Yelp Reviews & Ratings Prediction

Introduction

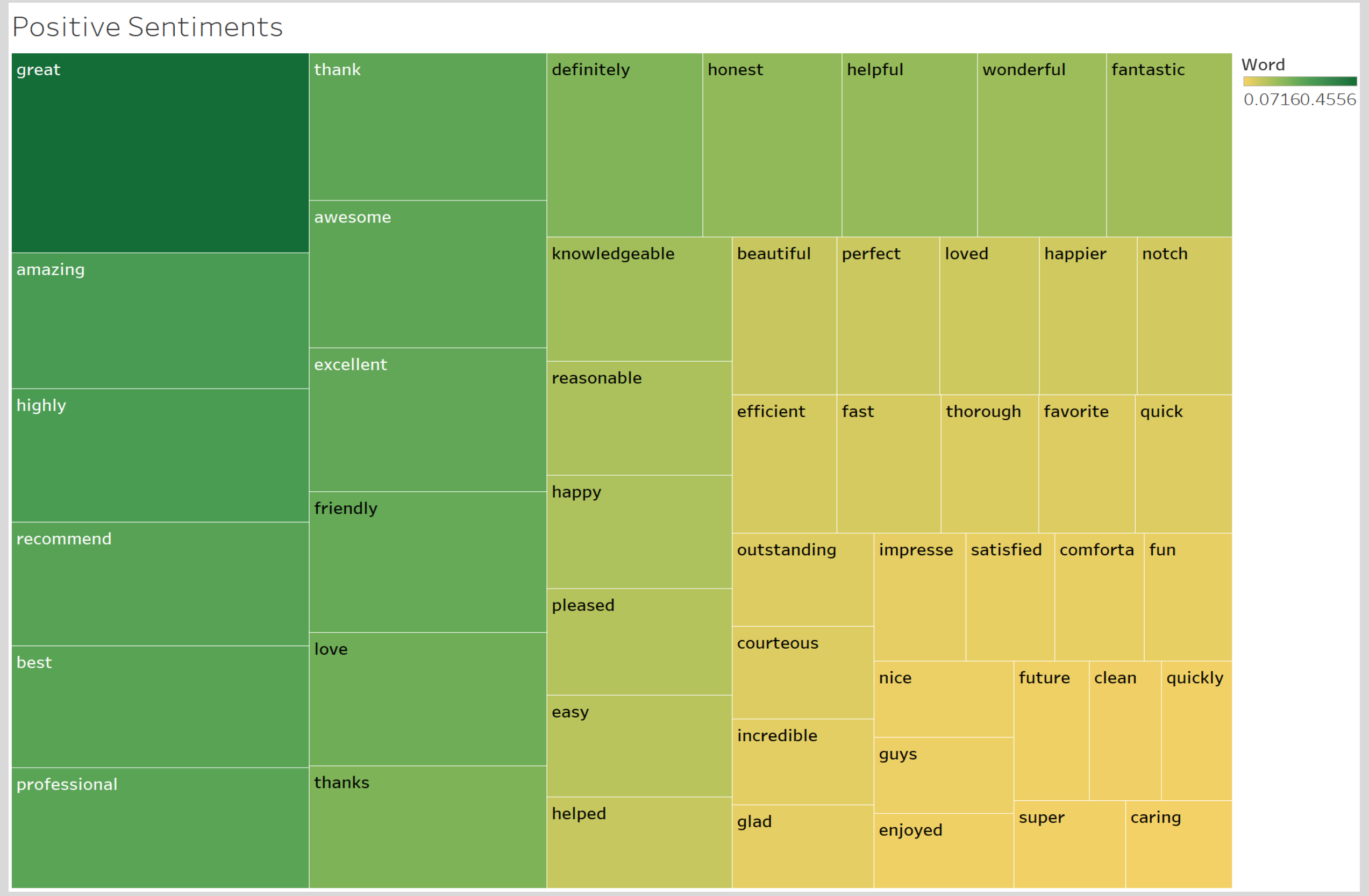
Yelp is one of the widely used platform by users to review businesses helping a user give a true review from an individual experience. The dataset consists of 4.1M reviews & 947K tips by 1M users for 144K businesses. It has 1.1M business attributes like hours, parking availability, ambience. The data also has aggregated check-ins over time for each of the 125K businesses. This dataset contains Tips, Users, Reviews, Businesses & Check-in tables.

