**Restaurant Entrepreneur: Predicting best type and location with Yelp Review Data**

Bill (an investor) went on a work trip recently to Las Vegas and Phoenix for a few days. During his stay, he really liked the restaurant options that were available. Being that he already had interests in opening up a restaurant for some time now, he wanted to know what the type of restaurant and where to open would be most successful.

As a first look, our group decided to use Yelp API to collect data on the mix of restaurants across cities and states within the United States. In this case, we decided to use the review data that Yelp is known for. Since Yelp provides a wide variety of restaurant categories, it is possible to get more information in regards to reviews and ratings for all different types in each city. After our initial attempts to use Yelp API, we found issues with importing the Yelp API data. When we created a dataframe from the API, it would only pull a small portion of the data on each run, not a sufficient amount of data to do a full analysis on. Fortunately, we were able to find some recent Yelp Review data on Kaggle.com that had a huge dataset to work with.

Before deciding on which Machine Learning Model we were going to choose to do our analysis on, we wanted to test a few to see what would be the best. We began setting up our code by importing various libraries, which included Random Forest Classifier and DeepLearning Machine Learning Models. Random Forest Classifier is a good model if you want high performance with less need for interpretation. Deep Learning Model is known for it's supremacy in terms of accuracy when trained with huge amounts of data and to get more neural network predictions. We've also imported train\_test\_split which will help us split our data for training and testing.

For the preprocessing, we imported StandardScaler and OneHotEncoder. The StandardScaler is needed to transform the data so that it has a mean of 0 and a standard deviation of 1. The OneHotEncoder is needed as it creates a binary column for each category type of restaurant.

Why we chose these specific models and how do they work specifically with our dataset?

* We chose Deep Learning, Random Forest Classifier and Logistic Regression as our Machine Learning Models. Random Forest is a good model for high performance with less need for interpretation. Deep Learning is known for it’s supremacy in terms of accuracy when trained with huge amounts of data and to get more neural network predictions. Logistic Regression is most useful when we want to predict the probability for a categorical response variable with two outcomes. In our case, we’re trying to decide on what category type of restaurant and where is best to open based off of reviews. A good review would be any scores between 3-5, bad would be 0-2.

Detailed description of preliminary data preprocessing

* We downloaded our data from Kaggle.com and cleaned it, then uploaded to postgres.

Description of preliminary feature engineering and feature selection

How was the data split into training and testing sets?

What was our model’s accuracy? What were the limitations and benefits of each model we chose?