GEORGETOWN MSBA 3

Airbnb Neural Network Predictive Model Analysis

Machine Learning II- Assignment 2

By: Najnin Sneha and Taylor Whitelow

Executive Summary/Problem Statement

Our team has been tasked with analyzing a sample of Airbnb's data that covers a few rental locations within the District of Columbia. The data was collected in September of 2022 and includes data from neighborhoods within DC, such as Capitol Hill, Takoma, Fort McNair, as well as several variables that may influence demand, price, and rating of each rental.

Our goal was to analyze the data provided and build a predictive model for price of the listing. Upon creating multiple models, we can conclude that model 2, using the Caret nnet package gave us a better output of the predictive price.

Processing and Engineering the Data

The Airbnb data was first manipulated to provide optimum modeling of our neural networks. To do that, our team removed several nominal categorical variables that would not contribute to our models, such as listing id and host since. As such, the super host feature (which is an ordinal categorical variable) was also removed from the dataset used for our models. Due to most of the super host feature observations needing to be corrected (i.e., most hosts were not super hosts), our models would have put too much importance on this feature.

After removing certain features, our team wanted to ensure that our dataset did not need up-sampling or down-sampling. Using the summary function and plotting a scatterplot of our target variable (price), we determined that no up or down sampling was needed. Next, to provide unbiased predictive logistical results, the 1,711 listings were split, with 70 percent in the training dataset and 30 percent in testing. Once our team split the data, we noticed that the features columns had spaces. We had to eliminate those spaces for our neural network models to run. So, a good portion of our data processing code is dedicated to removing those spaces and replacing them with underscores.

Next, our team wanted to see if any data needed to be included in our observations. We found 201 of our Airbnb observations in our training data and 80 missing portions in our testing data. Because of this, both datasets were inputted with the average answers from the training dataset. Features that need an average inputted were host acceptance rate, total reviews, average

rating, room type- entire home/apt, room type- private room, and room type- shared room.

Lastly, all the data was scaled with min-max standardization (including the target price feature), transforming the features to all have the same scale.

Neutral Network Models and Insights

Using three artificial neural networks (ANN) models to predict listing pricing from our Airbnb listing dataset, we used regression-supervised learning models by extracting functional patterns from the Airbnb dataset.

Model 1

Our first ANN model used the reLU activation function with two hidden layers with linear output set to true. When plotting our listed prices from our training data set with predicted price, as predicted by our first NN model, the plot appears to show a somewhat weak correlation between the first NN model's ability to predict price against the actual price, and it is underpredicting price (*Exhibit 1*). Similarly, when examining the R-squared for our first NN model, we only got an r-squared of 0.4082, which is relatively low.

Model 2

For our second NN model, we wanted to use Caret's nnet package. In this model, we wanted the nnet package to cycle through 10 different sizes (i.e., neurons) on a single layer with three decay rates (i.e.-learning rate). The model cycles through these different sizes and decay rates to see which combination of neurons and learning rate provides the model with the lowest RMSE. The model found that a neural network with a single layer, eight neuron network with a decay (i.e., learning rate) of 0.01 is the model that provides the lowest RMSE (*Exhibit 2*). Even with a cross-validation model, when plotting the predicted price against the actual price, the model is again under the predicted price with an r-squared of roughly .41.

Model 3

Lastly, for our third model, we wanted to go outside the box and use the tangent hyperbolic activation function with the modified globally convergent algorithm and a step max of 200,000.

Changing parameters such as step max and the activation function would significantly affect the better listing price prediction.

When modeling the predicted price with the actual price, the model could have done a better job than our other two models and under-predicted price once again (*Exhibit 3*).

