Analyzing Chicago Food Inspections Data to Predict Inspection Results

Capstone Project

Yelena Zadoyan

The Problem

- Health and Safety Issues
- Resource Allocation Issues
- Legal and Regulatory Compliance Issues

- Problem Solving Steps
  - data preprocessing,
  - EDA,
  - feature engineering,
  - model building,
  - model evaluation,
  - inspection result prediction.

The Dataset

- **Dataset** - food inspections conducted in Chicago.

- **Attributes** - the ID of the inspection, the name of the establishment, the type of establishment, the risk level, the address, the date of the inspection, the type of inspection, the results, and any violations found.

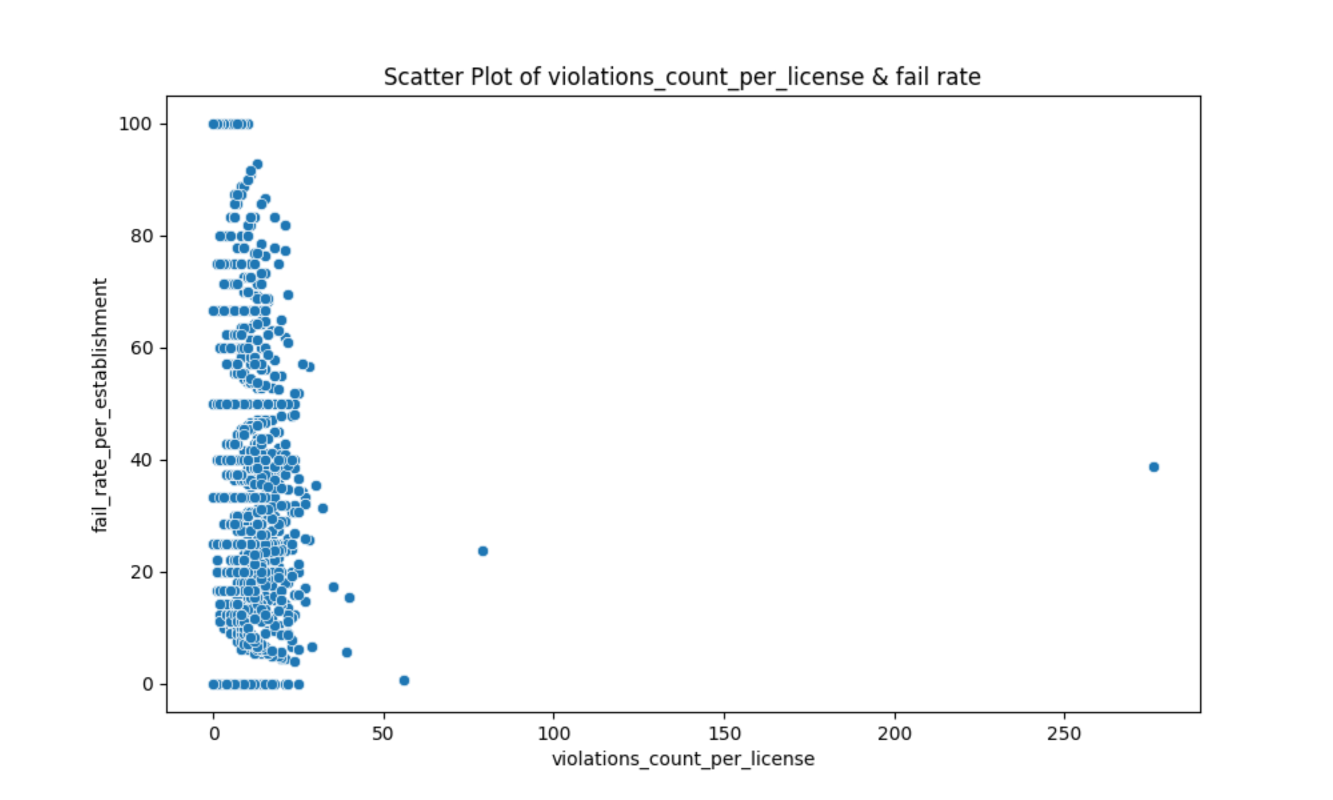
- **AIM** – predict the results of food inspections in the city of Chicago – the possible failure.

