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

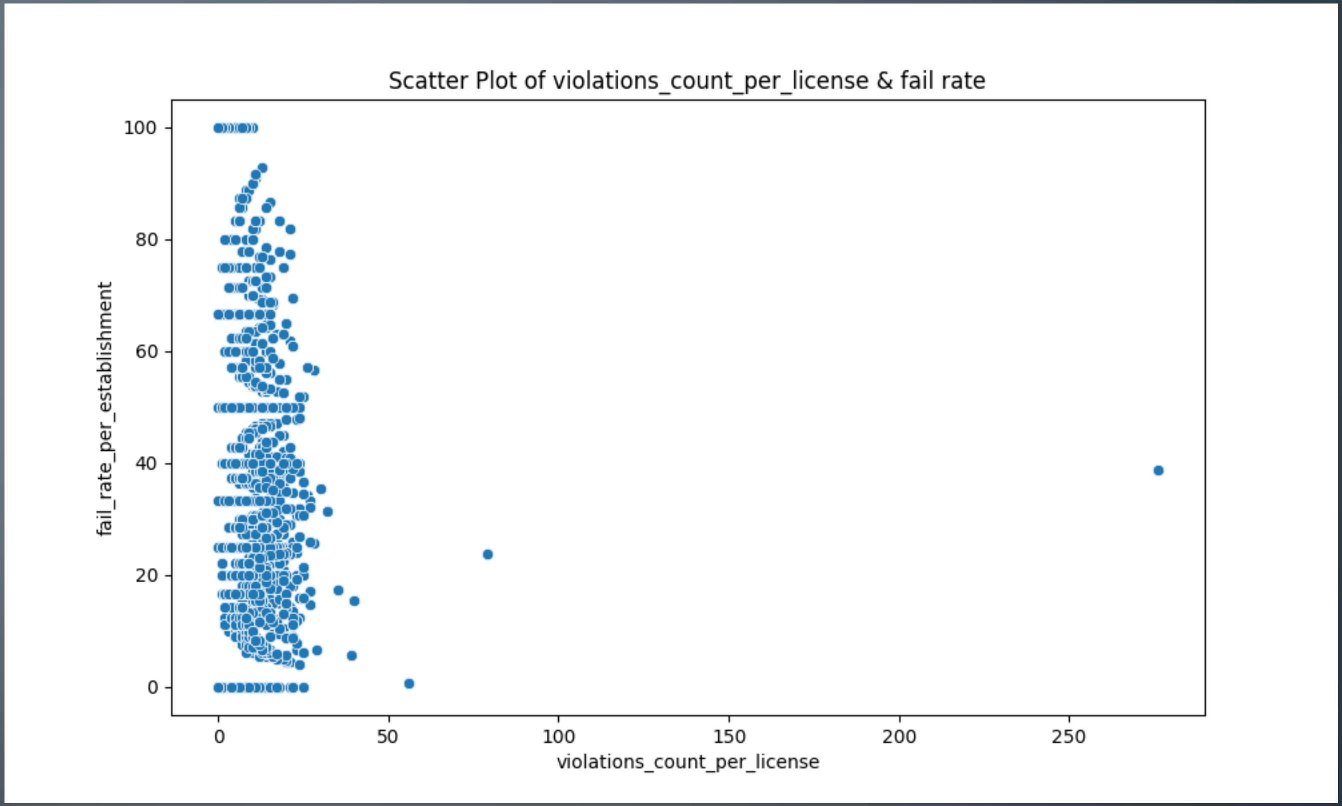
- **Dataset** - food inspections conducted in Chicago.
- **Attributes** - the ID of the inspection, the name of the establishment, the type of establishment, the risk level, the address, the date of the inspection, the type of inspection, the results, and any violations found.
- **AIM** – predict the results of food inspections in the city of Chicago – the possible failure.

