Prediction of Durham County Department of Public Health Inspection Scores

Capstone 2 Report

By Temilola Famakinwa

Problem Statement

Food safety has become top of mind for many food establishments following the SARS-CoV-2 outbreak in 2020. Despite food establishments' current interest in improving this aspect of their operations, it remains a challenge area for them.

Food service establishments are run by franchisees, corporate offices, or sole owners. The myriads of ownerships in franchises and corporate establishments means management varies at the store level even within a single company; despite having knowledgeable food safety teams within the corporate company structure, this does not translate to successful implementation of hygiene procedures for food safety at the store level. For locations where there is a sole owner (one owner-one shop e.g. 'mom-and-pop' stores), sometimes there is a lack of food safety expertise. These factors contribute to establishments incurring health department violations and low inspection scores.

This analysis attempted to predict inspection scores based on health inspection comments and store information. Some questions which were explored were:

- Determine the most frequent health department violation in establishments
- Predict health department scores or the next violation based on establishment information and inspection comments
- Find opportunities for improvement for the 5 worst performing food service establishments.

The United States uses the FDA Food Code as its basis for regulation of public health when food is presented to a customer. However, each state in the United States has autonomy to implement a specific version of the Food Code and its contents as the state government deems appropriate. Hence inspection content and method of scoring varies from state-to-state. In order to simplify the analysis, inspection data was limited to Durham county in North Carolina.

Data

There were five data files that were inspected for the project. Below is a description of each data file.

Link/Name of File	Source	Description
Food-inspection-violations_1.json	Data World	Food health inspection violation data up until 2/21/20217
Restaurant-and-services_3.json	Data World	Restaurant information as a list
Restaurant-and-services_4.json	Data World	Restaurant information as a dictionary
Food-health-inspections_3.csv	Kaggle	Restaurant and violation information
Durham_County_Boundary.shp	Durham, NC Open Data Portal	Durham County GIS Boundary

Data Wrangling

Three of the datasets were in JSON format, while the last two were SHP (for GIS data) and CSV file formats, respectively.

Food-inspection-violations_1 was intended as source of inspection comments, information about the violation, and the establishments in which they occurred. It was supposed to be joined to the restaurant-and-services data file on the 'id' column. There were 158,623 rows, and 7 columns.

- The JSON file was converted to a pandas data frame.
- It was originally a list of dictionaries, so it converted into a table format with a data frame transformation.


```
[{'rpt_area_desc': 'Food Service',
  'comments': '2-301.12 Cleaning Procedure - P2-301.14 When to Wash - PFOOD EMPLOYEES MUST WASH HANDS FOR AT LEAST 15-20 SECO
NDS AND TURN THE WATER FAUCET OFF WITH THE PAPER TOWEL.  FOOD EMPLOYEE WAS OBSERVED HANDLING READY TO EAT BREAD AFTER HANDLIN
G RAW GROUND BEEF WITHOUT WASHING HANDS AND CHANGING GLOVES. CDI',
  'item': '6',
  'inspection_id': '2385570',
  'critical': None,
  'weight_sum': 2.0,
  'id': '5218844'},
 {'rpt_area_desc': 'Food Service',
  'comments': '3-501.19 Time as a Public Health Control - P,PFWORKING SUPPLY EGG BIN WAS NOT LABELED IN ACCORDANCE WITH THE W
AFFLE HOUSE TIME AS A PUBLIC HEALTH CONTROL POLICY.  ANY EGG BINS WHERE IT CANNOT BE PROVEN WHEN THEY WERE REMOVED FROM REFRI
GERATION MUST BE DISCARDED.',
  'item': '22',
  'inspection_id': '2385570',
  'critical': None,
  'weight_sum': 1.0,
  'id': '5218847'},
```

- The unique values of establishment groups were found in this data frame. Those related to food establishments (‘Mobile Food’ and ‘Food Service’) were sliced out and saved as a separate data frame called df1_food.
- The percentage of missing values in df1_food was determined but none were dropped because the data was eventually not used for modelling.

Restaurant-and-services_3 and Restaurant-and-services_4 contained information on restaurants and services in Durham county. The addresses of the restaurants, their capacity, phone numbers, type of establishments, inspection frequencies, geolocations, smoking policies, etc, were included in this dataset. The intent was to join either with the food-inspections-violations_1 dataset using the ‘id’ column. Both datasets contained 2463 rows and 24 columns.

Restaurant-and-services_3

- The JSON file was converted to a pandas data frame; it was originally a list of dictionaries, so it converted into a table format with the data frame transformation.