Data Preprocessing

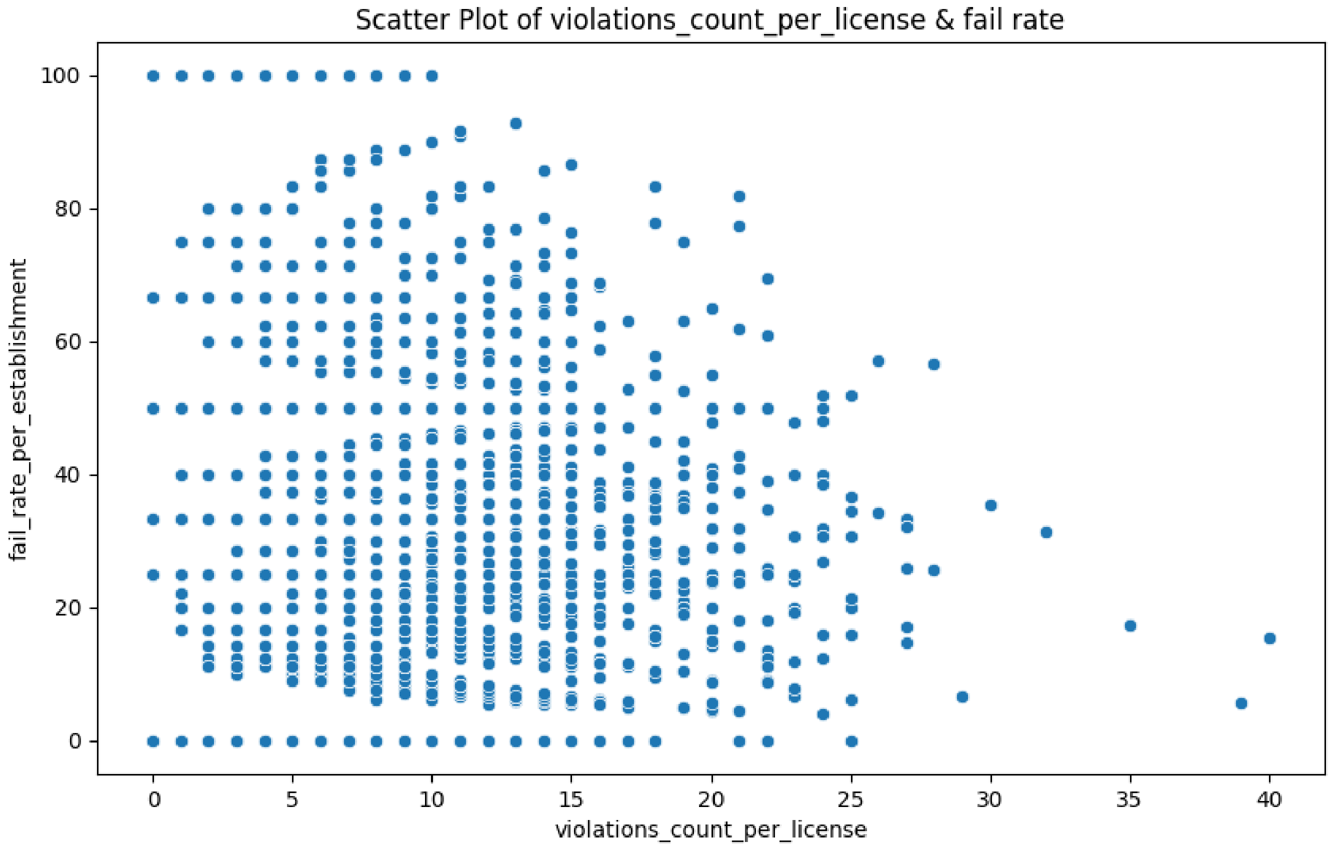
- As we deal with categorical data – the missing values can be replaced by the mode. But in our case, as the data refers to the safety standards, the rows containing missing values are dropped. The high weight of missing values has the Violations feature, for which the replacement by the model could impact significantly the results without increasing the accuracy.
- Only the inspections with Pass and Fail results are left by dropping the other rows.
- Those categorical variables that are considered in the scope of model building (have impact on failure rate and/or high weight in the dataset) are transformed into dummies.

