Models:

The team decided to take a couple different approaches for the predictive section of the project. First, the team ran a series of regression models to predict the reviews and then ran a series of classification models to predict if the review was good or bad. The team also evaluated a predictive classification model to determine if a restaurant was open or closed.

For the regression models, the team made use of both datasets provided. The restaurant statistics and the review statistics. The result was 18 variables to predict number of stars given to the restaurant. All the data was converted into integers and most are binary (1 – yes or 0 – no). For example, if a restaurant has restaurant hours on Tuesday, the team turned that data element into a 1, if there were no hours provided for the restaurant on Tuesday, the team turned that data element into a 0. From here, the data was prepped and ready to split into a 70% training (2,098,145 rows) and 30% testing (898,787 rows) data set.

4 Regression models were executed and evaluated (Linear Regression, Decision Tree, Random Forest, and Gradient Boosted Tree). All models performed relatively the same, the GBT had the lower RMSE, but if the team had to decide on the best model. The team would lean towards the less complex.

4 Classification models were executed and evaluated to predict if a restaurant would get a good or a bad review (Logistic, Decision Tree, Random Forest, and Gradient Boosted Tree). Of all the models, the best performing in terms of accuracy is the GBT, but between the best performing and worst performing there is only a 5% difference.

4 Classification models were executed and evaluated to predict if a restaurant is open or closed (Logistic, Decision Tree, Random Forest, and Gradient Boosted Tree). Of all the models, the best performing in terms of accuracy is the Decision Tree at 86.2%. There is a significant drop off for the more complex Random Forest and GBT models.