

Report

I think the model predicts the number of stars earned well. It seems to do so especially well when predicting at the level of 2 and 4 stars and appears to slack off a little at extremes and for businesses that rank in the very middle. Overall, the model works very well for predicting restaurants more generally and would likely 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 random forest 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` however the confusion matrix did not look as promising. The confusion matrix pictured here shows us that the model does better at predicting results for businesses around 2 stars and 4 stars and does relatively well overall.

A random forest model is a strong choice for this type of 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 performs well

comparatively, for example, I ran a linear regression model also and found that the predictive power of the linear regression model 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 as well as add some observations based on where stars cluster together (metro, downtown, financial). Although the model could be improved upon, it predicts the number of stars pretty well and would be very effective classifying businesses into buckets like good bad and okay.