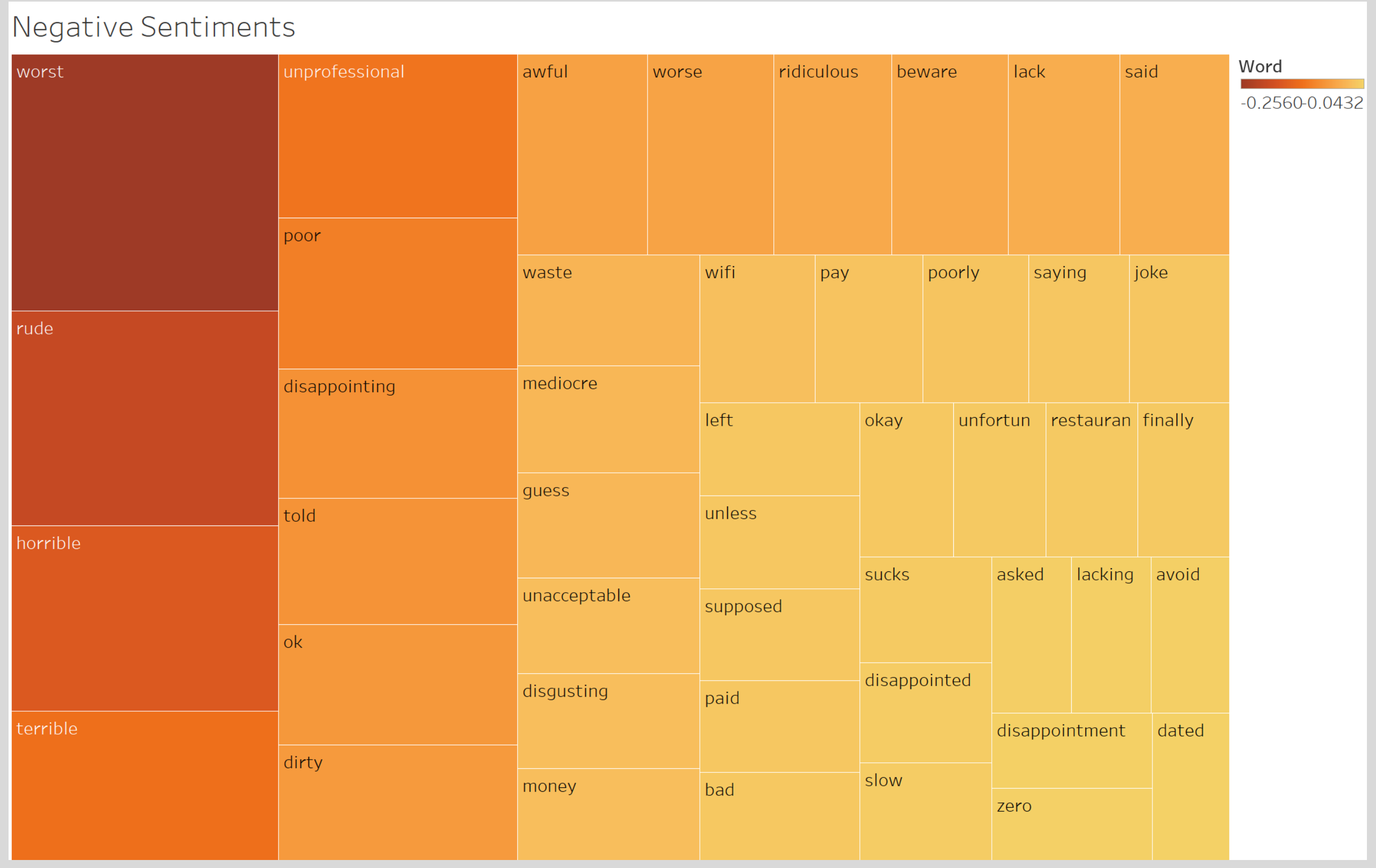
Business Case

How well can you guess a review's rating from its text alone?

What are the most common positive & negative sentiments used in the reviews?

What are the most popular businesses?