Data Preprocessing

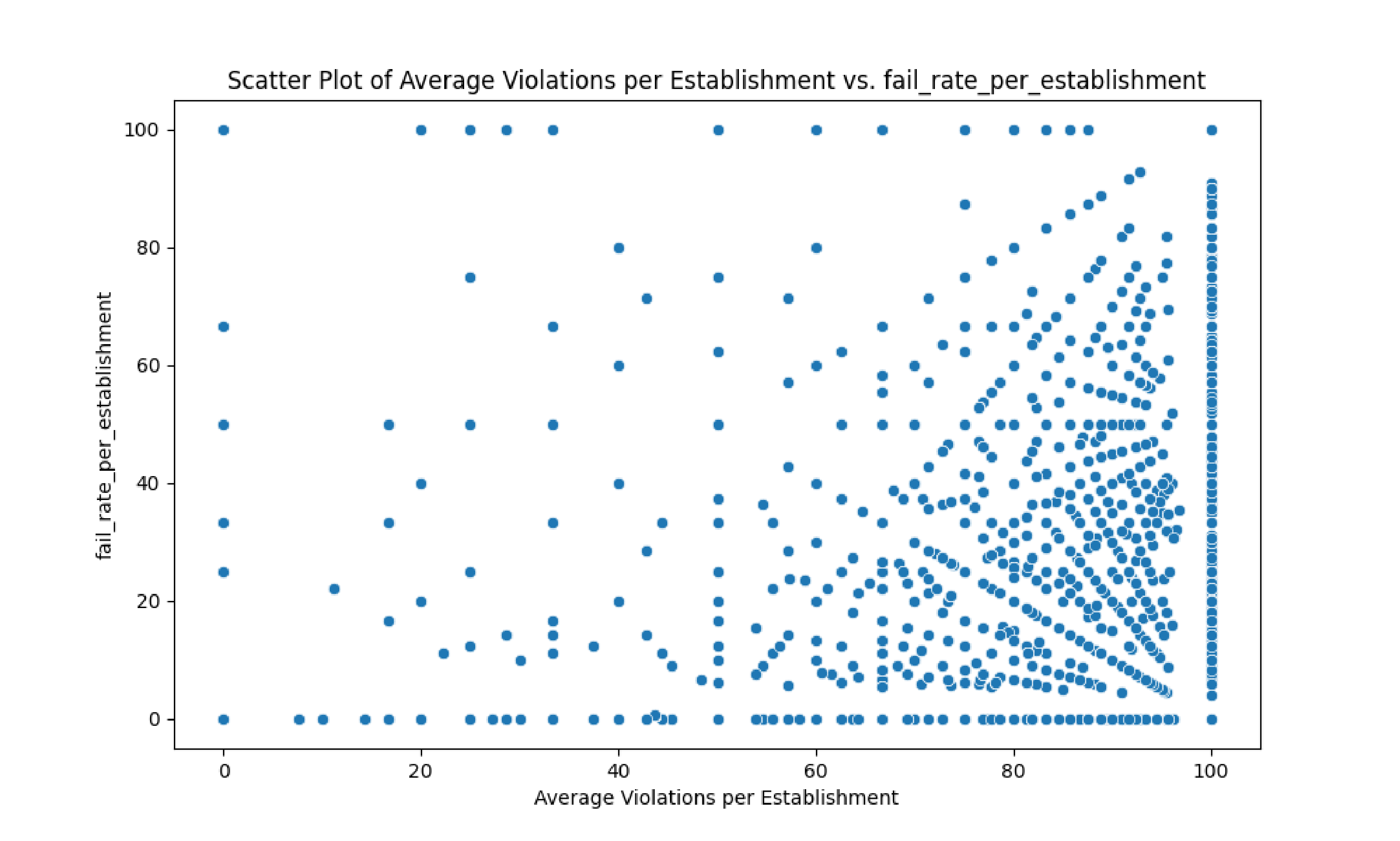
- Dropping outliers with high violation rate (>50)



