

*Machine Learning Bootcamp - 2020*

**Team No. 4**

**Project 3 Final Report**

**Submitted By:**

**Celine Cherian - Business Analyst**

**Miguel Mendoza - Developer**

**Fatmir Likrama – Data Engineer**

**Pratik Ghatake – Data Scientist**

1. **Problem Introduction**

The gig economy is a type of economy with a workforce based on "single projects or tasks for which a worker is hired, often through a digital marketplace, to work on demand."

The New Yorker calls the gig economy the "on-demand, peer, or platform economy." Embodied by companies like Airbnb, Uber, TaskRabbit, Handy, Thumbtack, and Fiverr, the gig economy operates by offering marketplaces based on ratings and payment systems routed through apps. Many indicators show that customers benefit from the gig economy not just by the enhanced and quality of services but also by paying less.

Airbnb, Inc. is an American vacation rental online marketplace company based in San Francisco, It offers arrangements for lodging, primarily homestays, or tourism experiences. Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. The dataset we used describes the listing activity and metrics in NYC, NY for 2019. We will use this dataset to predict how much the cost would be per night in NY based on several variables in the data provided.

We believe that such a model will help travelers and hosts alike in estimating what the supported price for lodging in NY would be based on location and other factors.

1. **Dataset**

We used an Airbnb dataset from New York residences provided by the company itself and hosted on Kaggle. The raw data consists of more than 48k rows split into columns of 'name', 'host\_id', 'host\_name', 'neighbourhood\_group', 'neighbourhood', 'latitude', 'longitude', 'room\_type', 'price', 'minimum\_nights', ‘number\_of\_reviews', 'last\_review', 'reviews\_per\_month', 'calculated\_host\_listings\_count', and 'availability\_365' variables. We considered the ‘price’ as the dependent variable to be estimated using the rest of the columns as the independent variables.

1. **Features and Processing**

Dataset is complete with the exception of very few missing name and host\_id data. About 15% of ‘last review’ and ‘reviews\_per\_month’ data is missing. Following our exploratory data analysis we removed this missing data from the analysis. We also removed outliers (distanced more than three standard deviations from the mean) on the ‘price’ column. We calculated the number of days from the date of last review and used that instead as a predictor in the data. To check whether it would improve prediction, we also categorized the price column into five categories as follows: 'very cheap', 'cheap', 'fair', 'expensive', 'very expensive', based on the quintiles of the values of the price column. Categorical data were one hot typed using dummy variables.

1. **Models and Techniques**

Neural networks using the Keras library were used to predict the price per night based on data such as neighborhood, number of rooms, days since last review, minimum nights, reviews per month availability, and host listings count. Half of the data was used to train the model and the other half was used to test it.

We tested networks with different numbers of hidden layers, different activation functions (rectified linear and sigmoid), with and without regularization, and tested prediction using a categorical value instead (with a categorical cross entropy loss function)

1. **Results**

Results obtained were not as satisfactory as initially expected. Predictions on the continuous price variable had a mean average error of about 33 USD on the test data. No significant improvement was observed using different network architecture, different activation functions or regularization.

Results on the categorical price were not satisfactory either. We got an accuracy of slightly higher than 47% which is not much better than the random 20%.

The poor results indicate that there may be other factors determining price besides location, reviews and other variables present in our dataset. One possible one is quality and attractiveness of the housing which often can be inferred from pictures--such data was not available.

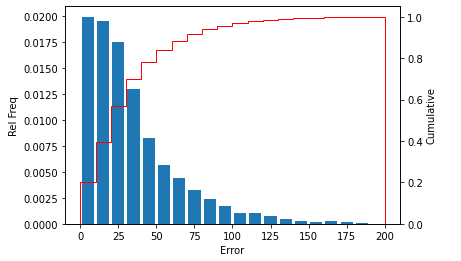


Fig 1. Graph displays that most errors are less than 15 USD

**Summary and Conclusions**

We used data analytics, statistics and machine learning methods to predict the price of housing per night on Airbnb data in New York City. Use of neural network modeling with Keras library, hidden layers, sigmoid activation function, and binning had little to no improvement on increasing the accuracy of predicting Airbnb prices based on the independent variables - with only 47% accuracy.

Results we obtained were not encouraging and we suspect that other factors besides the ones present in the data are responsible for the price.

**References**

https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data