Determining Important Factors for High Yelp Ratings in Los Angeles

Wyatt Walsh

Data

- Wikipedia Neighborhood Data for Los Angeles Region
- Geopy Geographic Coordinate Data for Each Neighborhood
- Foursquare Venue Data for Each Neighborhood Coordinate Pair
- Yelp Venue Data for Each Venue

Methodology (1 / 2)

Goal:

- Leverage Machine Learning to enable a better understanding of what factors play into high Yelp Ratings

Idea:

- Use regression to establish an accurate predictive model and then evaluate feature significance based on model

Methodology (2 / 2)

Solution:

- Process data and analyze for inconsistencies
- Create randomized train-test split
- Train ML regression models on training set
- Evaluate models using MSE, MAE, R^2, and maximum error for a single sample
- Determine feature significance from Random Forest Impurity

Results

- Random Forest was best performing model (0.234 MSE)

- Number of ratings is by far the most important determination of overall rating

Discussion

1. Modeling could be improved through cross validation
2. SHAP could be used for additional significance testing
3. These results may not generalize at all to other regions

Thank you!