Andrew Singh and Mike Johnson

OPAN 6603: Machine Learning II - Project 2

Executive Summary

Our team is providing an analytical summary to investors demonstrating the use of artificial neural networks (ANN's), to further understand the influential factors of price on Airbnb's short-term rentals and homes in Washington D.C. in July 2023. Our proposal demonstrates our main takeaways, comprehensive data analysis, and comparative analysis of our prior findings with our current results.

The key findings from our analysis include:

1) The **caret** model achieved a moderate R-squared value of 0.54, indicating that the machine learning model captures some of the variability in the data, but there is still a significant portion unexplained.

2) The **neuralnet** model achieved similar results with an R-squared value of 0.53.

Based on the two models developed, either could be used going forward. However, we recommend the **caret** model due to its slightly better predictive performance.

Introduction

Our data analytics team is performing a new data analytics project for prospective investors in the short-term rental market of Washington D.C. using ANN models to build a predictive model of the price of listings. Additionally, we are interested in understanding the comparative performance of ANN models to multivariate regression models introduced in a prior report. The dataset used for this report was scraped, compiled, and made publicly available for consumption on the website Inside Airbnb.

Methodology and Data

We are using a publicly available dataset provided by Inside Airbnb that contains 1,719 Airbnb listings with 15 individual characteristics about each listing. The key variables of interest are host statistics (acceptance rate, total listings, start date of offering Airbnb listings, total reviews, and average ratings) and the individual characteristics of each Airbnb location (bathrooms, bedrooms, beds, minimum night stays, and price).

As previously mentioned, for our methodology, two ANN models were developed by using the 'neuralnet' and 'caret' package in R.

From our initial data-processing steps, we took the following actions: 1) performed data visualization of relatable factors to find underlying insights and trends, 2) removed unnecessary variables from the train- and test- datasets, 3) handled missing values by imputation, 4) updated data types to reflect the underlying data for factor, numeric, and date records, and 5) modified the 'host_since' field to become 'host_days' referencing the number of days a host joined

Airbnb.

Artificial Neural Networks (ANNs) are a supervised machine learning model that could be adapted to unsupervised learning challenges. The model acts similarly to a generalized linear model, such as linear and logistic regression, fitting a curve through a series of points. ANN models are thought about as function approximators, mapping inputs to outputs, and are composed of interconnected computational units called neurons. The model is trained by using backpropagation (an algorithm to train ANN models by adjusting weights to minimize error) and gradient descent (an optimization function to minimize the error in a neural network), with the goal to minimize an error function (also called the loss function). The goal of ANN models are to detect hidden patterns, generalize insights from data for insights, and are mostly used for business challenges in forecasting, classification, clustering, and anomaly detection.

Model Selection

The training of the carer model included several hyperparameter tunings to find the best combination for the model. The best combination was selected through cross validation (Figure 1). The two hyperparameters included:

- **Size:** This represents the number of hidden layers in the model. Increasing the size can allow for the model to learn more complex patterns but runs the risk of overfitting. On the other hand, smaller sizes can lead to underfitting. The model selected had a size of **2**.
- **Decay:** This is a regularization technique that helps prevent overfitting. Higher decay values can improve generalization but might underfit if the value is too high. Lower decay values mean less generalization, but risk overfitting. The model selected had a decay value of **0.01**.

The neuralnet model does not support similar parameter tuning or setting a decay value. The model selected for neuralnet had a size of 2.

Key Findings

The models developed for this engagement demonstrated varying performance with the caret model performing slightly better than the alternative model, neuralnet.

The table below summarizes the key metrics for both models. The two models had similar performance, with the caret model having a slight edge in R-squared.

Artificial Neural Network (ANN) Metrics			
Model	Root Mean Squared Error (RMSE)	R-Squared	Mean Absolute Error (MAE)
caret	0.8	0.54	0.5
neuralnet	0.8	0.53	0.5

In a previous engagement, linear regression models were developed that attempted to predict price using the same data. In that engagement, simple linear regression models were built using each independent variable and a multiple regression model, using a select set of independent variables.

The best performing simple linear regression model, using bathrooms achieved an R-squared of 0.38. The multiple regression model using bathrooms, accommodates, bedrooms achieved an R-squared of 0.43 (Figure 3).

