New York City Airbnbs Final Paper

1. Abstract

How can one predict the price of a New York City Airbnb? We are trying to create a machine learning model that can predict the price of a NYC Airbnb given some factors with high accuracy. There are many factors that contribute to pricing an Airbnb. This model is important because it can help people list their Airbnb for a fair price and help a renter determine if a listing price is fair. We first found a dataset containing the price of New York City Airbnbs and many features about these houses. This dataset also contained many factors about the houses that could be important for predicting prices. We then condensed the data into only the data that we thought were important. We compared several regression models to see which would predict the price best. The random forest regressor predicted the price with about 27% away from the price on average. The SGD regression predicted the price with an error of about 42%. Our model wasn't very accurate when predicting the price. To accurately predict the price, more factors about the housing are most likely needed. Some of these could be square footage, bathrooms, and bedrooms. Though, these weren't on the dataset we used.

2. Introduction

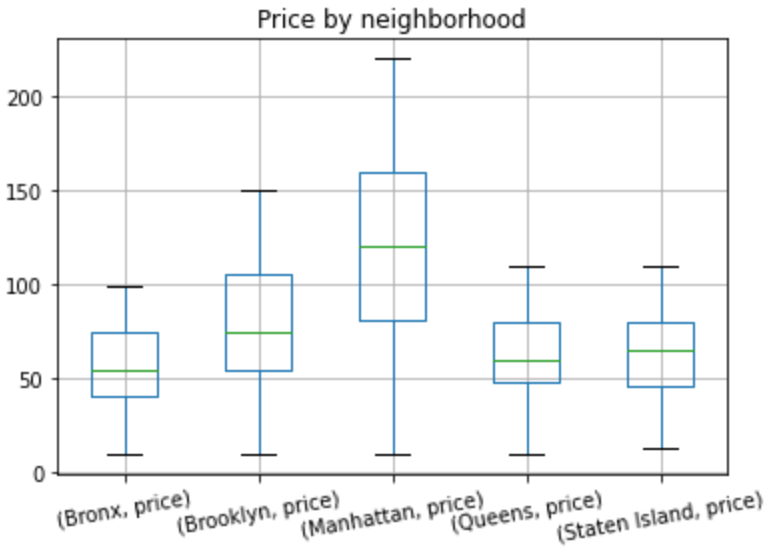
New York City is full of Airbnbs. Each person listing an Airbnb needs to give accurate pricing so that people will believe it is fair. This gives an overall better experience for the renter. Many different aspects of housing can affect prices, such as location, size of the house, and type of house. In New York City, they can be affected by the different boroughs and the type of room. Our goal is to try and make a machine learning model that can predict an accurate price given some of these factors. We used many different models like SGD regression and random forest regressor.Random forest regressor worked the best as it had the smallest mean error percentage. We used supervised learning because the data we used was labeled. Categorical values such as neighborhood and borough were given numerical values for training purposes. We did this by replacing each word with a number. The output of our project is the model we made to predict price. This corresponds to the price. If the estimate is closer to the actual price, the error is smaller, and vice versa.

3. Background

Prior research has used regression models to predict housing prices in places like Mexico City [1], Seattle [4], and Nashville [6]. In the Mexico City paper, the authors attempted to make a model that would predict Airbnb prices in the area. This is nearly identical to our goal aside from the different locations. We used this as a guide to how we should conduct our project. It showed me how we could use visual models to prove our point. For their paper, they showed many different types of models including box and whisker plots and bar graphs, comparing the different attributes of the houses and how they relate to price. Another source was the Seattle paper. Similar to the last paper, these authors attempted to make a predictive model on pricing Airbnbs. They did this using different types of regression models. One of these, we used for mine as well. This is called the Random Forrest Regressor. Moreover, we also used the paper that focused on Nashville. This paper helped me learn the more technical aspects of machine learning. It had many graphs and charts, depicting aspects of the model. Overall, all of these papers helped us immensely in making ours.

4. Dataset

The dataset we chose contained many different factors about Airbnbs in New York City. We accessed it from the website, kaggle[3]. The dataset contained features such as id(the listing identification), name(the name of the listing), host id, host name, neighborhood group, neighborhood, and price. Some of these, like price and id are numerical while others like neighborhood and neighborhood group are categorical. First, we got rid of some of the unneeded columns. We decided not to use these because they weren't factors that normally affect the price of a living area. For the categorical variables, We mapped them to numbers so that they were easier to work with. This made it much easier to use both them and the numerical data when putting them in the model. For the train test split, we used a 80/20 split. This makes it so most of the data is used for training and the remaining data is used for testing. "Neighborhood" described the specific neighborhood in New York City, such as Midtown, Harlem, or queens. This had many values as there are many neighborhoods in New York City. "Neighborhood Group" described the specific borough, like Manhattan, Queens, or Brooklyn. We decided to include neighborhood and neighborhood group because price could potentially be affected by location. There was also room type. This could be private room or entire home or others. This definitely affects price as a larger living area will ultimately drive up costs. We also included minimum nights and availability. For example, id, host id, and host name all don't correlate to anything about the house. This means that it would not be useful to me for predicting price. The processed dataset contained neighborhood, neighborhood group, availability, and room type.