Data Preprocessing

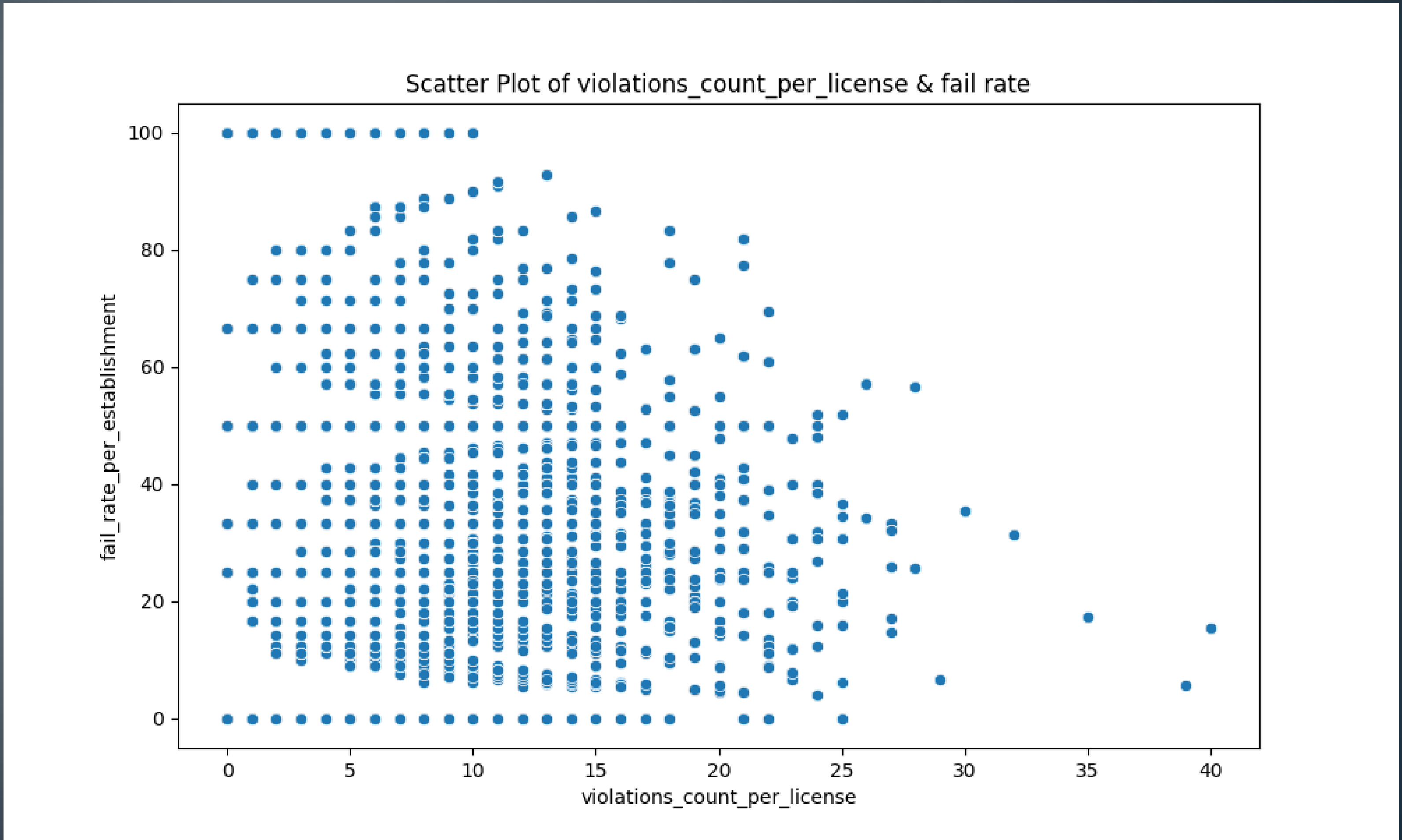
- As we deal with categorical data – the missing values can be replaced by the mode. But in our case, as the data refers to the safety standards, the rows