Model Comparison

When reviewing the results of each of our models from its measures of fit (via the postResample package), our second model performed better than the other two models (*Exhibit 4),* considering that we asked the caret package to find the optimum model with the lowest RMSE given a particular set of criteria within the size and decay parameters. However, while analyzing the three models, our team noticed that the r-squared numbers were weaker than expected, considering that all three models had r-squared amounts in the low to mid .40 range. Consequently, when plotting each model's predicated price with the actual price, every model is under the predicted actual price. Our team then went back and more closely examined the Airbnb dataset. Our team believes that the limited amount of data within the dataset may be causing the overall underperformance of all three models. Our neural network models would have significantly benefited from a much larger dataset that looks at Washington DC listings and Northern Virginia and Southern Maryland to capture the entire metropolitan area.

Conclusion

Our team wanted to look back on our previous analysis using univariant and multivariant regression, which analyzed the same data sent as our neural network models. In the end, those regression models had more predictive power than our neural network models. However, our team is cautious in comparing our regression and neural network models due to scaling and removing several variables within our neutral network models. Our team ultimately found that, with the feature we removed from the dataset and the three neural network models we created, we could only predict about 50% of Airbnb prices for the DC area. This could be due to the smaller size of the dataset provided by Airbnb for the public. Neural networks are known to be to need more data than the standard machine learning algorithms. As such, Airbnb should provide a dataset that includes not only listings in DC but also Northern Virginia and Southern Maryland, as that may provide higher predictive power for our neural network models.