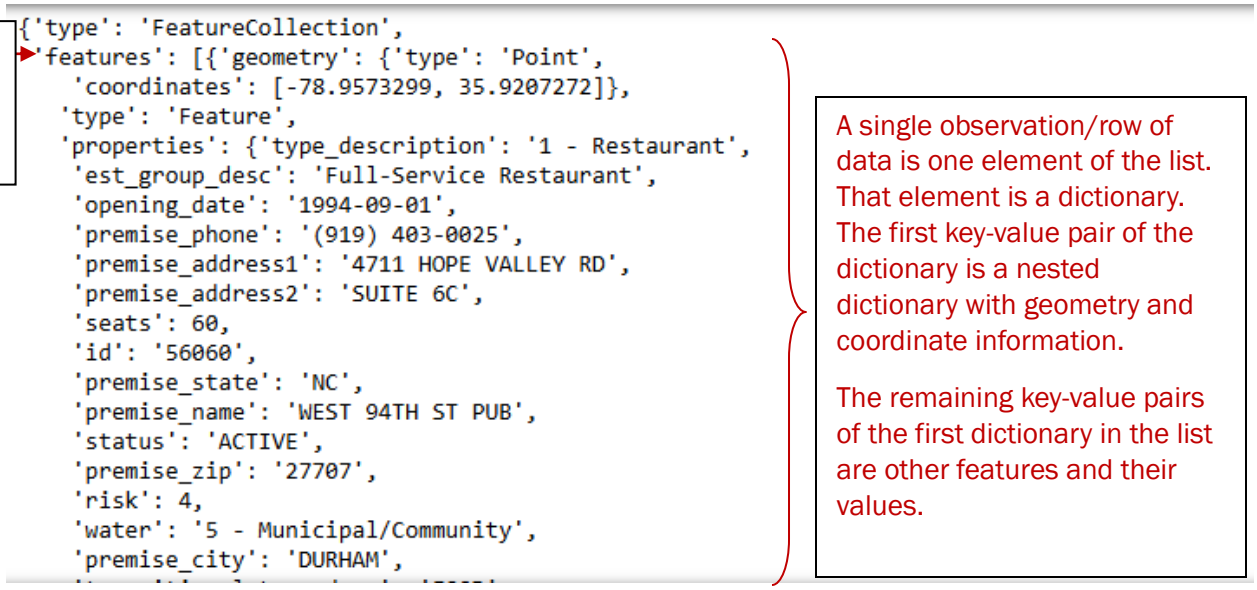

```
[{'geolocation': {'lat': 35.9285504, 'lon': -78.9237549},
  'type_description': '44 - School Building',
  'est_group_desc': 'Elementary School',
  'opening_date': '1994-11-15',
  'premise_phone': '(919) 560-3972',
  'premise_address1': '2320 COOK RD',
  'premise_address2': None,
  'seats': 0,
  'id': '58780',
  'premise_state': 'NC',
  'transitional_type_desc': 'N/A',
  'status': 'ACTIVE',
  'premise_zip': '27713',
  'risk': 0,
  'insp_freq': 1,
  'premise_city': 'DURHAM',
  'premise_name': 'SOUTHWEST ELEM SCHOOL',
  'water': '5 - Municipal/Community',
  'rpt_area_desc': 'School Buildings',
```


- The geolocation column was dropped because there were already two other columns, geolocation.lon and geolocation.lan, that had the geolocation information.
- The ‘geolocation.lon’ and ‘geolocation.lan’ columns were renamed to ‘longitude’ and ‘latitude’, respectively.

Restaurant-and-services_4

- The JSON file was converted to a pandas data frame; it was a dictionary. The highest-level dictionary had two key-value pairs. The value of the second element of the main dictionary was a list of nested dictionaries. One nested dictionary corresponded to one row of data. This was unravelled to create a data frame of observations and features.

2nd key of the highest-level dictionary



A single observation/row of data is one element of the list. That element is a dictionary. The first key-value pair of the dictionary is a nested dictionary with geometry and coordinate information.

The remaining key-value pairs of the first dictionary in the list are other features and their values.

- The values of the 'features' key were saved as a dictionary. This values list was converted from json to pandas data frame format.
- The 'properties.geolocation' column was dropped because it was empty and the information was already stored under the 'coordinates' key.
- The type, geometry, and geometry.type columns were also dropped because they did not contain useful information for the analysis.
- A function (split_list) was created which read each row of the geometry.coordinates column, and if it contained values, it created new columns named 'longitude' and 'latitude' and populated them with their respective values. If they were empty, it filled the rows with 'None'.

The columns and content of **Restaurant-and-services_3** and **Restaurant-and-services_4** were re-organized to match each other and the content was compared. They were also checked for duplicates but there were none. When they were tested for equality, the result came back as False. However, both data frames had the same shape and visual inspection of entries also looked identical.

Based on this, the data frames were treated as identical for future analysis; the difference seemed to be that one data frame had information originally stored as a list while the other was stored as a dictionary. Only information from **Restaurant-and-services_3** was used for further analysis.

