times as well. With this and further additional information, we could work to create more accurate models in the future.