

Report

I think that the model predicts stars well. It seems to do especially well when it is predicting stars at the level of 2 and 4 stars but appears to slack off some at extremes and for restaurants that rank in the very middle. Overall, the model works very well for predicting restaurants more generally. I think that the model would perform best classifying businesses into buckets like good bad and okay, I was able to get a little closer to this type of classification by rounding to the nearest whole star.

Below are prediction results a confusion matrix and the accuracy score.

```
[4. 5. 2. ... 5. 4. 5.]
[[ 43  240  99  548  124]
 [ 102 1470  714 4430  565]
 [  63  877  569 3429  363]
 [ 142 2486 1702 13980 2123]
 [  42  552  288  3477 1572]]
0.44085
```

The results window shows that the randomforest machine receives an accuracy score of 44%. I was able to get accuracy scores up to 54% by adjusting the `n_estimators` and `max_depth` but the confusing matrix did not look as nice.

The confusion matrix here shows us that the model does better at predicting results for businesses around 4 stars and 2 stars and does relatively well overall.

I think that a random forest is a good choice for this dataset because random forests do not suffer greatly from multicollinearity problems. I used a lot of dummy variables to convey the differences in the businesses to the machine, some of which are possibly heavily correlated like valet and classy, classy and upscale, and pizza and takeout. The model also seemed to perform

well comparatively, for example, I ran a linear regression model also and found that the predictive power was much lower (judging by the R^2 measurement, also the code for the linear regression is included in linearcomparison.py)

Overall, I am happy with the model. If I had more time, I would make the model more specific to a city or state by deciding map ranges based on latitude and longitude and cluster the data to see for which locations stars cluster together (metro, downtown, financial) to gain more predictive power. Although the model could be improved upon, I think that it does fairly well predicting stars, and would be very effective classifying businesses into buckets like good bad and okay.