- The unique values of establishment groups in Restaurant-and-services_3 were identified. Those related to food establishments (''16 - Institutional Food Service', '2 - Food Stands', '1 - Restaurant','3 - Mobile Food', '6 - Edu. Food Service', '15 - Commissary (Pushcarts/Mobile Foods)') were sliced out and saved as a separate data frame called df3_food.
- The percentage of missing values in df3_food was determined.

- df1_food and df3_food data frames were joined on 'id' and I found that the 'id' values did not overlap; Nan values filled both sides of the data frames where the ID did not match. I checked the range of values for the 'ids' and they were as follows:
    - df1_food: 6,882,767 - 744,890
    - df3_food: 193,403 - 55,461

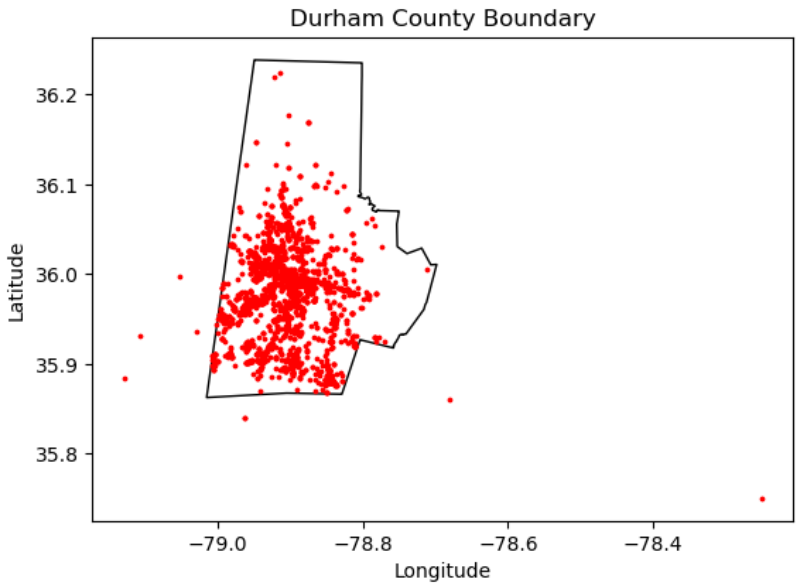
  Given this, a different data set (Food-health-inspections_3.csv) was found which contained both comments and restaurant information for the final analysis. However, exploratory data analysis was conducted on df1_food and df3_food.

Durham_County_Boundary.shp contained ARCGIS information about the boundaries of Durham county. It was overlaid with the positions of restaurants from **Restaurant-and-services_3** to show the distribution of restaurants in the county. No cleaning was necessary for this dataset.

Food-health-inspections_3.csv was used instead of Food-inspection-violations_1, Restaurant-and-services_3 and Restaurant-and-services_4 for the modelling steps because when it was found the latter datasets could not be joined by the 'id' column. Food-health-inspections_3.csv contained both inspection comments and restaurant information in one file. This dataset was used later during the project starting from the pre-processing stage before modelling. It contained 112,046 rows and 88 columns.

Exploratory Data Analysis

The Durham_County_Boundary file was used to map the boundaries of Durham County, and the map coordinates of each restaurant from df3 (Restaurant-and-services_3) was overlayed on it. The map showed that establishments are clustered in the lower middle section of the county. Towards the northern and eastern parts of the county, establishments become sparse. The data set also included a few establishments which are outside country borders. The center of the county corresponds to the location of Durham city.



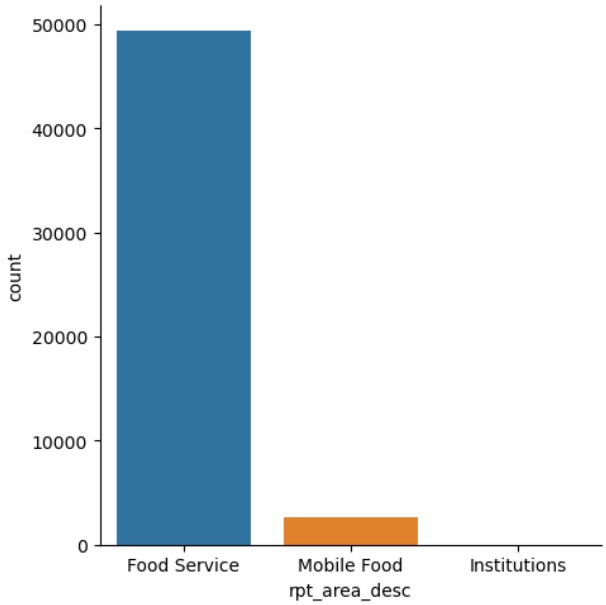
Next, df1 (Food-inspection-violations_1) was inspected. Many entries of the comments column started with a reference to a Food Code Section. However, these sections were not formatted uniformly. They contained special characters or incomplete section entries which lead to 16,000 unique sections. The following steps were taken to clean the food code sections

