# Predicting Star Rating based on User Reviews

Vinoth Shankar
Supervisor: Mr. Joseph Kehoe
M.Sc. in Data Science

## Introduction

- Online reviews play a significant role for most people to get judgments.
- Nowadays what to buy, what movie we should watch, what to eat. Everything is governed by the opinions of other people.
- Websites including Yelp, Google and TripAdvisor have now become a huge database for places, restaurants, planning a trip that includes feedback and thoughts written by people every day.
- Most reviews are biased towards high ratings but most of the ratings are practically worthless considering that the majority range from 3 to 4 ratings, with very few lower or higher.
- So predicting text analysis scores would have more quantitative meaning than a star rating.

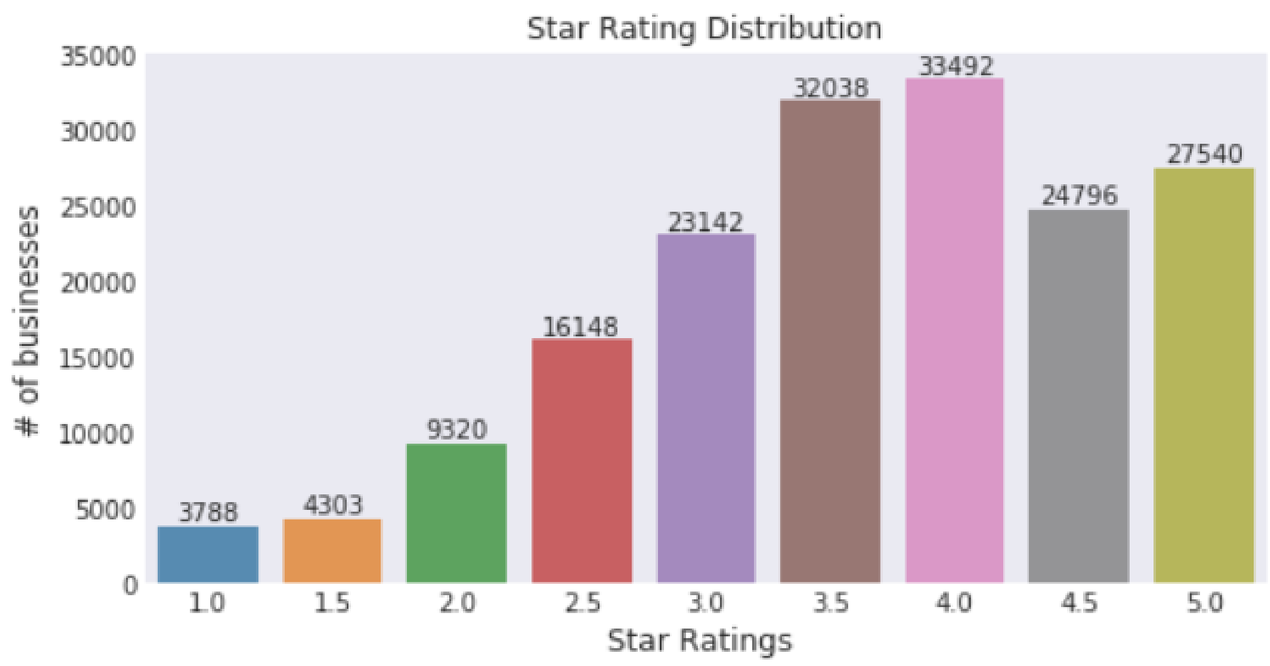


*Figure 1: Star Rating Distribution*

## Research Objective

- The project objective is to find whether rating predicted from the reviews and the actual rating provided by a customer correlate with each other.
- The ratings predicted from the reviews will be compared with actual star rating.

## Literature Review

- A novel method of classification of documents using a consistent metric called sentiment polarity score (sp-score) by using support vector regression (SVR).
- Suggest new ad hoc and regression-based recommendations that take both the textual portion of user feedback into consideration.
- Using textual information results in better average or customized performance estimates than the one derived from the users' numerical star ratings.
- In the existing two forms of review rating methods, one includes methods based on review text content and another based on collaborative filtering that collects information from past records in the reviewer-rating matrix.
- Proposing a framework for testing rating prediction on a neural network. The researcher presents user-word composition vector model (UWCVM) and documents the composition vector model (DCVM).
- That captures effectively how the consumer behaves as a mechanism that affects the continuous representation of words and incorporates them into a supervised rating prediction learning environment.

## The Dataset

- Yelp as the large dataset from businesses, reviews, and user data as JSON files.
- It has over 8,021,122 ratings, 1,320,761 suggestions from 1,968,703 users and over 1.4 million company attributes such as hours, parking, availability and aggregated time check-in atmosphere for each of the 209,393 companies.

## Methodology

- For my analysis, I will be targeting restaurant reviews. So separated the dataset based on their category.
- To train the dataset we used both text review and star rating to predict review rating and compare it with the star rating
- Algorithms such as CART model (Classification and Regression tree), LSTM (Long short-term memory), KNN, Naive Bayes are applied to dataset
- Accuracy of each model will be compared.

## Technologies

## Next Steps

- Extracting fake reviews from the dataset will add weight and the model will be more effective in training.
- On Future intended to generalize our model by looking at other areas of business such as fashion, beauty, and tourists.
- In future, the performance of the models needs to be improve by using Map Reduce.
- In addition, the GPU processor will significantly increase model performance.

Contact:

Vinoth Shankar
Phone: +353 894991194
Email: vinothshankar40@gmail.com

References:

- Yelp.com (2020). Yelp Open Dataset. [online] Available at: https://www.yelp.ie/dataset [Accessed 13 Apr. 2020]
- Bingkun Wang. ,Yongfeng Huang. ,Xing Li. (2016, January 3). Combining Review Text Content and Reviewer-Item Rating Matrix to Predict Review Rating.
- Daisuke Okanohara. ,Jun'ichi Tsujii. (2005). *Assigning Polarity Scores to Reviews Using Machine Learning Techniques.* Springer-Verlag.
- Duyu Tang. ,Bing Qin. ,Ting Liu. ,Yuekui Yang. (2015). User Modeling with Neural Network for Review Rating Prediction. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, (p. 1340).
- Gayatree Ganu. ,Noemie Elhadad. ,Amelie Marian. (2009). Beyond the Stars: Improving Rating Predictions using Review Text Content. *Twelfth International Workshop on the Web and Databases (WebDB 2009)*, (p. 1).