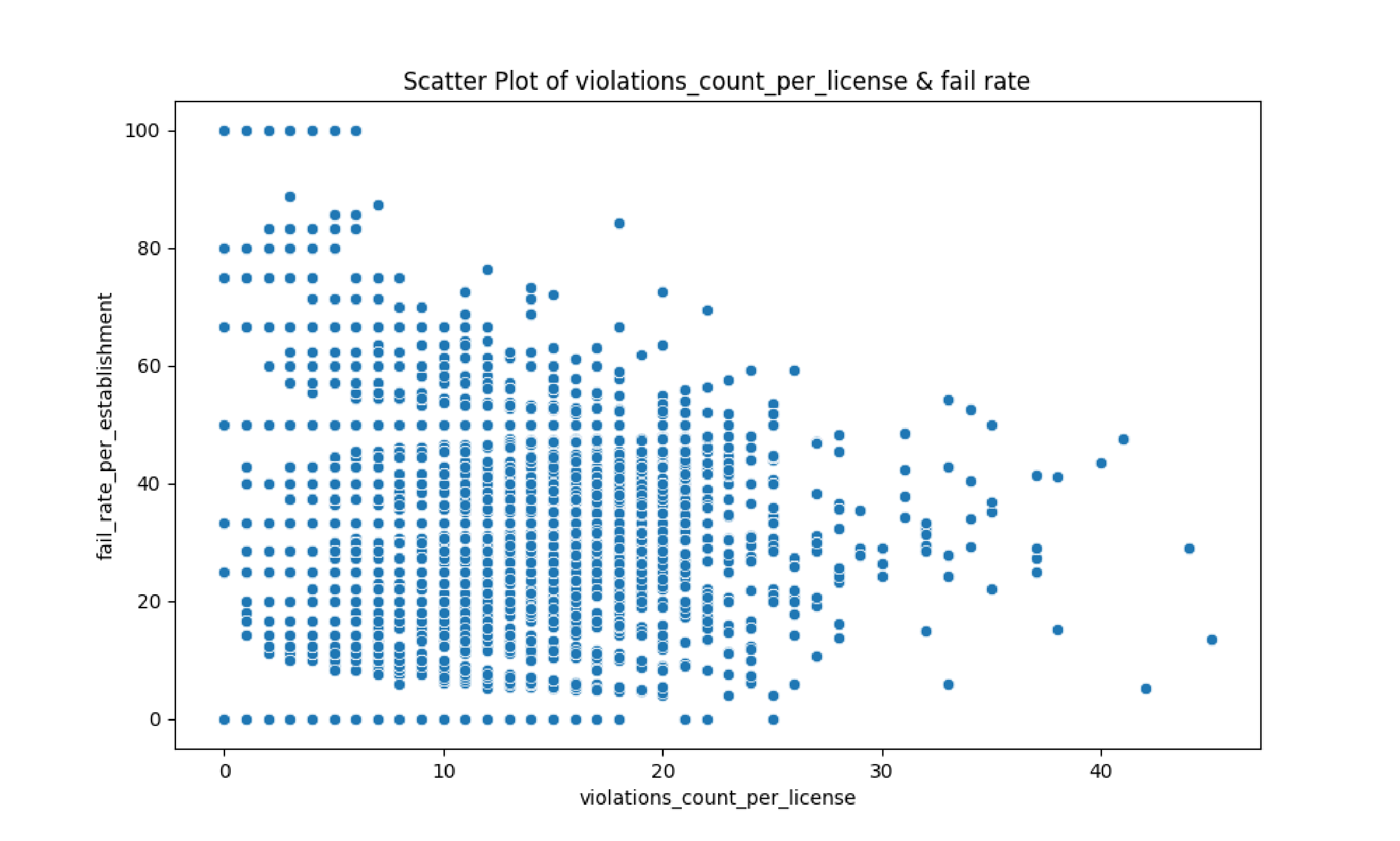
Data Preprocessing



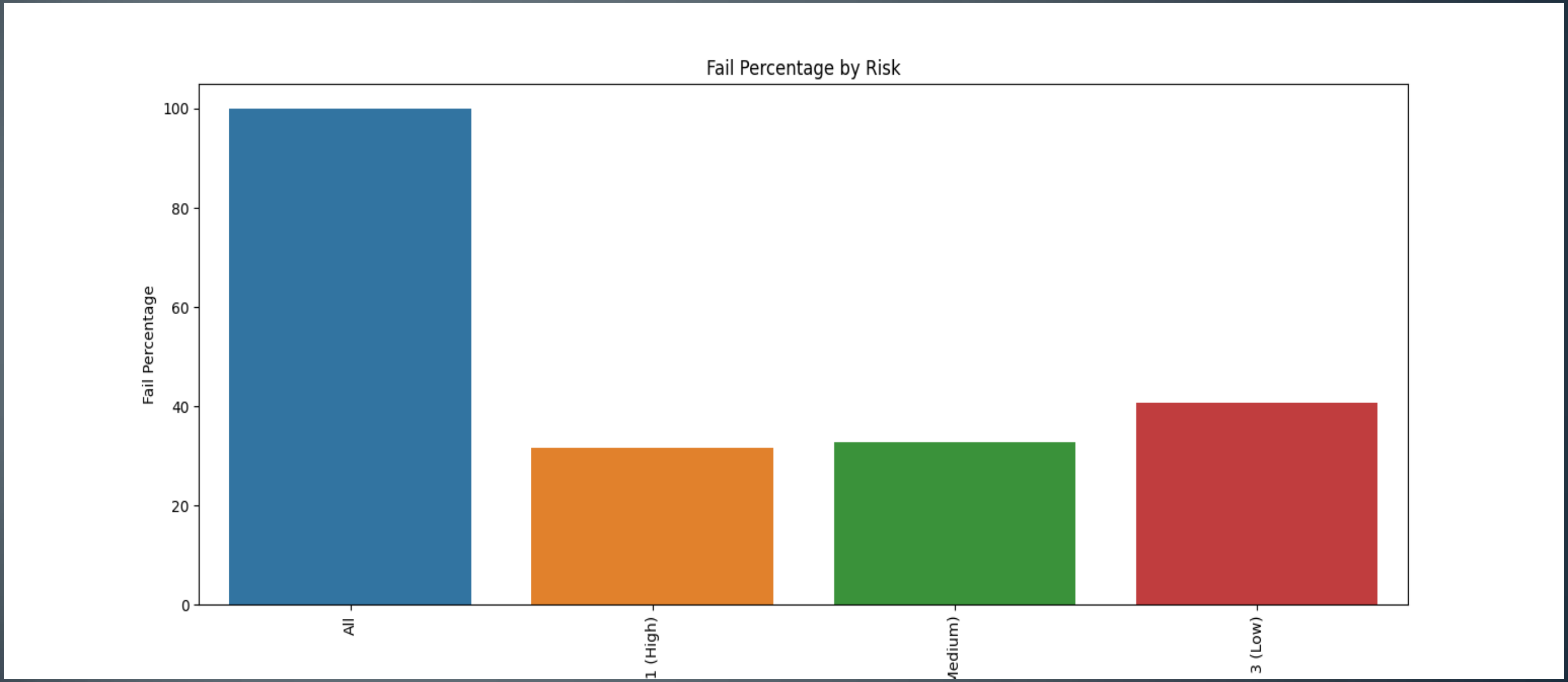
Feature engineering & EDA



Exploratory Data Analysis

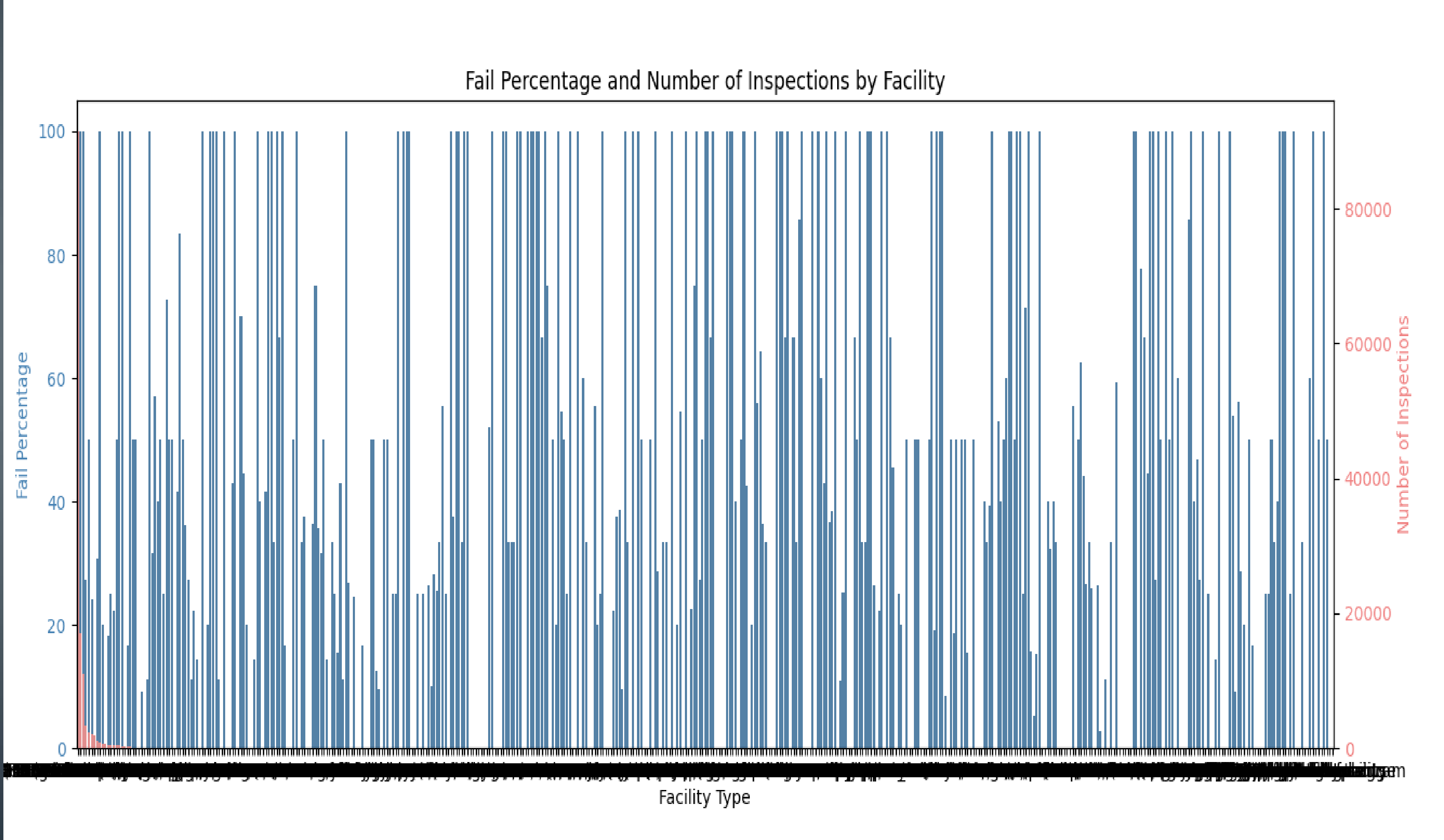


Exploratory Data Analysis



Risk feature – the high risk is overall in line with the high failure rate, but not much difference between the low and medium risk categories.

Exploratory Data Analysis



No essential relationship.

Exploratory Data Analysis

- From Facility type feature

only those types are transformed into dummies, which have >5% weight.