- The food code section was extracted from the 'comments' column using a str.extract method. The full section was listed and then broken down into Food Code Chapter, Food Code Subpart, and Food Code Section under the subpart.
- Rows containing Nan values under the new columns Food Code Chapter, Food Code Subpart, and Food Code Section were considered incomplete and dropped.
- The unique food code sections were inspected again. There were some values which were suspected to not be part of the food code so these were compared to actual food code sections from 2015 using the steps below.
- A csv file was created containing all the 2017 food code sections broken down into Food Code Part, Food Code Chapter Description and Food Code Subpart Description. This was saved as df2.

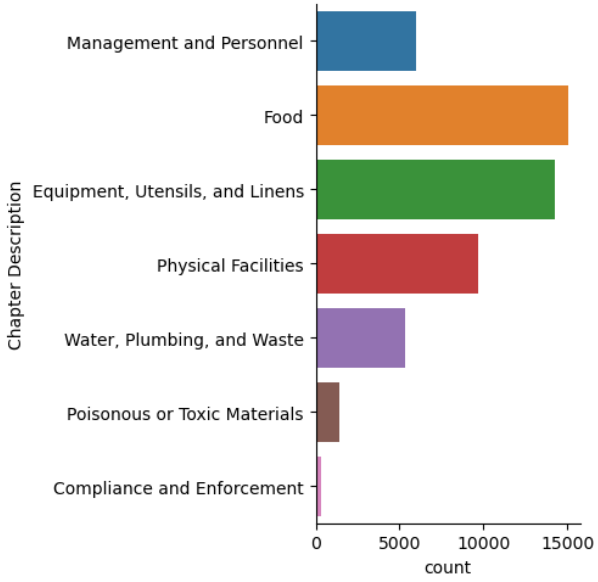
Part	Chapter	Chapter Description	Overall Subpart Heading	Subpart Heading
1-101	1	Purpose and Definitions	Title, Intent, Scope	Title
1-102	1	Purpose and Definitions	Title, Intent, Scope	Intent
1-103	1	Purpose and Definitions	Title, Intent, Scope	Scope
1-201	1	Purpose and Definitions	Definitions	Applicability and Terms Defined
2-101	2	Management and Personnel	Supervision	Responsibility

- Chapters, parts, and subparts from df1 which were in df2 were kept, filtering out sections that did not belong to the food code. This was saved as the cleaned df1.
- Df1 and df2 were merged by an inner join to get access to food code chapter and part descriptions.

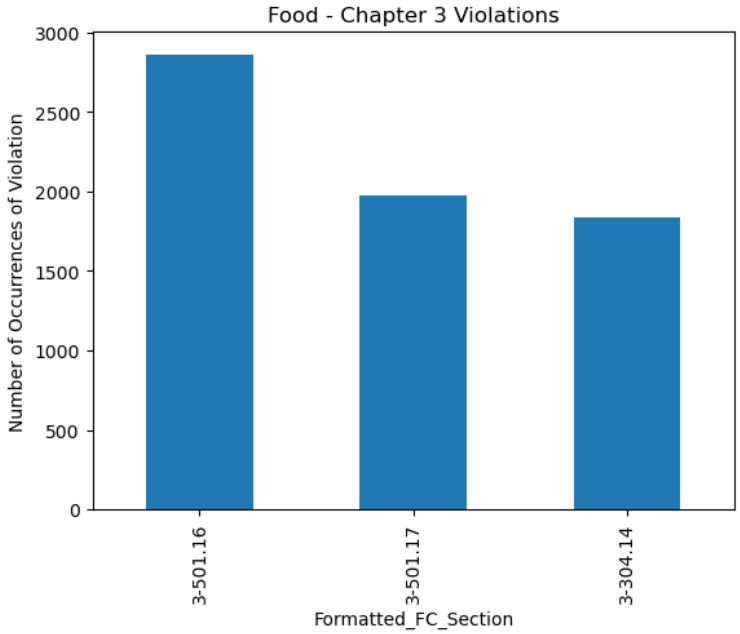
A plot bar plot of the establishment categories in df1 was made. It showed that majority of the establishments in the data frame were ‘Food Service’ type and then the rest were ‘Mobile Food’ and ‘Institutions’. This makes sense as df1 was limited to only violation sections dealing with the food code.