While the regression models performed worse, these models are interpretable with marginal effects on price that can be measured for each independent variable. This is not possible with the neural network models. The advantage in the neural network models comes from their predictive performance, and that is reflected in the model metrics when compared to the regression models.

Recommendations

The two neural network models developed had similar performance, but the caret model has a slight edge with a higher R-squared. Based on this model's performance compared to the regression models developed previously, we recommend using the caret model based on the models metrics that suggest more reliable price predictions.

Conclusion

Our analysis highlights the potential of artificial neural networks in predicting Airbnb listing prices in Washington, D.C. The caret model demonstrated slightly better performance, achieving an R-squared of 0.54, making it our recommended model for future use. Compared to the previously developed regression models, both ANN models showed stronger predictive capabilities, though at the cost of interpretability. Given these findings, we recommend that Airbnb's leadership consider implementing the caret model to enhance price prediction accuracy and support data-driven decision-making.

Appendix

Figure 1: Caret Model Cross Validation

Training of the Caret model included several hyperparameters to find the best combination for the model. The visual below summarizes the cross validation RMSE results for each. For RMSE, lower is better. In this case, the best tuned model has **2 neurons** in the hidden layer and a **weight decay parameter of 0.01**.

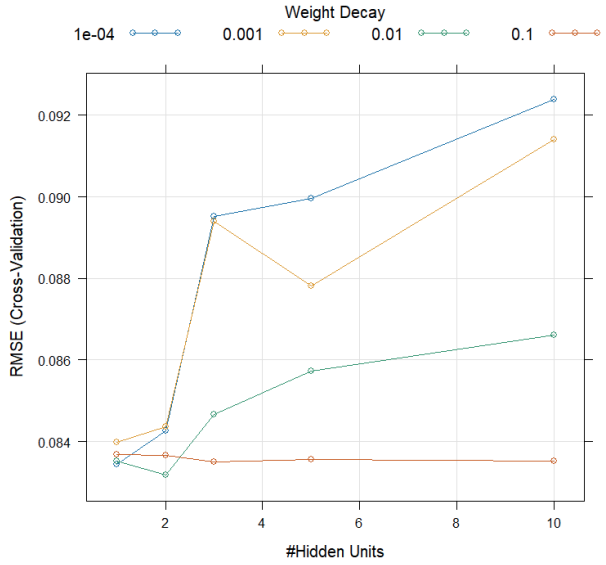

Figure 2: Neuralnet Model

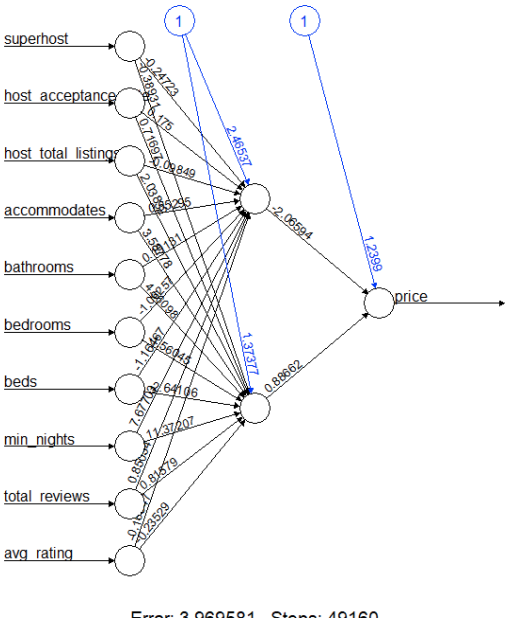

Figure 3: Linear Regression Models: R-Squared

The table below summarizes the R-Squared values for the linear regression models developed in the previous engagement.

Linear Regression Models: R-Squared	
Model	**R-Squared**
Multiple Linear Regression	0.43
Bathrooms	0.38
Accommodates	0.34
Bedrooms	0.34
Beds	0.27
Neighborhood	0.01
Superhost	0.01
Host Acceptance Rate	0.01
Room Type	0.01
Min Nights	0.01
Total Reviews	0.01
Host Since	0.00
Host Total Listings	0.00
Avg Rating	0.00