Appendix

Exhibit 1- Plotting predicted price with actual price from Model 1

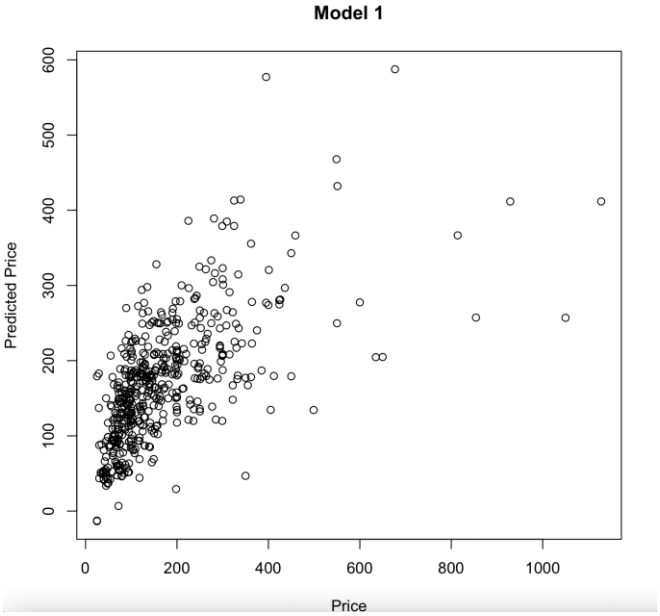

Exhibit 2- RMSE with different decay rates and hidden layers for Model 2

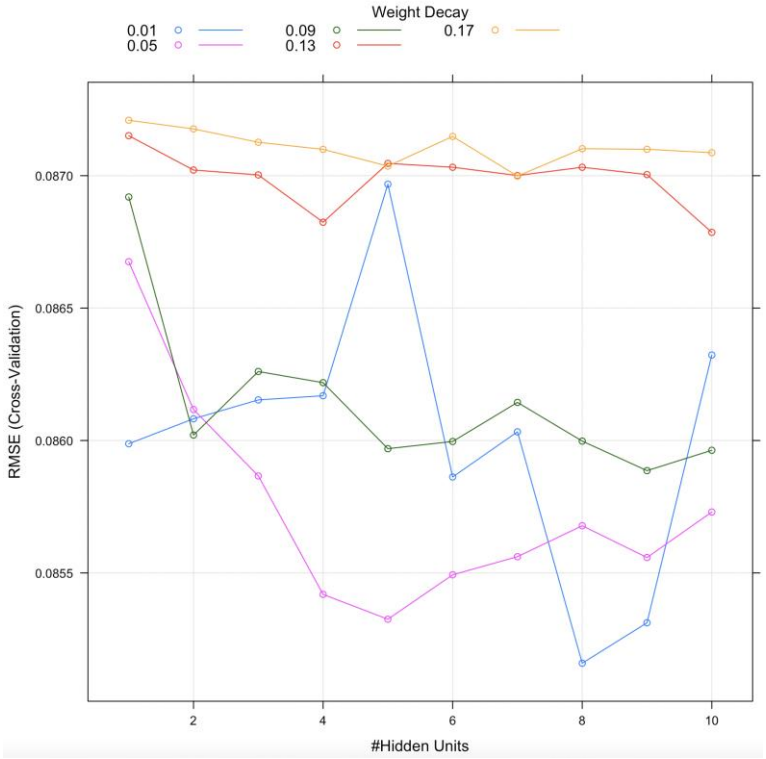

Exhibit 3- Plotting predicted price with actual price from Model 2

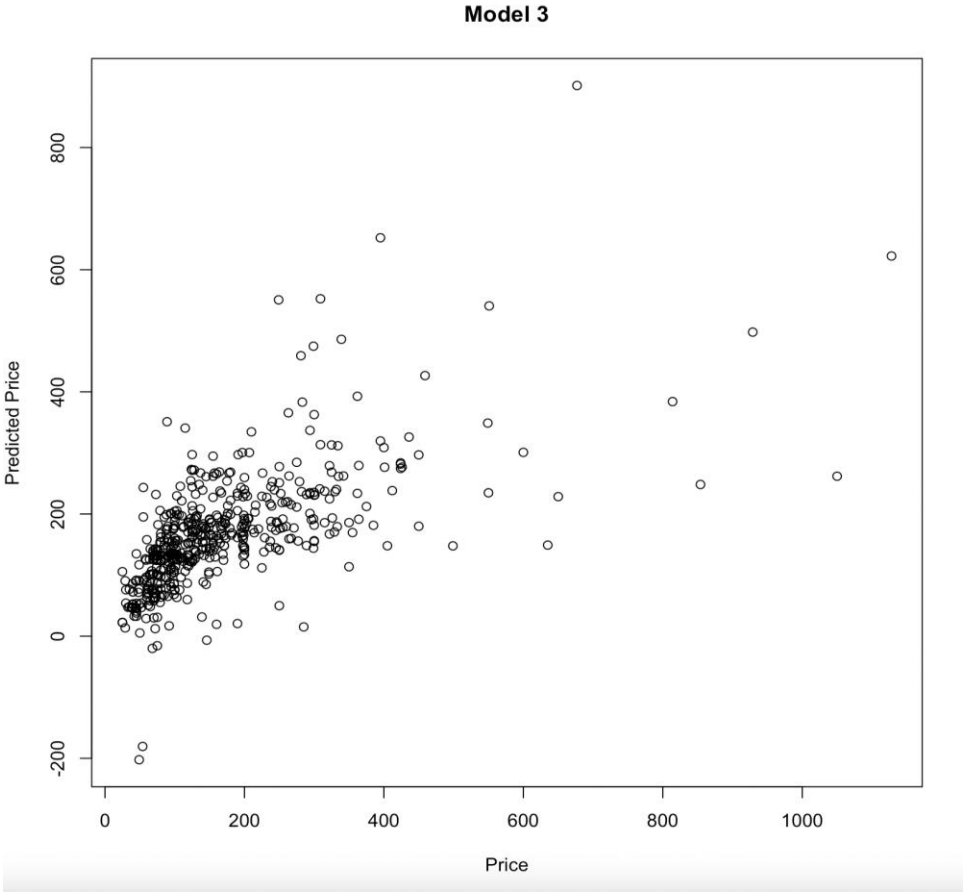

Exhibit 4- Comparing our model's measures of fit

Model Name	RMSE	Rsquared	MAE
Model 1	100.6971131	0.4237064	64.0299745
Model 2	98.1899378	0.4533296	61.0931079
Model 3	105.2712903	0.3907379	65.0505948