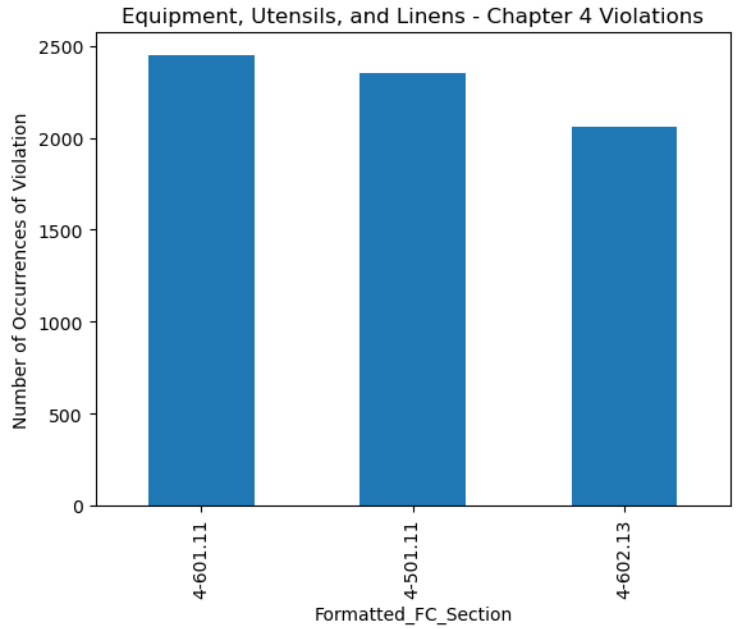


A bar plot of the food code chapters associated with the violations showed that most violations were from the ‘Food Chapter’. The second most chapter cited was the ‘Equipment, Utensils, and Linens’ chapter.



The top 3 food code sections cited for each chapter were plotted. For the food category, the most cited section was ‘3-501.16’ for **'Time /Temperature Control for Safety Food, Hot and Cold Holding'**. For the Equipment, Utensils, and Linen category, the most cited section was 4-601.11 for **'Equipment, Food-Contact Surfaces, NonfoodContact Surfaces, and Utensils.'**





A bar plot of the sections cited for critical violations was also plotted. It showed majority of the critical violations (15 %) came from '3-501.16' for **'Time /Temperature Control for Safety Food, Hot and Cold Holding'.**

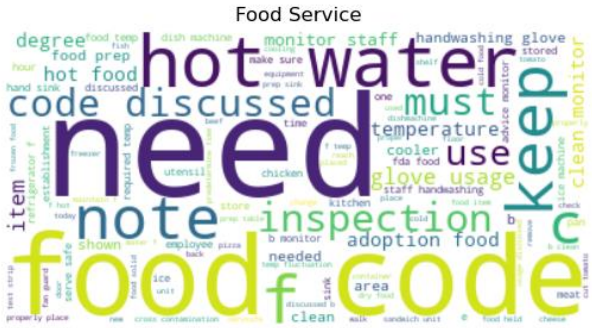
Pre-processing

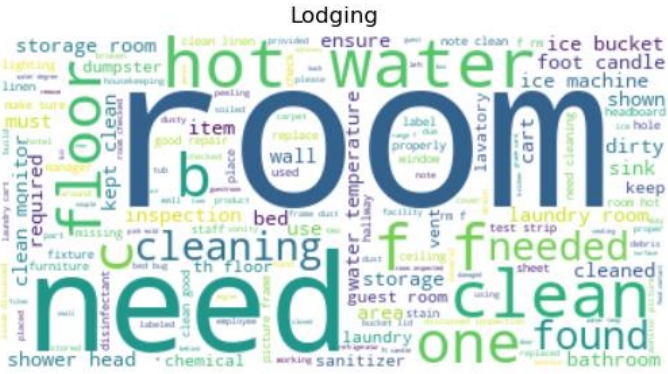
As mentioned above, the data sets, **Food-inspection-violations_1** and **Restaurant-and-services_3** could not be matched based on 'id'. A new data set, **Food-health-inspections_3.csv** was used instead for modelling in the project. Food-health-inspections_3.csv contained both inspection comments and establishment information. Note that establishments here were not limited to food establishments; the data set

included establishments such as daycares, swimming pools, tattoo parlors, etc. The intention was to use inspection comments and establishment features to predict and inspection score or grade.

There were 87 features in this dataset. Null values were dropped from inspection grades and scores then percentage of null values in other features were checked. Features with null values greater than 5% were dropped. Features that were redundant or did not add pertinent information were also dropped.

Word clouds were generated from all inspection comments for each establishment to investigate whether there were unique violations seen for each establishment. These word clouds were generated by removing stop words, tokenizing words, and then lemmatizing words in each establishment group. There was a lot of overlap between the word clouds which shows additional processing was needed (see example below).





It would be useful to use Part – of – Speech (POS) tagging to further refine the words used to build up the word cloud. Another approach would be to explore if flag embedding or sentence similarity can be used to better improve use of context for generating the wordclouds.