Positive Sentiments


Negative Sentiments

Platforms Used

Models: Linear Regression, Random Forest Regression

Language: Python

Packages: Pandas, Numpy, Matplotlib

Tools: Apache Spark, Jupyter, Hadoop

Sentimental Analysis

We used sentimental analysis to predict the most common key words found in positive & negative reviews using Logistic Regression

The Normal Approach

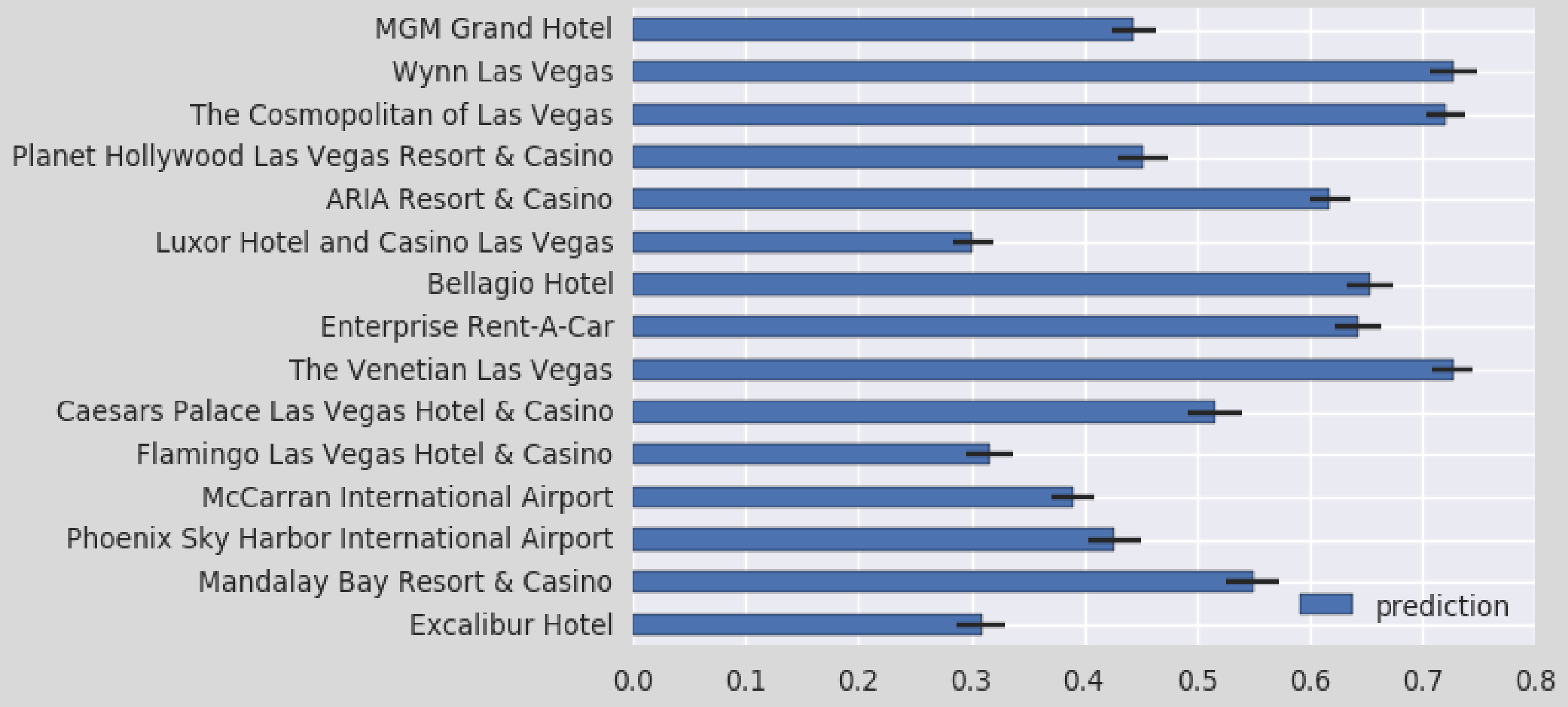
We used the normal approach to compare the words with positive and negative dictionary from Stanford.

Accuracy: 70.334%

Data Science Approach

We then applied Logistic Regression to the reviews & compared them with user ratings generating the best fit model.

Accuracy: 87.579%



Bar Chart showing businesses having high ratings & corresponding sentiment predictions

Review Ratings Prediction

We used Linear Regression model to predict the review ratings based on the features in the user, business & the review datasets. We split raw data into training, validation & testing part by a 60:30:10 ratio.

The MSE for model containing both user & business rating: 1.1930

The MSE for models containing only user rating: 1.6504 & only business rating: 1.5474

The MSE for model containing the user, business, funny, cool & compliment average ratings : 1.1439

Conclusion

We recommend the Yelp listed businesses to focus more on professional, be friendly, knowledgeable about customer & have clean & quick service to the customers.

