CS663 Data Challenge, Kei Fukutani, May 13th, 2020

Predict inspection scores

1. Latitude and Longitude
   I filled the missing Latitude and Longitude with their median value.

2. Creating Grading Point ('grading_point')
   Grading Point is a point that is the mean of the inspection scores of the same business name. Min-max scaling is applied to this column.

3. Defining the ordinal scale of 'risk_category' column
   I filled the missing value with 'Good'.
   Map 'High Risk' to 0, 'Moderate Risk' to 1, and 'Low Risk' to 2, and 'Good' to 3.

4. Creating 'street_name'
   I roughly extracted street name from 'business_address' column. Before one-hot encoding, The rare street name, whose number is small, is grouped together.

5. One-hot encoding
   I used One-hot encoding for several columns.

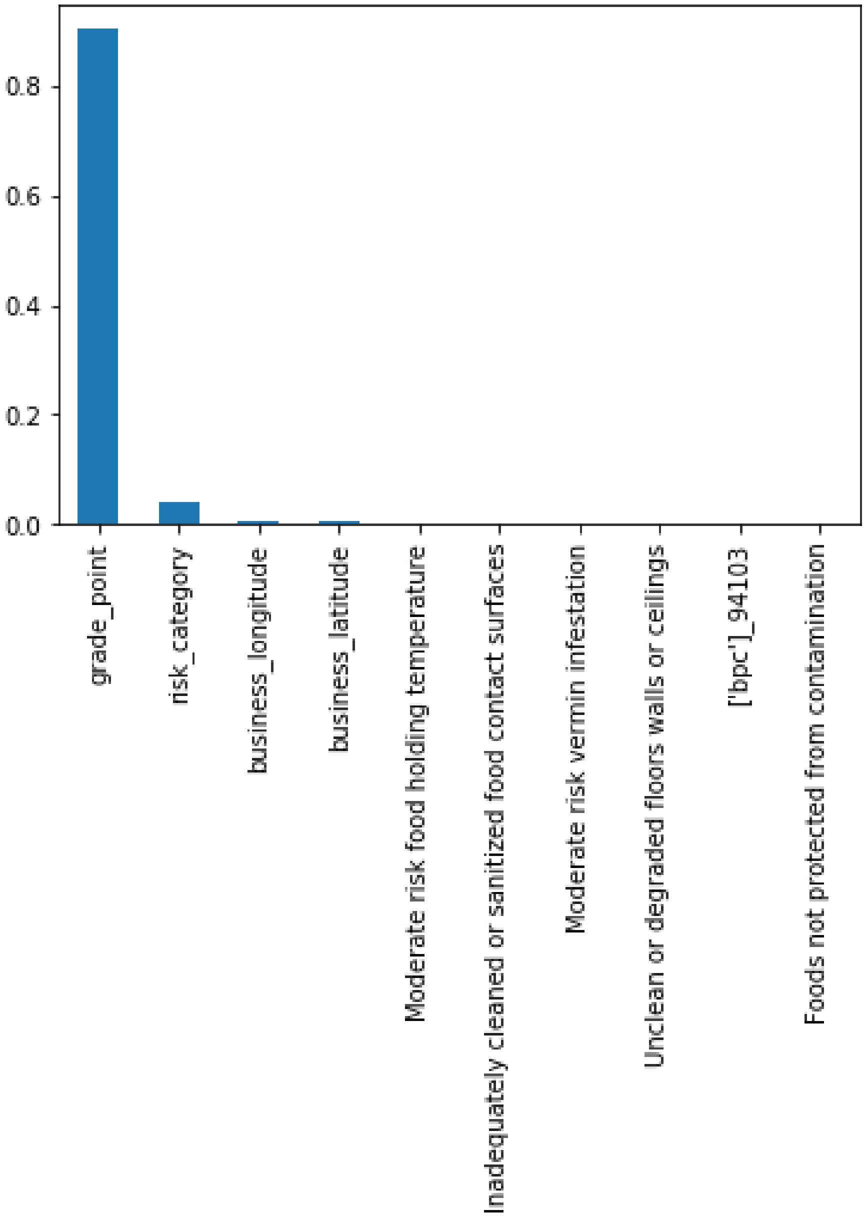
6. Random Forest
   I used Random Forest regressor to predict the scores. Grid search was conducted using the parameters:

   - 'n_estimators': [400, 800],
   - 'max_depth': [20, 40, 80],
   - 'min_samples_split': [4, 8, 16],
   - 'min_samples_leaf': [1, 2, 4].

   The best parameter were:
   - 'n_estimators': 800,
   - 'max_depth': 20,
   - 'min_samples_split': 16,
   - 'min_samples_leaf': 4.

Explain inspection scores

The Random Forest model provides feature importances attribute, which indicates what features contributed the scores relatively. Looking at the figure below, the ‘grading_point’ is the most important feature. The next is the ‘risk_category’.


grade_point
risk_category
business_longitude
business_latitude
Moderate risk food holding temperature
Inadequately cleaned or sanitized food contact surfaces
Moderate risk vermin infestation
Unclean or degraded floors walls or ceilings
['bpc']_94103
Foods not protected from contamination