TfidfVectorizer was used to create vectors from the inspection comments for each row of data. The feature matrix including the output from Tfidfvectorizer was as scaled using StandardScaler and split into training and test sets. The inspection scores were used as the target variable.

Modeling

Three models were evaluated for prediction of the inspection grades. They were Support Vector Regression (SVR), RandomForestRegression (RFR) and XGBoost Regression (XGBR). The hyper parameters space was explored by using RandomizedSearchCV and Bayes Optimization; both methods of hyper parameter tuning were run in conjunction with a 5-fold cross validation. The Hyper parameters explored are below.

Support Vector Regression Hyperparameters

Hyperparameter	Values
Kernel	'Poly', 'rbf'
Epsilon	0.001, 0.1, 1
C	1, 50, 100

Random Forest Regression Hyperparameters

Hyperparameter	Values
n_estimators	50, 100, 150
max_features	'sqrt', 'log2'
max_depth	3, 6, 9
max_leaf_nodes	3, 6, 9

XGBoost Regression Hyperparameters

Hyperparameter	Values
max_depth	3, 6, 9
min_child_weight	1, 5, 10
subsample	0.5, 0.7, 0.9
learning_rate	0.05, 0.5, 1

The performance of the models was evaluated by measuring the R-squared values. The model parameters selected by RandomizedSearchCV and Bayes Optimization, associated r-squared values and model run times are shown in the table below.

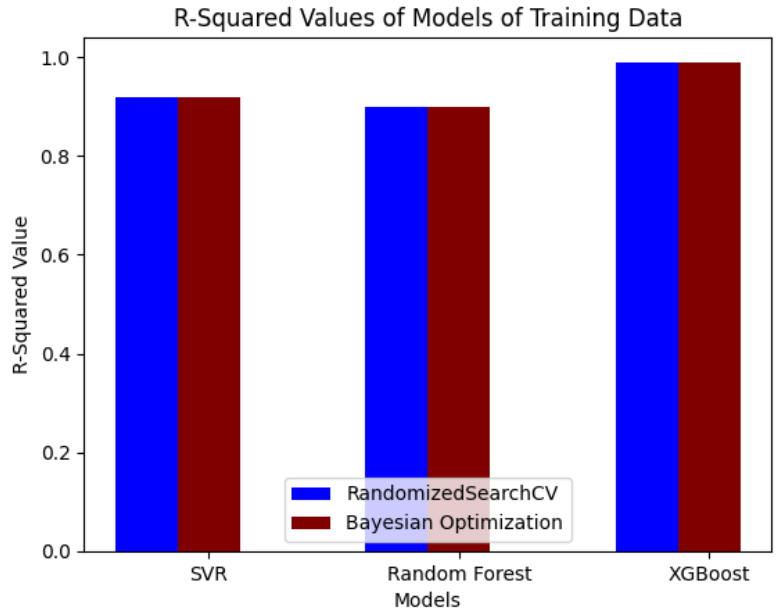
Hyper tuning Output – RandomizedSearchCV

Model	Best R-Squared Value	Best hyperparameters	Optimization Time (min)
SVR	0.92	'kernel': 'rbf', 'epsilon': 1, 'C': 100	10.16
RFR	0.90	'n_estimators': 50, 'max_leaf_nodes': 9, 'max_features': 'sqrt', 'max_depth': 9	0.31
XGBR	0.99	'subsample': 0.5, 'min_child_weight': 1, 'max_depth': 3, 'learning_rate': 0.5	2.46

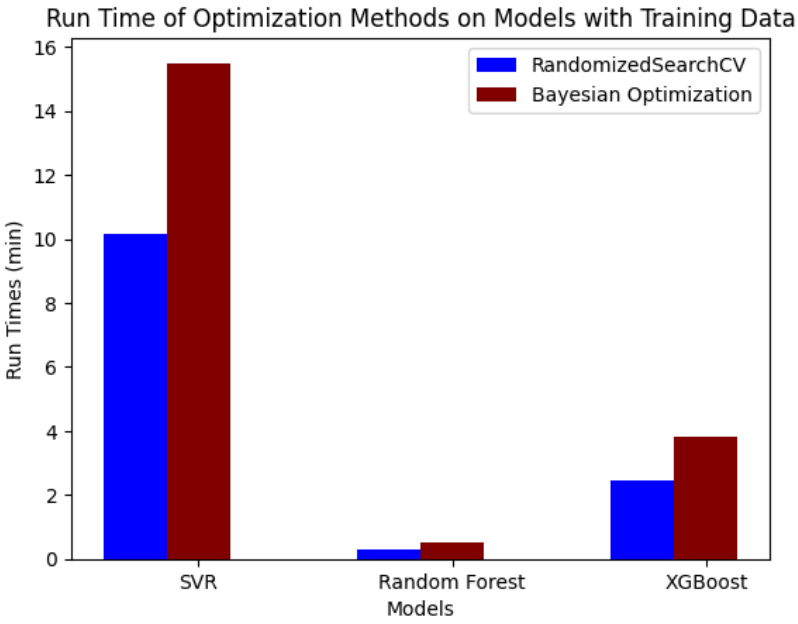
Hyper tuning Output – Bayes Optimization