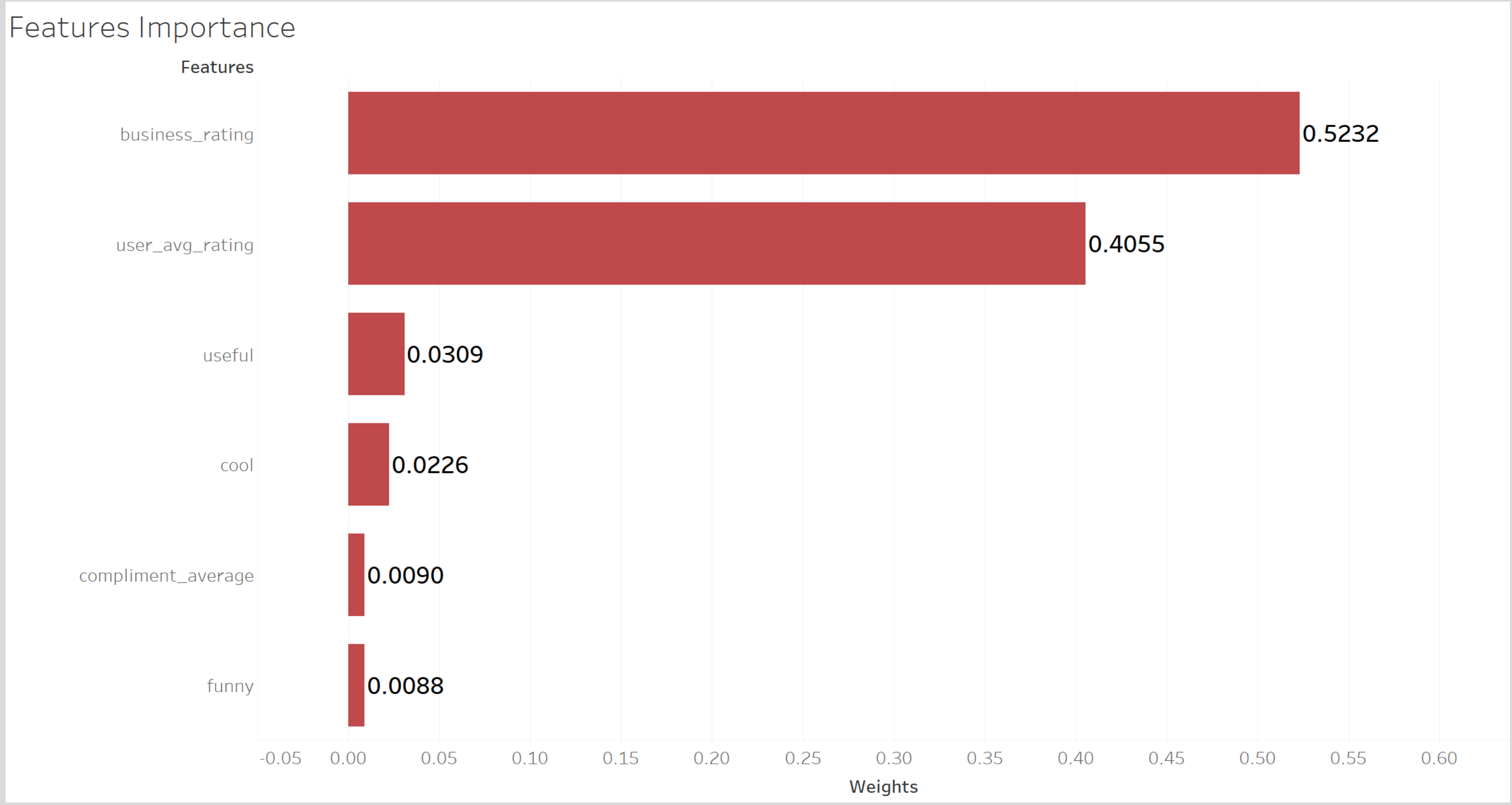
Alternately, users avoid restaurants & companies having mediocre & unprofessional services, no wi-fi, high charges & slow/disappointing services as noted by Yelp user reviews.

Average user ratings & business ratings are taken into consideration 9/10 times, when a review is posted.

