## CKME 136 – Data Analytics Capstone Project

## Ahmedur Rahman

## Fall 2018

Abstract

Yelp has always been a trusted source for finding high rated restaurants around the places. Their dataset provides a huge list of data containing each restaurant’s features and their ratings along with some user details, review details and check-in details. A lot of studies have been conducted to do the sentiment analysis on review text and predicting the star rating. But none of them are focused on if there is a relationship between high rating and any of the restaurant features. In this project I will try to analyze which features of a restaurant are more likely to contribute directly towards higher ratings; this result could become a great source for business owners or entrepreneurs to improve their ratings.

The theme for this project is Classification and Regression. The Yelp Dataset contains over 1.4 million business attributes like hours, parking, availability, and ambience for 188,593 businesses in 10 metropolitan areas1.

I plan to use Hadoop with pig/ hive to load the dataset, use R language to preprocess the data. Then build at least two/ three classifier models using different methods (Logistic Regression/ Naive Bayes/ Random Forest) to find out important features (i.e. hours/parking/availability/location etc.) of a business to get higher rating. After that I am planning to mine association rules on high-rated businesses to get more insight about the features and how they are correlated. At a later point, I plan to run some statistical test (i.e. ANOVA) to find the significant difference between classifier models built earlier.