Model	Best R-Squared Value	Best hyperparameters	Optimization Time (min)
SVR	0.92	'C': 69.7, 'epsilon': 0.01	15.5
RFR	0.90	'n_estimators': 53, 'max_leaf_nodes': 9, 'max_depth': 9	0.5
XGBR	0.99	'subsample': 0.77, 'min_child_weight': 1.6, 'max_depth': 8.0, 'learning_rate': 0.05	3.8

The R-squared value for each model was the same using the RandomizedSearchCV or Bayes Optimization method despite slightly different hyperparameters. This could be because the function being optimized does not vary a lot in the ranges of hyperparameters that were chosen.

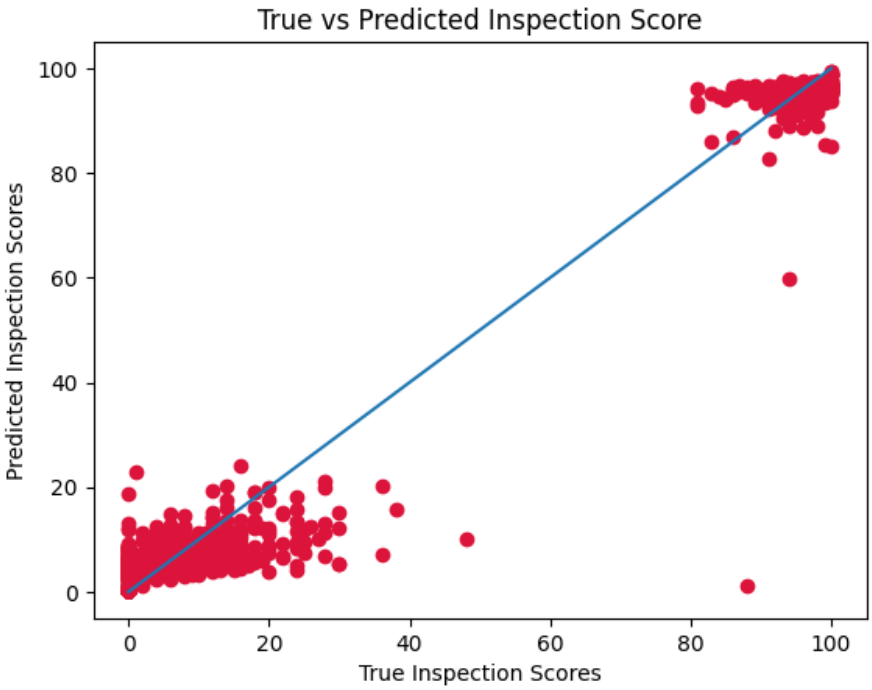


Overall, Bayes Optimization had a longer run time than RandomizedSearchCV. Many articles mention that Bayes Optimization is a more efficient method of hyperparameter tuning than RandomizedSearchCV; this is not true in this example but might be true when investigating a larger hyperparameter space.

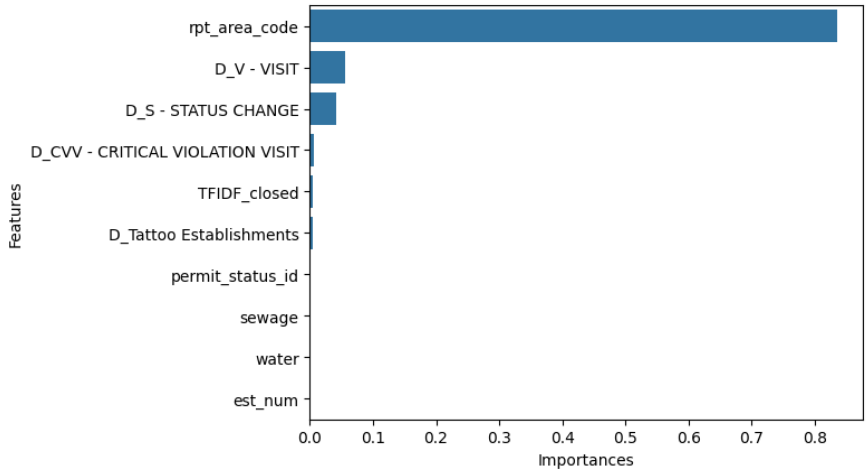


Exploration of non-numerical hyperparamters was not implemented in the Bayes Optimization method of hyper parameter tuning; a future exercise should look at including categorical hyperparameters for Bayes Optimization and see if that results in hyperparameter values that are closer to those selected for RandomizedSearchCV.

