Predicting Durham County Department of Public Health Inspection Scores

Temilola Famakinwa

Email: tefamakinwa@gmail.com

LinkedIn: https://www.linkedin.com/in/famakinwatemilola/

The Problem

- Incorrect implementation of hygiene procedures at food services establishments leads to:
  - Brand risk for the establishments.[1]
  - Risk to public health and safety for consumers.[2]

How can we help establishments understand where and how to improve in their hygiene procedures?

1. https://www.fastcompany.com/91137429/addressing-brand-damage-and-legal-risk-in-food-safety-issues
2. https://www.cdc.gov/foodborne-outbreaks/index.html

Approach

- Determine the **most frequent health department violations** in Durham county.
- Predict health **department inspection scores**
- Find **opportunities for improvement** for the 5 worst performing food service establishments.

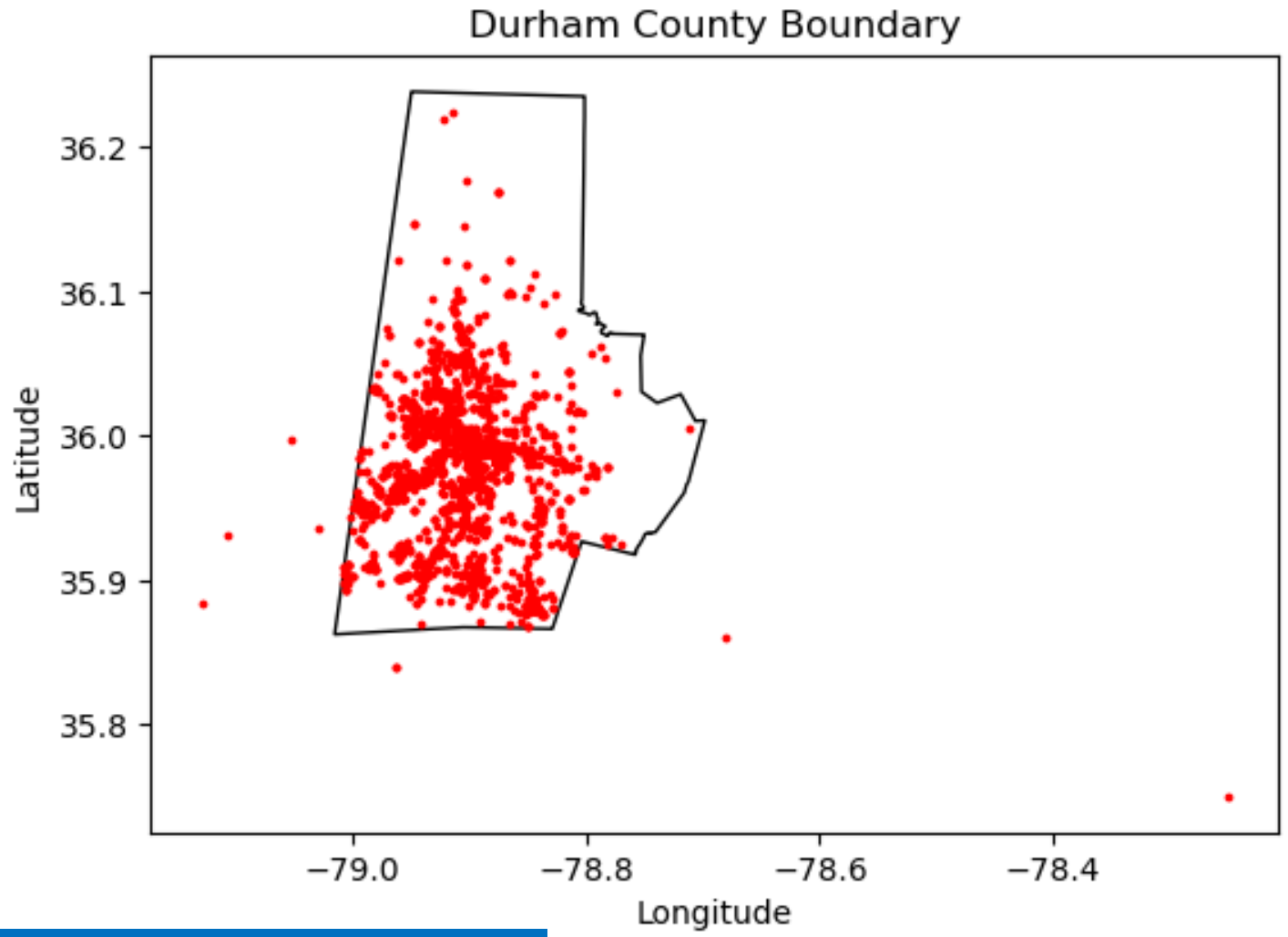
Data Sources and Their Uses

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
Food-health-inspections_3.csv	Kaggle	Restaurant and violation information
Durham_County_Boundary.shp	Durham, NC Open Data Portal	Durham County GIS Boundary

- **Food-inspection-violations_1:** To learn about the types of inspection violations.
- **Restaurants-and-services_3** and **Durham_County_Bounday:** To map out the location of restaurants in Durham county
- **Food-health-inspections_3:** To model inspection scores.

