

*Machine Learning Bootcamp - 2020*

**Team No. 4**

**Project 3 Abstract**

**Submitted By-**

**Celine Cherian - Business Analyst**

**Miguel Mendoza - Developer**

**Fatmir Likrama – Data Engineer**

**Pratik Ghatake – Data Scientist**

**Abstract**

Airbnb is an online marketplace that allows hosts to accommodate guests all over the world, but with risks. Especially during the pandemic where Airbnb had to cancel accommodations to reduce the virus from spreading. Airbnb was created in 2008, where guests and hosts have used Airbnb as a way to help travelers have a place to stay in a local area that is cheaper than hotels. However the risks that both hosts and guests face when using Airbnb have impacted how Airbnb users travel and provide accomodations. For instance, hosts provide their home, or any available space, that provides guests a place to rest at a price the host sets. Hosts risk having their property damaged by these guests and reduce the value of their available space. Guests, on the other hand, risk choosing a host accommodation that may not be the same as what the guest saw on Airbnb. Hosts and guests need a way to reduce the risks and provide a trustworthy business.

This motivates us to tackle this business problem and find possible solutions for it. We can do this by learning from 2019 New York City Airbnb data. For this project, our focus would be predicting why guests choose a certain host and their accommodation through Airbnb with the use of different data attributes. In order to achieve the predictive analysis, a model will encompass various statistical techniques from Supervised Machine learning specifically, deep learning algorithms. These techniques will study the 2019 data and will try to make future predictions. This detection will help both guests and hosts in making well assessed decisions for proper accommodation setup in the online marketplace. The developed model can be used in real world scenarios with the purpose of reducing the risks for both hosts and guests and allowing more hosts to get paid and guests to travel more economically and safely.

**Data Preparation and Analysis**

We will be using an Airbnb dataset from New York residences. 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', 'availability\_365' variables. We will consider the price as the dependent variable to be estimated using the rest of the columns as the independent variables. We suspect price would likely depend strongly on location/neighborhood and room type but it will be interesting to see other patterns and dependencies in the data.

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 will decide on how to deal with the missing data.

**Hypothesis**

Once the data wrangling is completed, Exploratory data analysis will take place upon the dataset in order to find patterns and develop data understanding before choosing the best method to adjust the dataset to any algorithms. This may involve answering common questions such as which hosts are the busiest, why there is more traffic in a certain location, are there correlations between price and host, and other questions that may develop during the data analysis.

Feature selection techniques, such as dropping columns that may have weak correlation to the target label, will be applied to identify most informative attributes for the prediction. After which, the deep algorithms will be applied. We will try to implement the deep learning neural network keras with RMSprop optimizer and cross-entropy loss. If we find any loss of validation in the dataset, it can be a categorical cross-entropy and can be used to predict the range in prices.