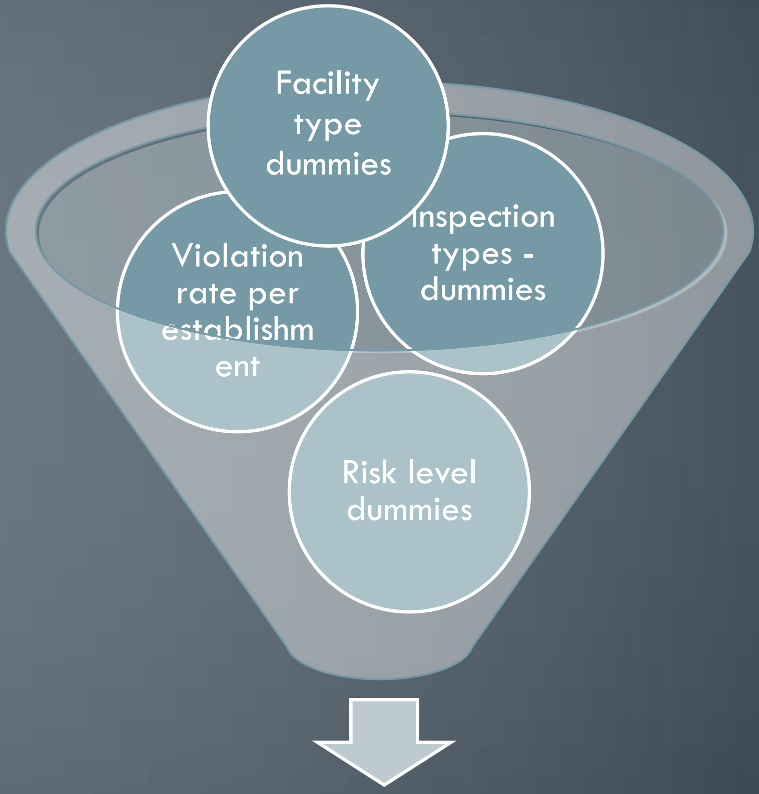
Facility Type	Weight
Restaurant	66.090488
Grocery Store	12.360622
School	7.989421
Children's Services Facility	2.409255
Daycare Above and Under 2 Years	1.594126
…	
PALETERIA /ICECREAM SHOP	0.000723
GROCERY/LIQUOR	0.000723
PRODUCE STAND	0.000723
RESTAURANT/GROCERY	0.000723
Kids Cafe'	0.000723

EDA

Fail rate analysis by inspection types

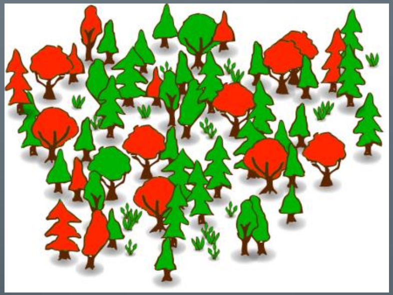
Inspection_Canvass	28.42%
Inspection_Suspect	**37.02%**
Inspection_Task	**57.99%**
Inspection_Consultation	20.57%
Inspection_Complaint	**37.41%**
Inspection_other inf	24.10%