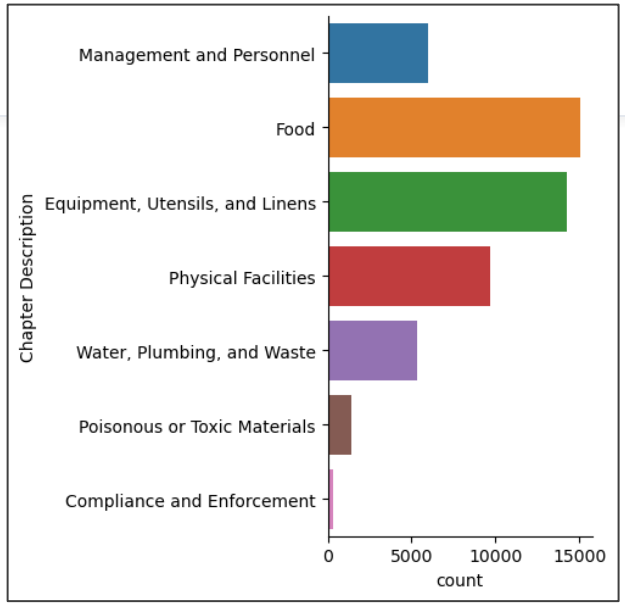
EDA Findings

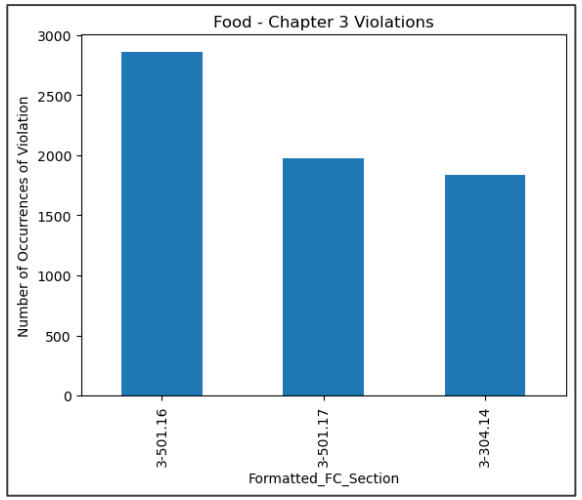
Restaurant Locations

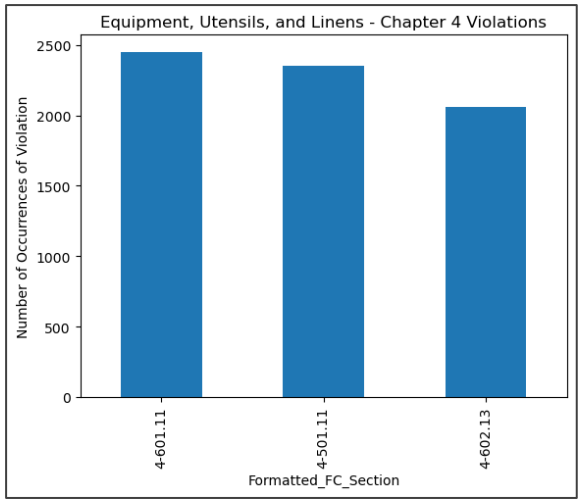


- Establishments are concentrated toward the center and lower parts of Durham county (close to Durham city).
- Additional questions: is there is a difference in number of restaurant visits per year between restaurants on the outskirts of the county, and those in the center?

General Food Code Violations







Violations related to **'Time/Temperature Control for Safety Food, Hot and Cold Holding'** and **'Equipment, Food-Contact Surfaces, NonfoodContact Surfaces, and Utensils.'** occurred most frequently in each chapter.

Modeling

Modeling Plan

Feature Engineering

Features > 5% null values were dropped

Non-pertinent feature dropped