XGBoost Regression resulted in the best model score. When the XGBoost parameters were used to predict the test data set, the R-squared value was 0.99. The plot below shows the predicted scores plotted against the true scores. A gap in inspection scores exists between scores of 30 – 80. In this range, the error between true and residual values is greatest. This might be due to the lack of inspection scores in this region for learning. It would be useful to find out why inspection scores are typically extreme values, and repeat the modelling exercise on a dataset with a normal distribution of scores. This abnormal distribution of scores could have been caught earlier if a histogram of scores had been plotted. It is recommended that for future analysis, EDA should include a histogram of target features which are numerical.



Finally, the top 10 features of importance for the model were plotted. The area code of the establishment had the highest score. This code does not correspond to a phone number. It would be key to understand what this corresponds to in inspection reports by contacting the Durham County Department of Public Health. Other features of importance were the reasons for the inspection, whether it was a regular visit, a status change, or a critical violation visit.



Other interesting features of importance are the sewage type and water type used at the establishment.

Conclusion

The first inspection comments dataset showed that **'Time /Temperature Control for Safety Food, Hot and Cold Holding'** and **'Equipment, Food-Contact Surfaces, NonfoodContact Surfaces, and Utensils.'** are the most cited areas for food services establishments and that 15 % of critical violations are also attributed to **'Time /Temperature Control for Safety Food, Hot and Cold Holding'**.

Bayes Optimization and RandomizedSearchCV were effective for hyperparameter tuning and returned the same model scores for each model but RandomizedSearchCV had a shorter run time than Bayes Optimization. XGBoost Regression resulted in the highest model score with an R-squared value of 0.99 on the training set. It also had a higher score (R-square = 0.99) when used on the test set. However, the model performs well on inspection scores at the extreme ends of the score range. The score data is lacking within the score range of 30 – 80 which is likely why the model shows bad fit between true and predicted scores in this region.

The rtp(Research Triangle Park)_area_ code scored the highest importance as a feature of the model. Some of the reasons for inspection visits also scored high for model importance. XGBoost Regression performed well in predicting inspection scores based on inspection comments and specific information about the establishment visited.

Future

Recommendations for Clients

- Food service establishments, and companies who provide food safety and hygiene services for food service establishments should focus on training and simple

procedures related to time and temperature control for food safety and maintenance of equipment, food contact surfaces and nonfood contact surfaces.

- Companies who offer services to food establishments should explore opportunities for automating time and temperature control of foods and maintenance of equipment, food contact surfaces, and nonfood contact surfaces through hardware and software solutions.
- Further investment should be made in data analytics and modelling to understand:
    - Specific infractions under the 'time and temperature control' and 'equipment, food contact surfaces and nonfood contact surface' categories. It would be good to understand if there are systemic factors that drive infractions in certain customers, markets or geographical locations or if the infractions are one-offs
    - what drives certain inspection scores in a geographical location. It was surprising to see that the rpt_area_code had a much higher importance score compared to other features. An example of a question to ask to determine what drives inspection scores is "what does the rpt_area_code truly represent? a specific inspector, a socio-economic group?" and so on.

Recommendations for Future Analysis

- Before moving forward other steps in EDA or modelling, understand the distribution of values for the target variable.
- Reach out to the Durham County Health Department to understand
    - The extreme values in inspection scores
    - The meaning behind the rpt_area_code
- Work to understand if the extreme values in inspection scores are unique to this data set or if this can be seen in data sets in other counties and states. When I have visited restaurants in various cities in the US I have seen that they display high inspection score ratings (> 90 or A grades), so it would b good to understand if this is a norm across states.
- Explore using Principal Component Analysis (PCA) to reduce the dimensionality of the predictive features instead of relying on contextual knowledge to drop features.
- Build on text manipulation methods used for the inspection comments by using sentence similarity to understand 'closeness' between comments from establishments, then see if there is an underlying pattern (e.g. method of operation, location, etc.) in establishments which are clustered together.
- Investigate the use of vectors from flag embedding instead of TFIDF Vectorizer as features for modelling.
- Use output from PCA, sentence similarity of flag embedding to perform a clustering analysis on establishments to see if there is an underlying factor among establishments that makes them cluster.

- Incorporate categorical hyperparameters in the hyperparameter tuning step for XGBoost Regression and evaluate whether that changes the recommended best parameters.
- Apply the XGBoost Regression model to another data set with similar features (e.g. if another county in NC records inspection data in a similar format, how successful will the model be in predicting the inspection scores for that county?)
- Group the observations from **Food-inspection-violations_1** and **Restaurant-and-services_3** on the type of establishment and check if the two data frames can be joined on establishment type. Determine if a working model could be built on this need aggregated data frame.