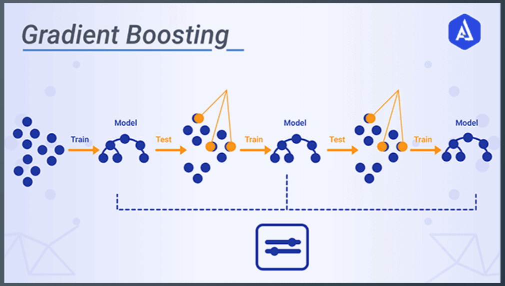
Data selection for model estimation



Fail 1/0

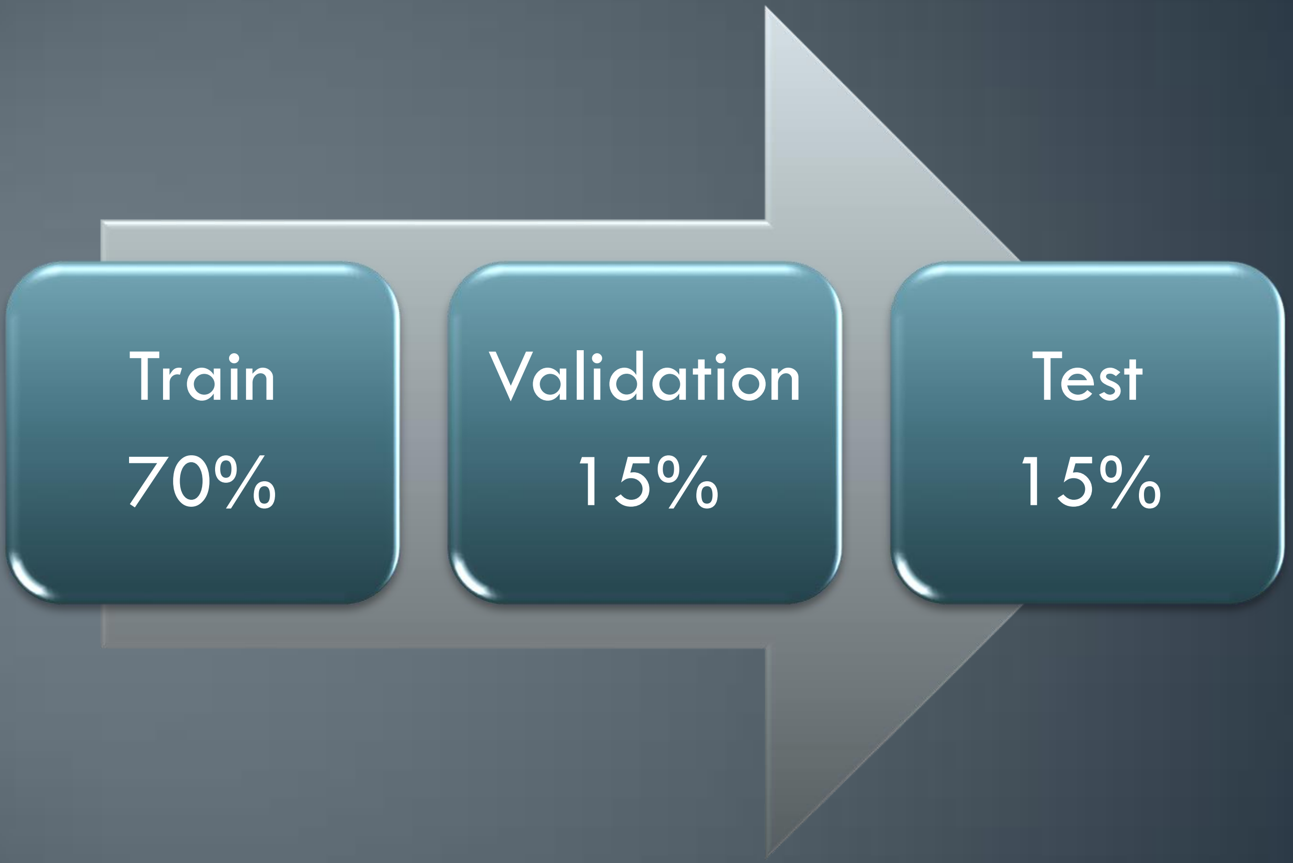
Model building





Four types of binary classification models considered.

