Add on for the predictive modeling part:

One important output of the random forest model is feature importance. Feature importance provides important information about how each features contributes towards the final model and quantify this contribution based on average information gain for each learner in the forest. Following graph shows the feature importance in a subset of data.



Based on above plot, the top three important features listed above are average price, noise\_level and gym. The top features correspond well with the association rule mining and exploratory data analysis. However, another observation is that there is no strong determining factors in drawing the classification decision boundary.

Discussion and Conclusion

To sum up, the restaurant data is collected in three major dimensions, restaurant facilities, nearby locational information, and demographical information within the neighborhood. In addition, business review data are collected for text analytics. The data analytics are conducted in multiple dimensions including clustering, association rule mining, PCA analysis and multiple classification algorithm including SVM, Logistics Regression, Naïve Bayes, Decision Tree, Random Forest. Moreover, sentiment analysis and topic modeling are performed on the review data. Following insights are drawn from those analytics:

1. Out of the three dimensions, restaurant facilities are the most important dimension, followed by nearby facilities, and the demographical information contributes litter to the success of the restaurant.
2. Among all the restaurant internal features, price and noise level has the most significance. The higher the price, the higher the rating could be. This meets the assumption of the data is biased towards positive review. In addition, customers are sensitive about the noise level, which means they pay attention to the privacy and atmosphere during dining.
3. Among all nearby faculties data, the most important factor is number of gyms. And the association is positive. This might relate to the increasing consciousness about healthy food.
4. The ratings of the restaurant are predictable, but the decision boundary is only clear between good restaurant and the rest. No models achieve good separation between moderate and poor restaurants.
5. Out of all predictive models, Random Forest achieved best restaurants, AUC value for good restaurants prediction is 0.90. ‘Price’,’Noise\_level’ and ‘Number of gyms’ are the most important factors towards the good ratings of restaurants.
6. Sentiment analysis achieved 92% prediction precision, however the model is biased towards positive reviews. As there are much more false-positive predictions than false-negative predictions.
7. Topic modeling results reveal that the top topics generally contains positive adjectives, which suggest that the business review data is biased towards positive.

Based on the analysis above, there is no most significant factors determines the success of a restaurants and the data being collected is somewhat biases towards positive side.

Future analysis will focus more on extracting meaningful information from the business review data per restaurants using text mining and natural language processing techniques. Business reviews can be given more specifically by analyzing the positive and negative reviews from customers.