Caroline Tomasik

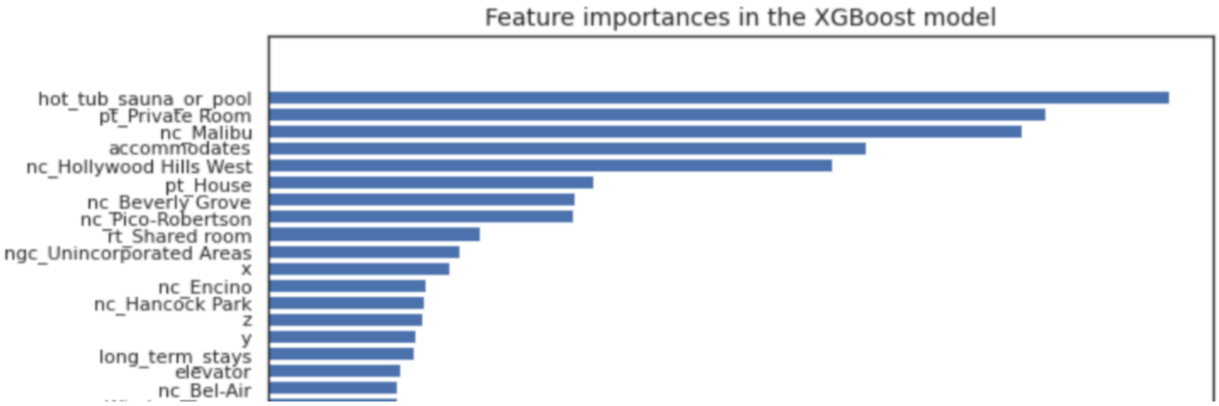
AANN Final Report: Examining Airbnb

Founded in 2008, Airbnb has quickly become the world's leading short term rental business, with over 150 million users across 191 countries. Airbnb is an online platform that allows users to find and book properties. These properties vary widely and can range from a shared room in a house, to a private apartment, to exotic stays, such as a castle or tree house. The purpose of this project is to predict the price of Airbnb listings using machine learning models and artificial neural networks. The data was gathered from insideairbnb.com, which contains open source data web-scraped from Airbnb.com. Two recent datasets (web-scraped April 9, 2021) containing listings from Los Angeles County, California were examined, the first dataset was extensive and contained 74 columns, and the other was simplified and contained only 16 columns.

Extensive Dataset:

The extensive dataset contained data for 35,757 listings in Los Angeles County, California. The first step was data cleaning, which was quite extensive due to the messiness of the data. Data cleaning included: removing uninformative columns, removing price outliers, changing boolean features to binaries, converting datetimes to integers, data imputation, removing collinear columns, and creating one-hot encodings for categorical features.

After data cleaning, machine learning models for regression (linear regression, XGBoost, and random forest) were used as baselines to compare to the artificial neural networks. The simple linear regression model performed very poorly, but the random forest model and XGBoost model performed better (0.58 and 0.61 explained variance, respectively). According to the feature importance scores from the XGBoost model (see below), some of the top price predictors regarding the listing were: having a hot tub/sauna/pool, if the listing was a private room, and the number of people the listing accommodates for. Among the top predictors also included the listing being neighborhoods such as: Malibu, Beverly Hills, Bel-Air, and Hollywood Hills. This comports with reason, as these are very expensive neighborhoods.



Various neural network models were also run: Model 1 and Model 2 used the same architecture, but Model 1 used the SGD optimizer, and a learning rate of 0.05, and had a train loss of 0.004 and a test loss of 0.005. Model 2 utilized the Adagrad optimizer, and a 0.1 learning rate, and also resulted in a train loss of 0.004 and test loss of 0.005. Model 3 implemented a three layer architecture with Adagrad and a 0.05 learning rate, and resulted in a training and test loss of 0.004. Model 4 utilized the sigmoid activation function (as opposed to ReLU like the previous architectures), Adagrad, and a 0.1 learning rate, and resulted in the best performance: training loss 0.002, test loss 0.004.

Scaling: When constructing the artificial neural networks, the type of scaler used was very influential: using StandardScaler produced far worse results (train loss: 0.1, validation loss: 0.47) compared to using MinMaxScaler (train loss: 0.004, validation loss: 0.005) which produced far better results.

Optimizer: Adagrad generally performed well (train loss of 0.003; test loss of 0.005). SGD performed similarly, with train and test loss at 0.004. RMSprop had the worst performance, as the model got stuck at a train and test loss of 0.014. Adam performed slightly better, but similarly plateaued at a train and test loss of 0.007.

Dropout: Including or not including dropout, as well as the dropout rate had very little influence on the performance of the model.

Simple Dataset:

A simple dataset of far fewer features was also examined for comparison. This dataset required very minimal cleaning, but had worse results, presumably due to the fewer number of features. XGBoost: training loss of 0.008, test loss of 0.007. Random forest regressor: training loss of 0.001, test loss of 0.007. Various ANN models were conducted on the simple dataset, but the training loss remained close to 0.009, test loss at 0.008.

Conclusion:

In conclusion, the extensive dataset provided better results than using the simple dataset. With the extensive dataset, overall, the machine learning models were very comparable to the artificial neural networks. The XGBoost model had a training loss of 0.001, test loss of 0.005; and the random forest model had a training loss of 0.0007, test loss of 0.005. Although the test loss was slightly lower with the best ANN models, the results are very similar.

A future model would likely benefit from including other features regarding an Airbnb listing, such as photos of the listing and the sentiment analysis of the individual reviews. For the purposes of this project, only the review scores were retained, however investigating the impact of written reviews may be worthwhile.