IBM Capstone Project

Wyatt Walsh

February 2021

Contents

1 Introduction

In many cases, customer decisions to purchase products or services can hinge on a business's online ratings. However, it can be difficult to parse what sort of business strategies can help bolster ratings, and furthermore, what key factors play into higher or lower rating outcomes. For this project, a Yelp rating prediction tool based on machine learning paradigms has been created as well as feature significance testing to determine which predictors are most influential for high ratings in the Los Angeles area of Southern California. Stakeholders interested in boosting their business's online ratings are certainly encouraged to read further.

2 Data

There are four sources of data utilized within this project. First, Wikipedia neighborhood data is scraped for the Los Angeles area. Next, Geopy is used to find geographical coordinates for each neighborhood. Using these coordinates, the Foursquare API is then queried to provide basic information on the businesses within each neighborhood. The Foursquare data contains business names and geographical coordinates, which can then be used to query additional business data from the Yelp API, which included ratings, categories, number of ratings and price information.

3 Methodology

In order to better understand the data, histograms and boxplots were created for the continuous variables. For the case of the response, a grouped histogram was also created to evaluate any changes in distribution across price levels. Furthermore, top categories for the region were also accessed. Finally, a map was generated with coloring depending on a venue's category and neighborhood.

In order to properly model the data, dummy variables were created for each categorical feature, dropping one of each group to avoid the dummy variable trap. From there, a train-test split (with a test size of 20%) was conducted and a slew of models was applied to the training set to see which may have the best fit. These models were then analyzed through the lens of four metrics as applied to the testing set: mean squared error (MSE) , mean absolute error (MAE), R^2, and maximum error across all testing samples. Feature importance was then assessed through the impurity scores of the Random Forest model.

4 Results

Out of the models tested, it appears that the Random Forest performs best in terms of MSE and R^2 (0.234 and 0.558 respectively), the Extra Trees model performs best in terms of MAE (0.309), and the Adaboost model had the least maximum error (0.186).

The top twenty-five most important features influencing ratings based off the Random Forest model can be found in the Appendix, table 7.

5 Discussion

From an examination of the most important features, it seems that review count plays most highly in determining a business' rating. I would venture to say that this finding is somewhat location agnostic, however more analysis is needed to make a final determination. Furthermore, the important features show a few commonalities: customers tend to rate middle of the road priced businesses more highly (this could be due to the fact that they have more reviews), businesses close to the center of their respective neighborhoods fair better in ratings, more angular, specifically defined businesses (businesses that exist in a single category) seem to perform better, and convenience in general is quite important.

6 Conclusion

This project certainly seems to have value, however improvements can be made. The Random Forest model (and perhaps Adaboost and Extra Trees) should have cross-validation applied such that more optimal hyperparameters can be sought out, thus improving the modeling accuracy. Further, additional data should be collected to increase the robustness of the findings.

7 Appendix

	Features	Significance Scores
0	Venue Review Count	0.26351479027159397
1	Venue Category: Fast Food Restaurant	0.05225571231122478
2	Distance to Neighborhood 0	0.03622942558070114
3	Venue Price: $$	0.03505524683968805
4	Venue Category 0: servicestations	0.0224543705783712
5	Venue Category 1: convenience	0.021031089971745763
6	Venue Category: Pharmacy	0.019997302590209466
7	Venue Category: Discount Store	0.019881923538804493
8	Venue Price: $$$	0.019809889334400817
9	Venue Category 1: hotdogs	0.018674949084110767
10	Venue Category 0: drugstores	0.018002785022634568
11	Venue Category 2: none	0.016268271264053925
12	Venue Category 0: mexican	0.013049188719269411
13	Venue Category 0: hotdogs	0.012651158125353097
14	Venue Category: Pizza Place	0.011760487905228099
15	Venue Category 0: deptstores	0.011749912041672257
16	Venue Category: Grocery Store	0.010687408729728835
17	Neighborhood 0: Panorama City	0.008569294572579989
18	Venue Category 0: pizza	0.007646726459372984
19	Venue Category: Ice Cream Shop	0.00673511071774266
20	Venue Category 1: none	0.0065578866155325734
21	Venue Category 0: grocery	0.006100590540913762
22	Neighborhood 0: University Hills	0.005819573107668515
23	Venue Category: Convenience Store	0.005118249342810683
24	Venue Category 1: tradamerican	0.004894900146275748

Table 1: Top twenty-five most important features by Random Forest impurity