Yelp Restaurant Rating Prediction in NYC

J.Dong   Z.Hu   Y.Pan   Z.Wang

BROWN

Introduction

Since the coronavirus outbreak has brought many negative impacts on the economy, it will be crucial for a future restaurant owner to choose a restaurant business plan that is expected to bring good profits and reputation, both are dependent on restaurant rating. Therefore, we are interested in the following prediction task: given a restaurant's longitude, latitude, transaction types, price level, categories and how common the categories are in the same region, we wish to predict its Yelp rating.

Data

We collected our data from the Yelp website using its official business search API and scraped business attribute information of all restaurants in NYC because of the city's diverse and dense restaurant population. We used zip code as our location parameter and requested up to 1000 restaurants for each zip code. Removing duplicate restaurants and those with missing values left us with 5977 samples for modeling. We coded transaction types and 23 popular categories into indicator variables, and calculated 23 category ratios (number of restaurants in each category in the zip code area/total number of restaurants in a zip code area). Our final data table has rows with the following 53 fields of interest: rating, longitude, latitude, reservation, pickup, delivery, price level, 23 categories and their ratios.

Methods

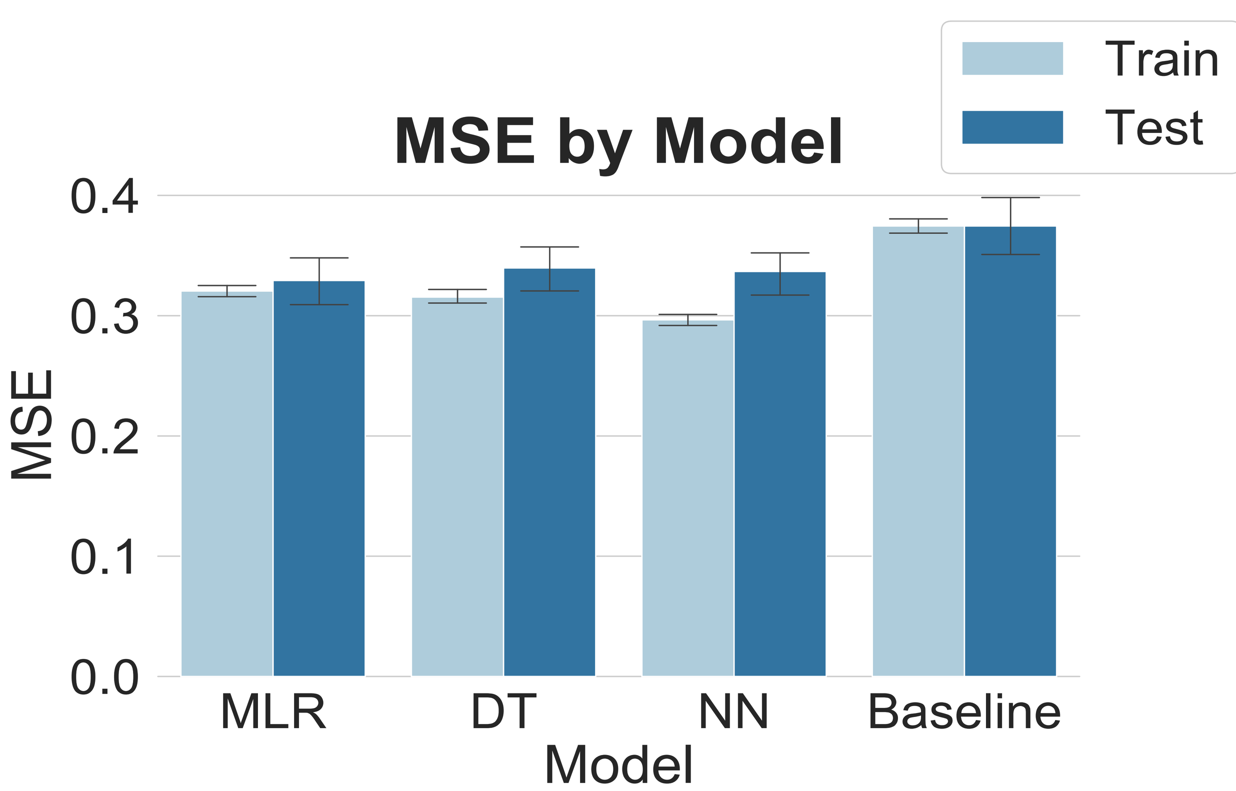
We focused on regression models and used three different types of models for this rating prediction task - Multiple Linear Regression (MLR), Decision Tree Regression (DTR), and Neutral Network (NN). MLR models were used to explore linear relationships between rating and different combinations of features. DTR models with a maximum depth of 5 were used to explore more complex relationships between rating and the features. In addition, we used a two-layer fully-connected NN trained to capture any non-linearities within the data.