Comments (stop words, tokenization, lemmatization, TFIDFVectorizer)

StandardScaler

Supervised Learning: Regression*

Predict Inspection Scores

Support Vector Regression (SVR)

RandomForest Regression (RFR)

XGBoost Regression (XGBR)

Model Parameter Tuning

RandomizedSearch CV

Bayes Optimization

5-Fold Cross Validation

*See appendix for model hyperparameters

Model Performance

Hyper tuning Output – RandomizedSearchCV

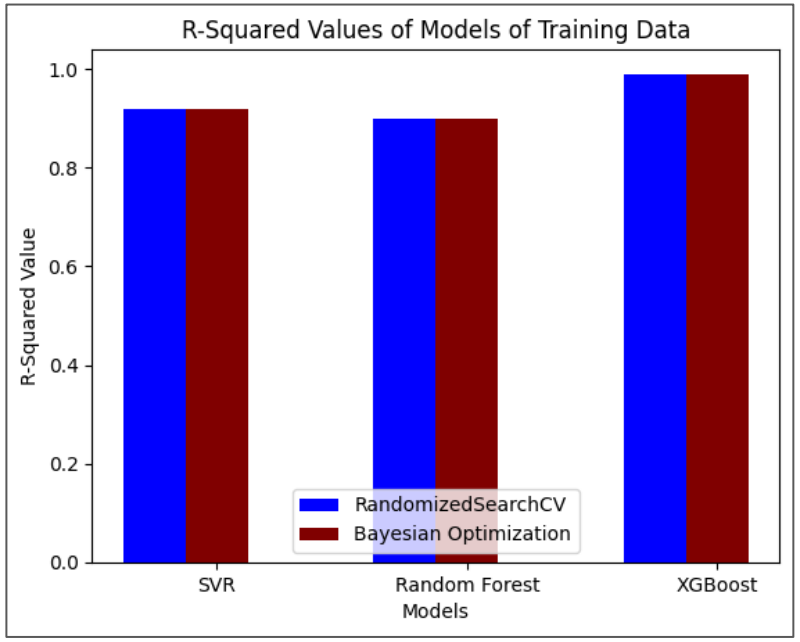
Model	Best R-Squared Value	Best hyperparameters	Optimization Time (min)
SVR	0.92	'kernel': 'rbf'	10.16
		'epsilon': 1	
		'C': 100	
RFR	0.90	'n_estimators': 50	0.31
		'max_leaf_nodes': 9	
		'max_features': 'sqrt'	
		'max_depth': 9	
XGBR	0.99	'subsample': 0.5	2.46
		'min_child_weight': 1	
		'max_depth':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

