**Project Proposal**

Airbnb is an online platform that connects people who want to rent out their homes with people who are looking for accommodations in that locale. The project is about improving Airbnb host ratings. The existing system doesn’t help the hosts in Airbnb to recognize which features to contribute more to get a higher rating from their customers.Currently, the hosts have too many inputs to manage such as prices, facilities, policies, response rate and so on that they don’t have a clear idea on what should be the focus.

The model we are going to develop is kind of an optimization model using regression for the existing hosts to scale up business as well as a proposal for the new investors. The output of our model will provide insights for the hosts to figure out which features to concentrate on the most, also improve their ratings and their business.

Our regression models will help build a rating system in sharing economy business and appeal more potential customers effectively. For example, in the case of Uber, we can use a similar approach but different insights to build a model for their driver’s rating system to increase the revenue and customer satisfaction.

The goals are build several regression models to identify the key factors that contribute to host ratings and based on the result, provide hosts in NYC with actionable insights for future operations and take these insights as a reference to the Airbnb hosts across the world. The task flow is to collect historical Airbnb transaction data of NYC and move on to data pre-processing step including feature engineering such as imputing missing data, transforming data, dealing with skewed data, etc. As a model development process, we define proper cross-validation strategies, dividing the dataset into training, cross-validation and test datasets also comparing MSE as a model assessment approach comes next. Based on the model assessment, understanding which of the variables in the dataset are the key drivers for the ratings and are deriving insightful results would conclude the flow.

We quantify the importance of each variable to ratings by modeling techniques. Based on the output of the model, we would figure out how those variables would influence the host rating and which variable can bring the most impact. We can also generate insights for the hosts of Airbnb to increase their accommodation reservation rate at most.

**Expected problems and possible solutions:**

* The dataset is relatively large. Meanwhile, there are missing values in columns and have inconsistency in format, which might need a lot of data preprocessing work before building regressions. The ideal way to solve this problem is sampling data and handling missing values by removing or replacing them.
* The customer’s  comments are unstructured data and cannot fit in the model directly. Therefore, it’ll be challenging to detect its importance. We try to perform sentiment analysis to determine how positive or negative the review is.
* There are some outliers or noise in the dataset, and this will influence the accuracy of models. We find them by observing statistics of the data an standardize the data by normalization.
* The models might be overfitting or underfitting. For now, we plan to use cross-validation to prevent overfitting and underfitting.

**About the data:**

Dataset : Inside Airbnb’s New York City’s  ‘Detailed listings data & Review data’

Complied Date :12/06/2018  
Tool : Spark ML / scikit-learn / Pandas / PySpark / Tensorflow

Models : Regression Models with features of Airbnb Host & Amenities  
     ex) Multiple, Random forest, Gradient boosting, Deep Neural Network

Few other important ideas: comparing MSE, cross-validation

Dataset URL : <http://insideairbnb.com/about.html>