We used MSE in our evaluation metric since we were interested in how close the predicted ratings would be to actual ratings. Considering our small sample size, to achieve higher accuracy we calculated each MSE as the average of a 5-fold cross validation. Since there was no sign of overfitting, we didn't focus on regularized models.

Performance Analysis

We used all 53 features to train our prediction models and the result shows that they outperform the baseline model (a mean predictor of rating).

We can see from the figure below that all three types of models have smaller train and test MSEs than the baseline.



Predictive Power Comparison

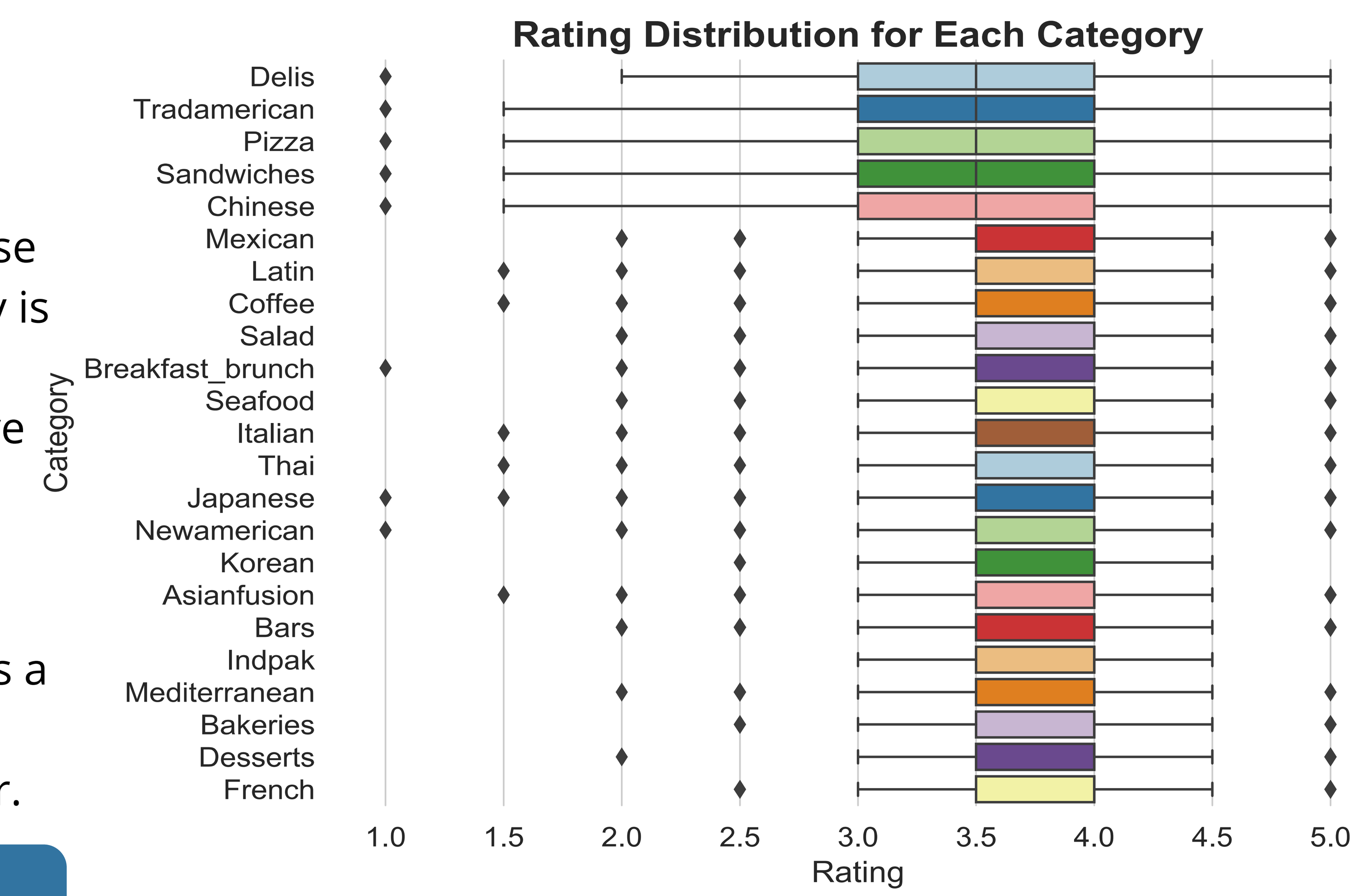
To examine how different features contribute to rating prediction, we splitted features into 5 groups: category, location, transaction type, price, and category ratio. 5 versions of all three types of models were built, with each version having one group of features taken out from the full model.