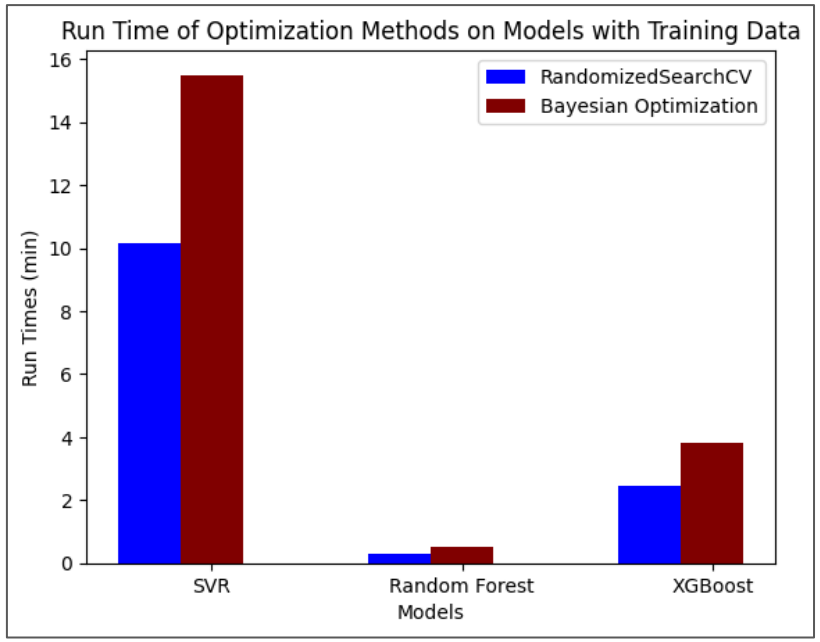
Hyper tuning Output – Bayes Optimization

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

⟫ Categorical and numerical hyperparameters were used in RandomizedSearchCV while only numerical parameters were used in Bayes Optimization.