Model Evaluation



Train

70%

Validation

15%

Test

15%

Model Evaluation

Model Group #1

- ALL Data: Inspection and re-inspection categories

Model Group #2

- Cutting the results from the re-inspection

Model Results – Group #1

Over-fit possibility

Logit Train

Accuracy 0.7268

Precision: 0.6398

Recall: 0.0142

F1: 0.0278

Logit Val

Accuracy: 0.7279

Precision: 0.6446

Recall: 0.0145

F1: 0.0283

DF Train

Accuracy 0.7288

Precision: 0.6641

Recall: 0.0276

F1: 0.0530

DF Val

Accuracy: 0.7293

Precision: 0.6250

Recall: 0.0284

F1: 0.0544

RF Train

Accuracy 0.7289

Precision: 0.7060

Recall: 0.0236

F1: 0.0458

RF Val

Accuracy: 0.7293

Precision: 0.6629

Recall: 0.0237

F1: 0.0457

GB Train

Accuracy 0.7268

Precision: 0.8114

Recall: 0.0082

F1: 0.0162

GB Val

Accuracy: 0.7281

Precision: 0.8171

Recall: 0.0091

F1: 0.0179

Model Results – Group #2

No over-fit possibility

Logit Train

Accuracy 0.6846

Precision: 0.6012

Recall: 0.0628

F1: 0.1137

Logit Val

Accuracy: 0.6736

Precision: 0.5931

Recall: 0.0599

F1: 0.1089

DF Train

Accuracy 0.6882

Precision: 0.5757

Recall: 0.1218

F1: 0.2011

DF Val

Accuracy: 0.6771

Precision: 0.5677

Recall: 0.1221

F1: 0.2010

RF Train

Accuracy 0.6885

Precision: 0.5809

Recall: 0.1191

F1: 0.1977

RF Val

Accuracy: 0.6774

Precision: 0.5726

Recall: 0.1182

F1: 0.1959

GB Train

Accuracy 0.6867

Precision: 0.5817

Recall: 0.0978

F1: 0.1675

GB Val

Accuracy: 0.6755

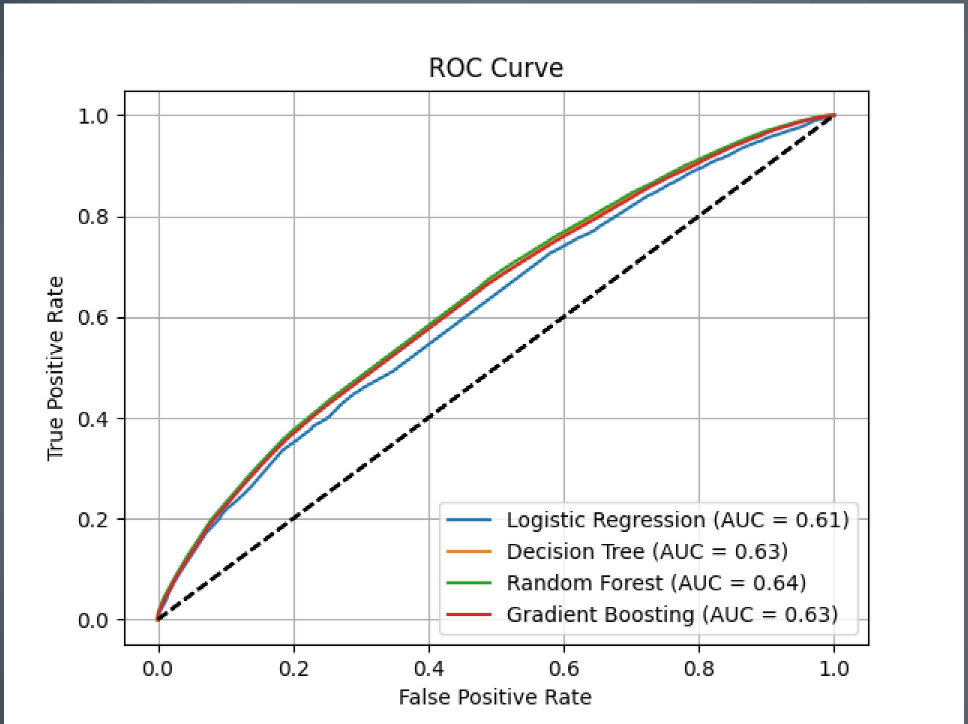
Precision: 0.5727

Recall: 0.0965

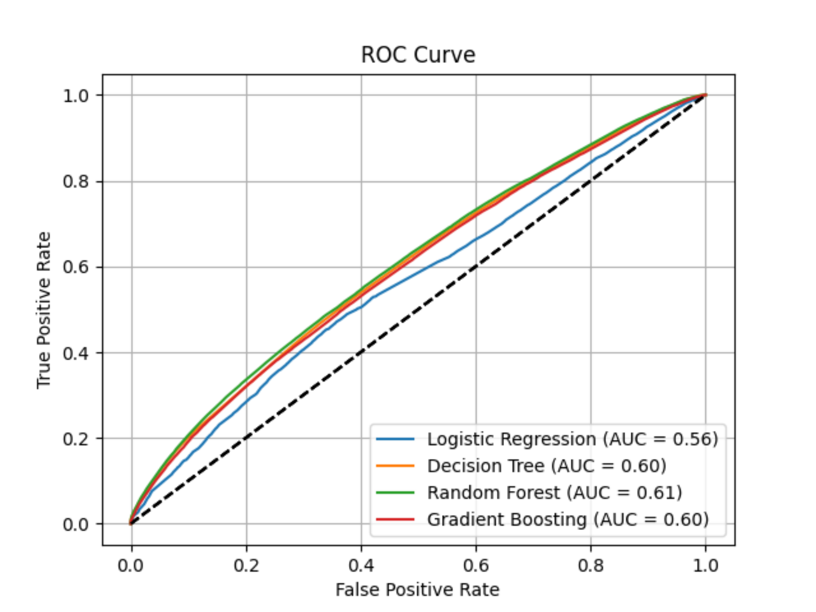
F1: 0.1651

Model Evaluation

Model Group #2



Model Group #1



Hyper-parameter optimization

Logistic Regression - Best Parameters:
{'C': 10, 'penalty': 'l2'}

Decision Tree - Best Parameters:
{'max_depth': 10, 'min_samples_split': 10}

Random Forest - Best Parameters:
{'max_depth': 10, 'n_estimators': 50}

Gradient Boosting - Best Parameters:
{'learning_rate': 0.1, 'n_estimators': 200}

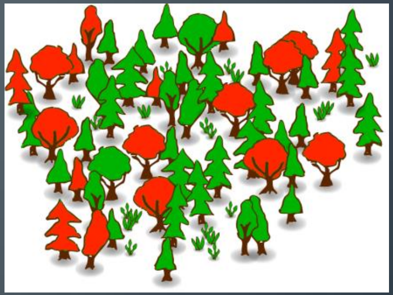
Model Selection

Models #1

- Higher Accuracy & Precision
- Lower Recall & F1, ROC-AUC
- Over-fit possibility

Models #2

- Lower Accuracy & Precision
- Higher Recall & F1, ROC-AUC
- No over-fit possibility



Model Result prediction

Logit test

Accuracy 0.6784

Precision: 0.5929

Recall: 0.0591

F1: 0.1074

DT test

Accuracy: 0.6789

Precision: 0.5466

Recall: 0.1173

F1: 0.1931

RF test

Accuracy 0.6797

Precision: 0.5550

Recall: 0.1149

F1: 0.1905

GB test

Accuracy: 0.6776

Precision: 0.5490

Recall: 0.0921

F1: 0.1578

Testing sample

Conclusion

- All the models have high accuracy and precision rates and very low recall and F1 score. Thus, the results can be used mostly in case of **Resource Allocation issues**, when limited resources are directed toward inspections that are more likely to identify actual failures, reducing unnecessary inspections on compliant establishments, and for **Legal and Regulatory compliance issues**, when precision might be prioritized to minimize the risk of wrongly penalizing compliant establishments. But the model **cannot be considered in case of Health and Safety concerns.**