Test MSE by Model table shows that removing category from the full model leads to the largest increase in test MSE in MLR and DTR; removing location and transaction type leads to the largest decrease in test MSE in NN. However, none of the smaller models imply significant increase or decrease in test MSE. The insignificant increase after category is removed also aligns with the Rating Distribution by Category table - that category does not seem to have an impact on rating.

Restaurant category has a slightly better predictive power than other features, and transaction type has a slightly worse predictive power, but none of the features seem to have a significant predictive power.

Test MSE by Model

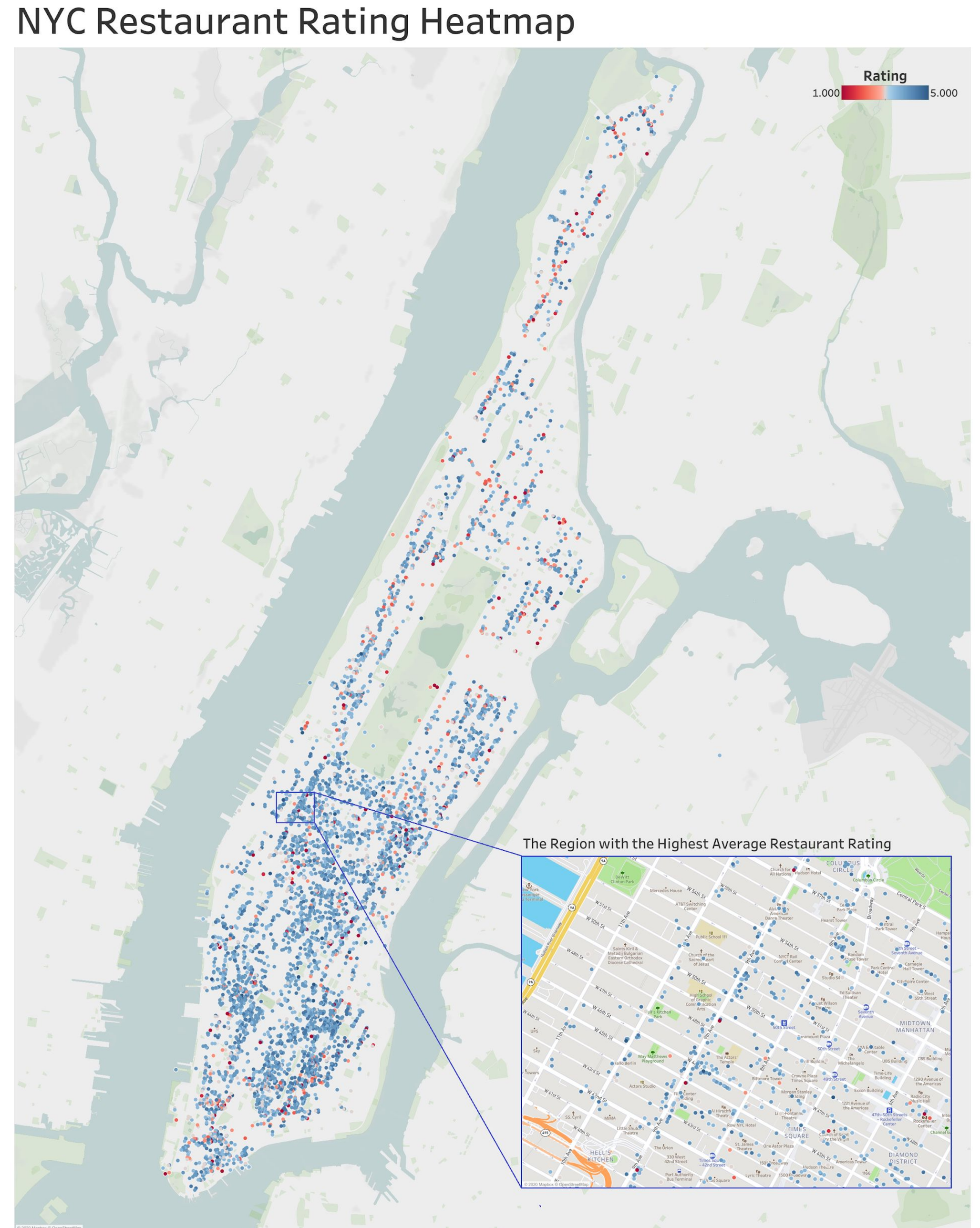
	All Features	W/O Category	W/O Category Ratio	W/O Price	W/O Location	W/O Transaction Type
MLR	0.329	0.342	0.336	0.33	0.33	0.334
DT	0.34	0.345	0.342	0.341	0.337	0.34
NN	0.316	0.353	0.344	0.341	0.401	0.333
Baseline	0.375	0.375	0.375	0.375	0.375	0.375