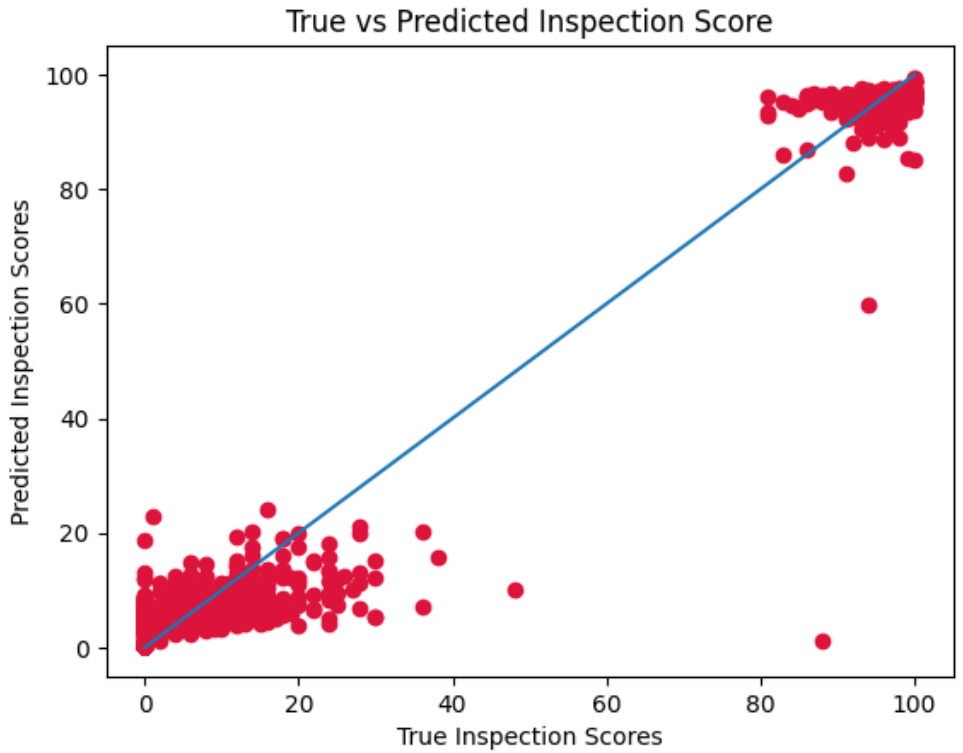
Model Scores and Run Time





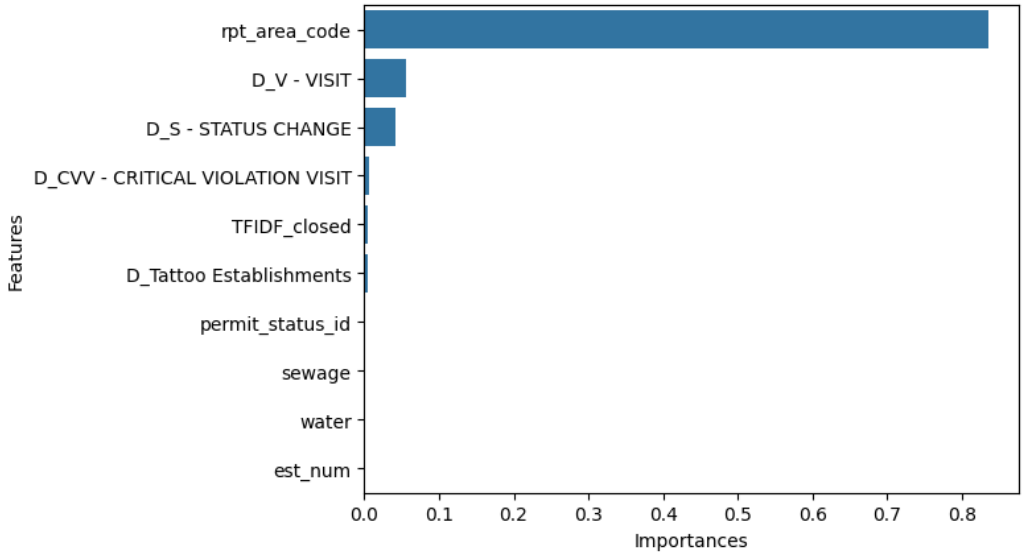
- XGBoost Regression resulted in the best model score. R-squared = 0.99).
- R-squared value for each model was the same using the RandomizedSearchCV or Bayes Optimization.
- Bayes Optimization had a longer run time than RandomizedSearchCV

Model Performance on Test Data



- XGBoost Regression along with its hyperparameters were used to predict scores on test data. R-squared value was 0.99.
- Majority of recorded inspection scores are at the extremes of the score range.
- The error between true scores and test scores is greatest in score range between 20 – 80, possibly due to lack of data.

Model Features of Importance



- The feature with the highest importance score was rpt_area_code description. Additional work should be done to understand what it corresponds to
- The reason for inspection visit were also important features.

Business Recommendations

- Companies who provide food safety and hygiene services for food service establishments should **focus on training on 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:
    - if there are **systemic factors that drive infractions** in certain customers, markets or geographical locations or if the infractions are **one-offs** for infractions under the 'time and temperature control' and 'equipment, food contact surfaces and nonfood contact surface' categories.
    - what drives certain inspection scores in a geographical location. 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.

Future Modeling Recommendations

1. Before moving forward other steps in EDA or modelling, understand the distribution of values for the target variable.
2. Reach out to the Durham County Health Department to understand
   - The extreme values in inspection scores
   - The meaning behind the rpt_area_code
3. 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.
4. Explore using Principal Component Analysis (PCA) to reduce the dimensionality of the predictive features instead of relying on contextual knowledge to drop features.
5. 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.

Future Modeling Recommendations

6. Investigate the use of vectors from flag embedding instead of TFIDF Vectorizer as features for modelling.
7. 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.
8. Incorporate categorical hyperparameters in the hyperparameter tuning step for XGBoost Regression and evaluate whether that changes the recommended best parameters.
9. 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?)
10. 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.

Appendix

Model Hyperparameters

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