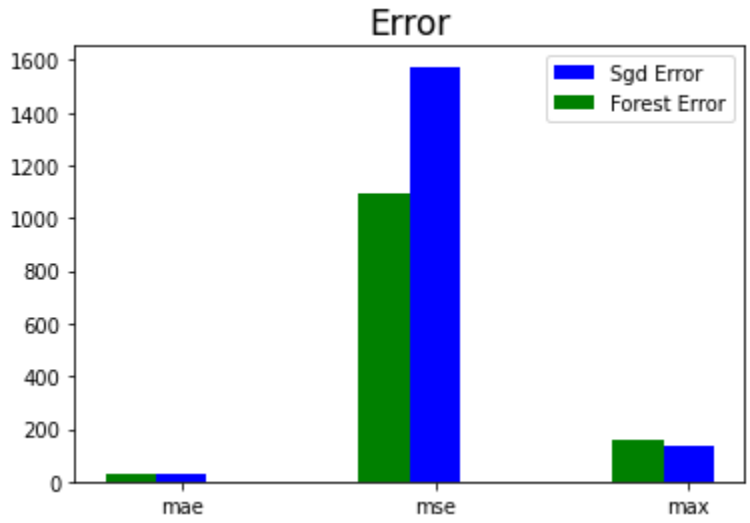
5. Methodology/Models

We chose regression models because regression models require numerical input and all of the data we used was numerical after we converted the categorical variables. One of the main regression models we used is called SGD(scholastic gradient descent) [5] Regressor. Another model we used was a random forest regressor. This model is very different from SGD regressor. It uses multiple different decision trees to try and make a prediction of the price. Each tree is trained on a random subset of the data, hence why random is in the name. The model then evaluates all of the trees and takes the average of the results. We decided on using this as one of our models because it is more complex than the models like linear regression. Linear regression uses the data to make a line of best fit and makes predictions based on that. I chose not to use it because of the simplicity of the model and the fact that my data didn't seem to have a linear relationship. SGD regression works by estimating the difference of loss across samples and is updated across the sample. This model didn't work as well as the random forest regressor.

6. Results and Discussion

The more accurate model was the random forest regressor. The mean absolute error for this model was 22.80. This number is the absolute error over the course of the test and train sets. The mean absolute percentage error was 0.27. This means that the model was about 27% off. The r squared score was 0.58. This value explains how the dependent variables vary based on the

independent variables in a regression model. The mean squared error was 917.85. This number outlines the fit on the data with the regression line. This is not good for the model because the number is better when it is closer to 0. This number is a large distance from 0, meaning that there is a lot of error/. Lastly, the max error was 160.7. This is also not a good score because it means that one of the points was 160.7 off of where it should be. The other model we used is called the SGD regressor. The mean absolute error was 30.74. This is greater than the random forest regressor which means it is worse off. The mean absolute percentage error was 0.42. This means that the model was about 42% off which is worse than the random forest regressor which was 27% off. The r squared score was 0.25. This means that there was less variation than the other model. The mean squared error was 1575.05. This number is much further from 0 indicating that is didn't fit as well. The max error was 138.23. This is less than the random forest regressor although this one doesn't matter as much as the others because it only describes a certain point.



Model	Mean absolute Error	Mean Absolute Percentage Error	R squared score	Mean Squared Error	Maxmimum Error
Random Forest Regressor	22.80	0.27	0.58	917.85	160.70
SGD Regressor	30.74	0.42	0.25	1575.05	138.23

7. Conclusion

The purpose of this project was for predicting the price of a New York City Airbnb. My goal at the start of the project was to get the model to accurately predict the prices of the Airbnbs. We did this using SGD regression and random forest regressor. Overall neither were very successful though we believe the project has more potential in the future. First off, one could try more and different models. It may be that the models we chose weren't a good fit for the data we chose. There are many different models online and there is certainly one that could improve our work on the project. Furthermore, there may be other datasets with better factors to fit the code. For example, if a dataset had a square footage column, the model may be able to fit the data much better if there is a better correlation. Overall if these are taken into account, the error can become even smaller making the model much more accurate.

8. Acknowledgements

[2-3 sentences max] In this section, you should briefly thank or acknowledge any individuals or institutions that provided significant help and/or advice on your research, analysis, and report writing

[Your Acknowledgements Section Here]

I would like to thank Tomer Arnon for assisting me on writing the paper.

9. References

Works Cited

[1] Alejandra Gomez Cravioto, Daniela, et al. "Mexico City's Airbnb Listing Price Analysis Using Regression." *Reasearchgate*, https://www.researchgate.net/publication/348975238_MEXICO_CITY'S_AIRBNB_LISTING_PRICE_ANALYSIS_USING_REGRESSION.

[2] Choudhary, Paridhi, et al. "Unravelling Airbnb Predicting Price for New Listing." arXiv, 2018, https://doi.org/10.48550/arXiv.1805.12101.

[3] Dgomonov. “New York City Airbnb Open Data.” Kaggle, 12 Aug. 2019, https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data.

[4] Keating, Joshua. “Predicting Airbnb Prices in Seattle.” Josh Keating, 29 Aug. 2020, https://joshuakeating.com/.

[5] “Sklearn.linear_model.Sgdregressor.” *Scikit*, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html.

[6] Zhang, Zhihua, et al. “Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach.” MDPI, Multidisciplinary Digital Publishing Institute, 14 Sept. 2017, https://www.mdpi.com/2071-1050/9/9/1635.