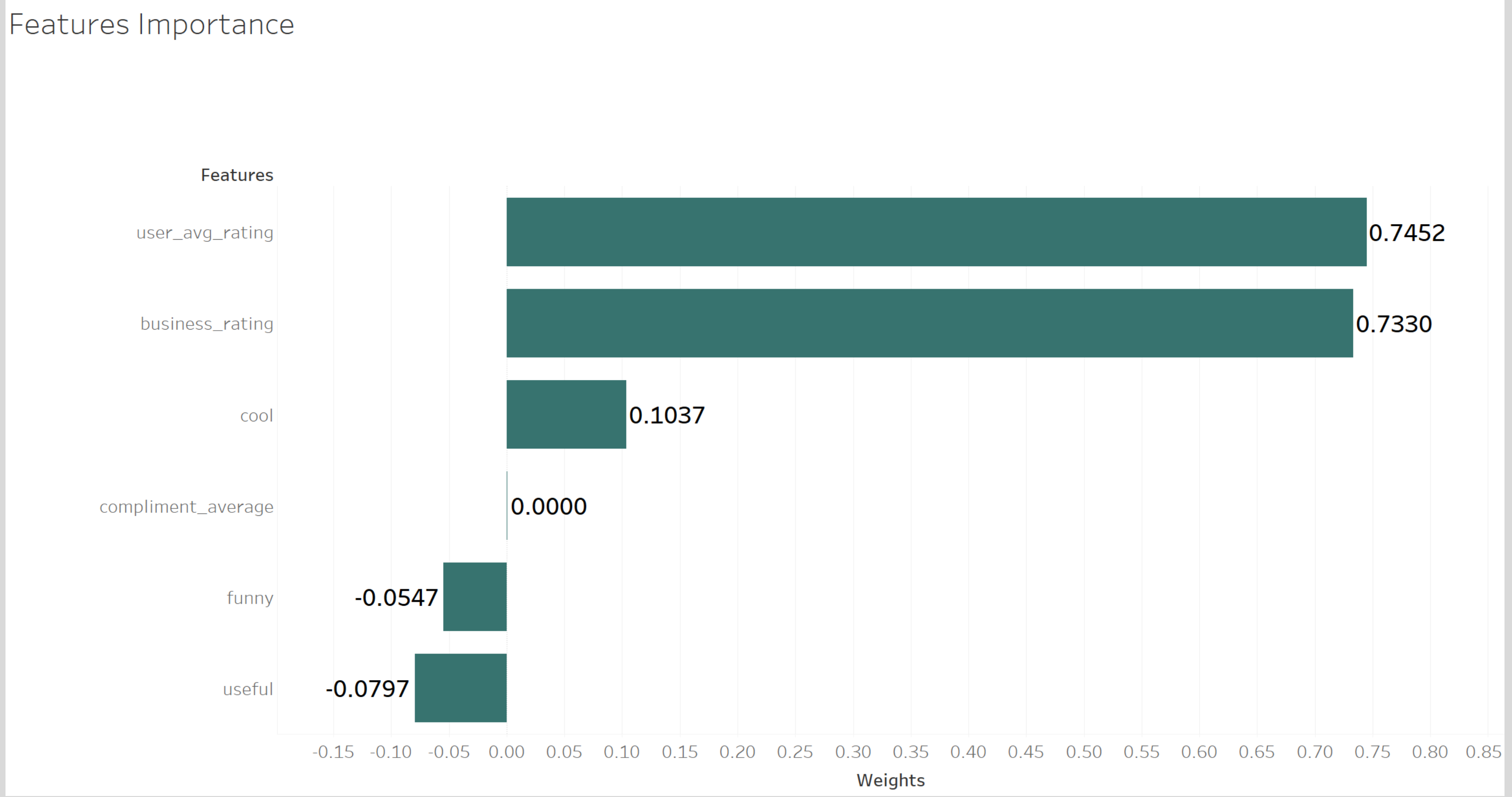
Review Ratings Prediction using Random Forest

We further applied Random Forest Regression model to predict the review ratings based on the features in the user, business & the review datasets. We split raw data into training, validation & testing part by a 60:30:10 ratio.

The features used to predict the review ratings are shown in the diagram below. We calculated the Binary Classification Evaluator for the model and the model accuracy achieved was 90.2151%


Features Importance

Bar Chart showing features corresponding to the weights predictions for Random Forest


Features Importance

Bar Chart showing features corresponding to the weights predictions for Logistic Regression