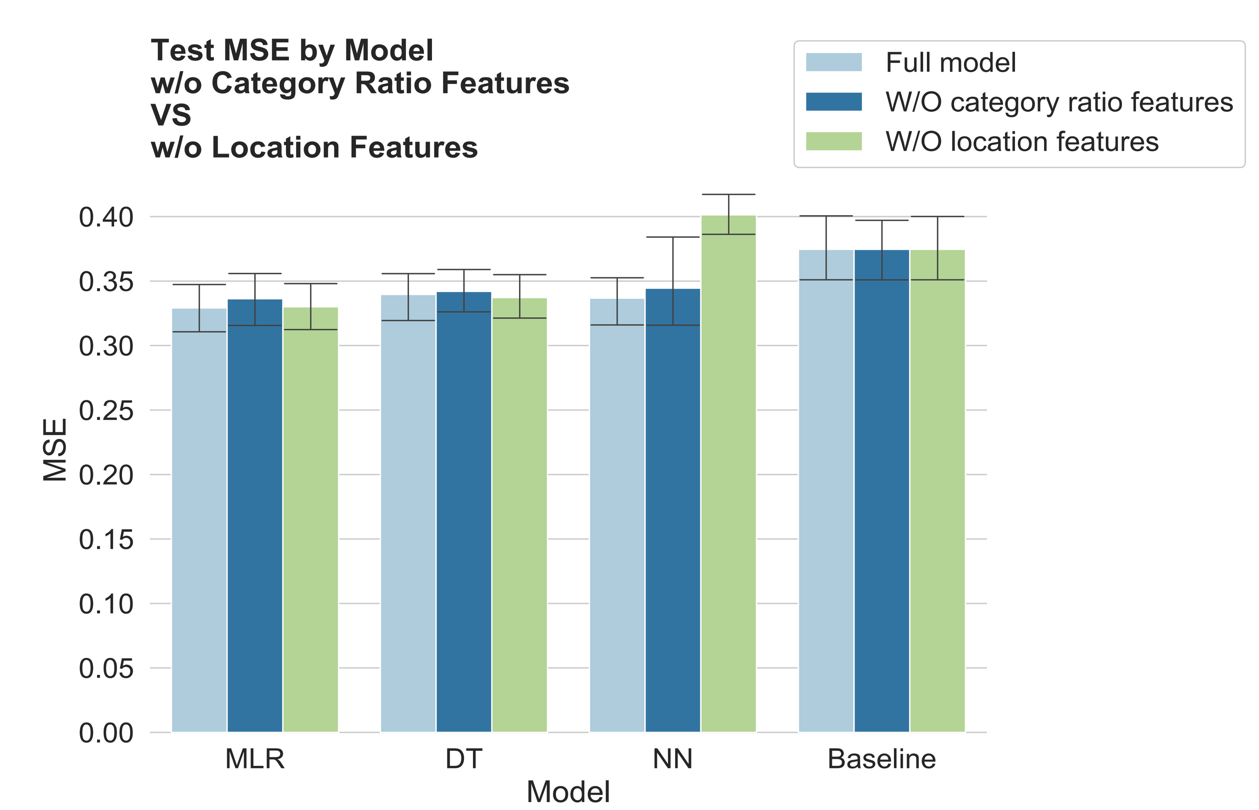


Location Analysis

The rating heatmap below suggests a correlation between simple location features (longitude and latitude) and rating.

NYC Restaurant Rating Heatmap





Besides simple location features, regional preference could be another location factor affecting restaurant rating. We looked into the distribution of restaurant categories in each region by calculating category ratios for each zip code area. The motivation is that category ratios could be useful for defining the type of a region in terms of category preferences, and might be able to replace longitude and latitude. We can see from the Test MSE by Model figure that, compared to the full model, the model without location features does slightly better in MLR and DT, but has an overfitting problem in NN. Therefore, even though ratio does help improve prediction performance, it cannot replace longitude and latitude.

Conclusions and Limitations

Our analysis shows that none of the existing features contain powerful signals for rating prediction, which explains why all three types of models (both linear and nonlinear) just slightly outperform the baseline model. We believe the bottleneck lies in three aspects:

- The basic version of Yelp API we used to scrape data only provides the most basic restaurant information.
- Business attributes alone might not contain key information for rating prediction.
- NYC is too small to demonstrate how local preferences and behaviors affect review rating.

Future Works

- More business attributes that might have a significant impact on restaurant rating may be explored, such as the amenities a restaurant provides, and whether it has parking lot, accepts credit card, or has WiFi.
- Non-business information such as demographics data could also be incorporated into the models.
- Inclusion of more cities and regions in the sample may provide greater insight into the significance of physical location on restaurant rating.