containing missing values are dropped. The high weight of missing values has the Violations feature, for which the replacement by the model could 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

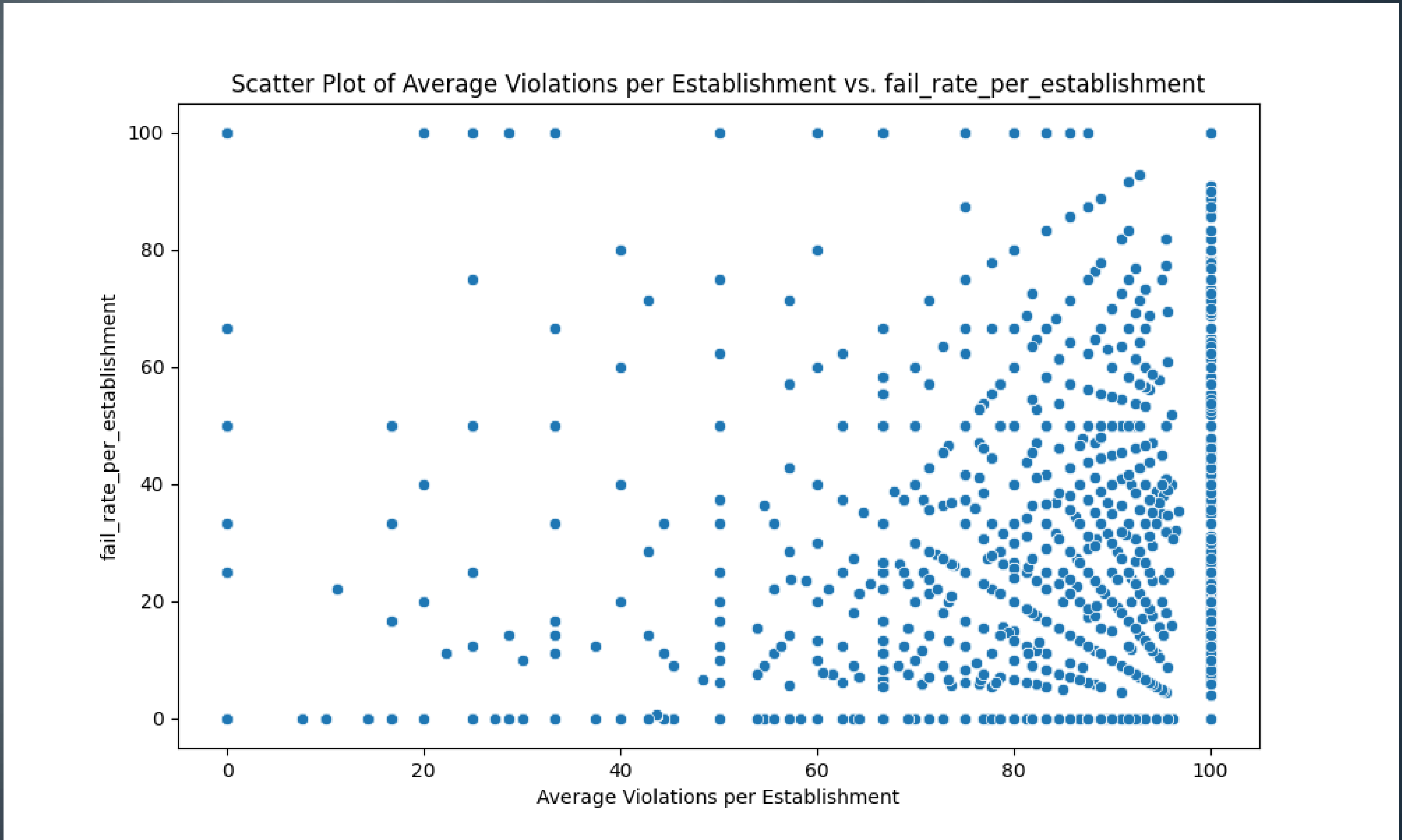
- Dropping outliers with high violation rate (>50)



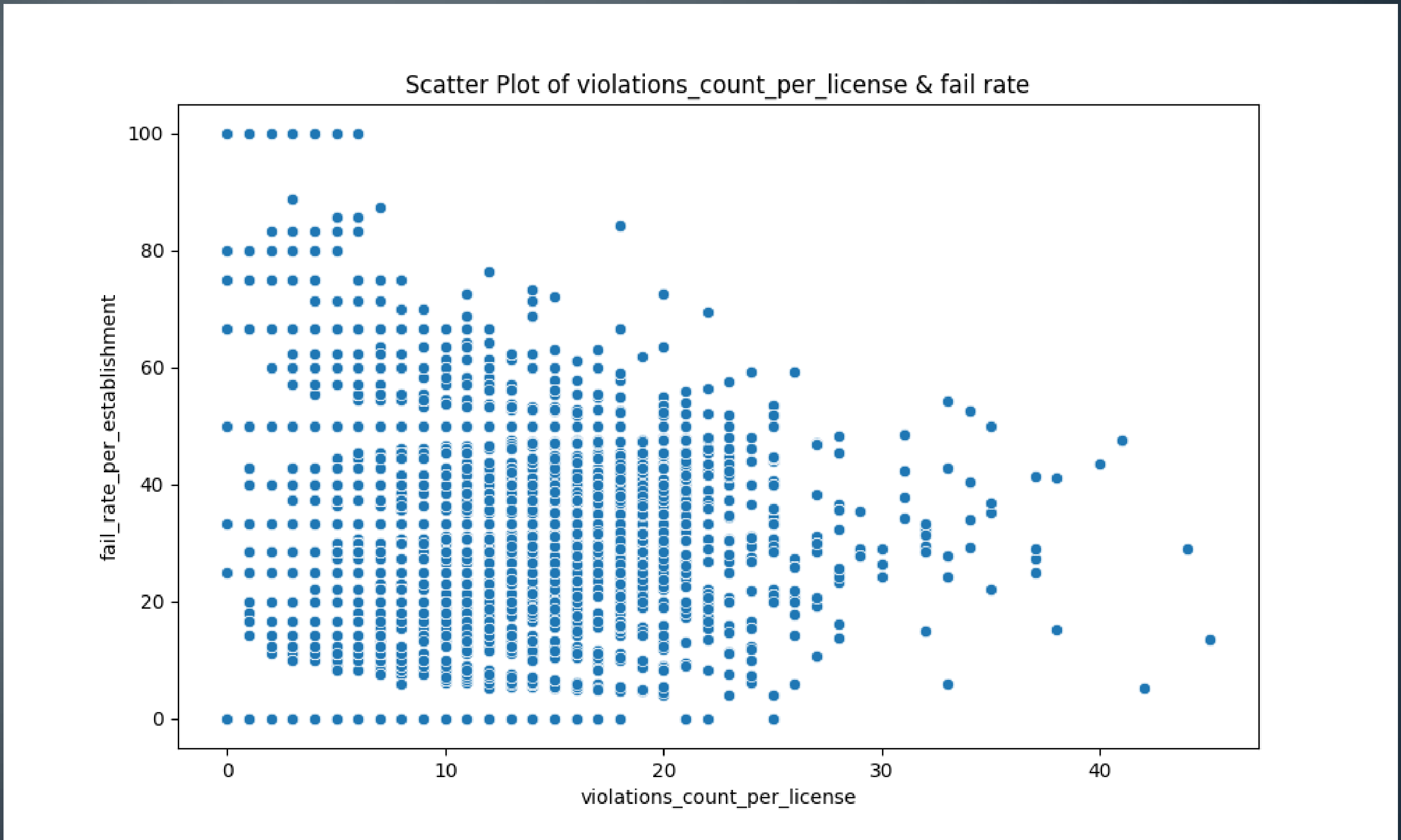
Data Preprocessing



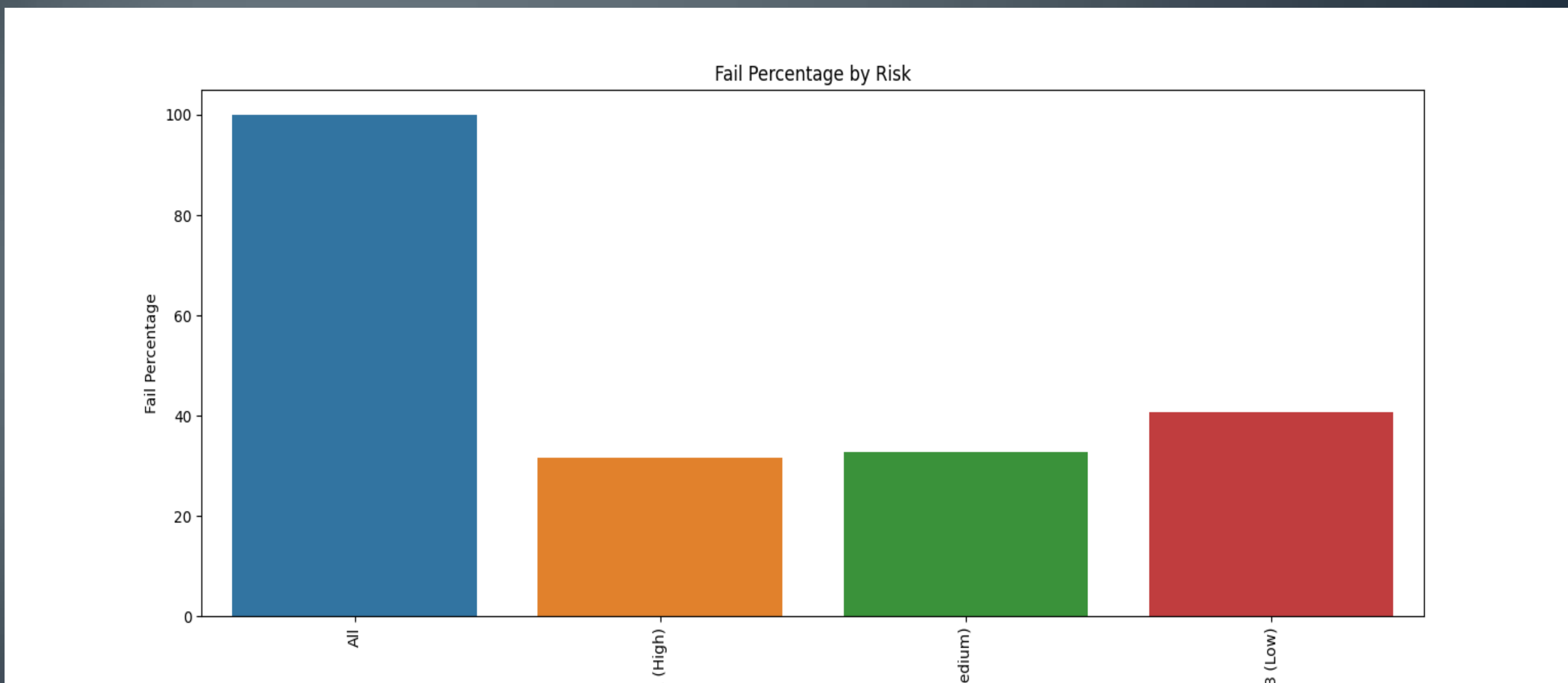
Feature engineering & EDA



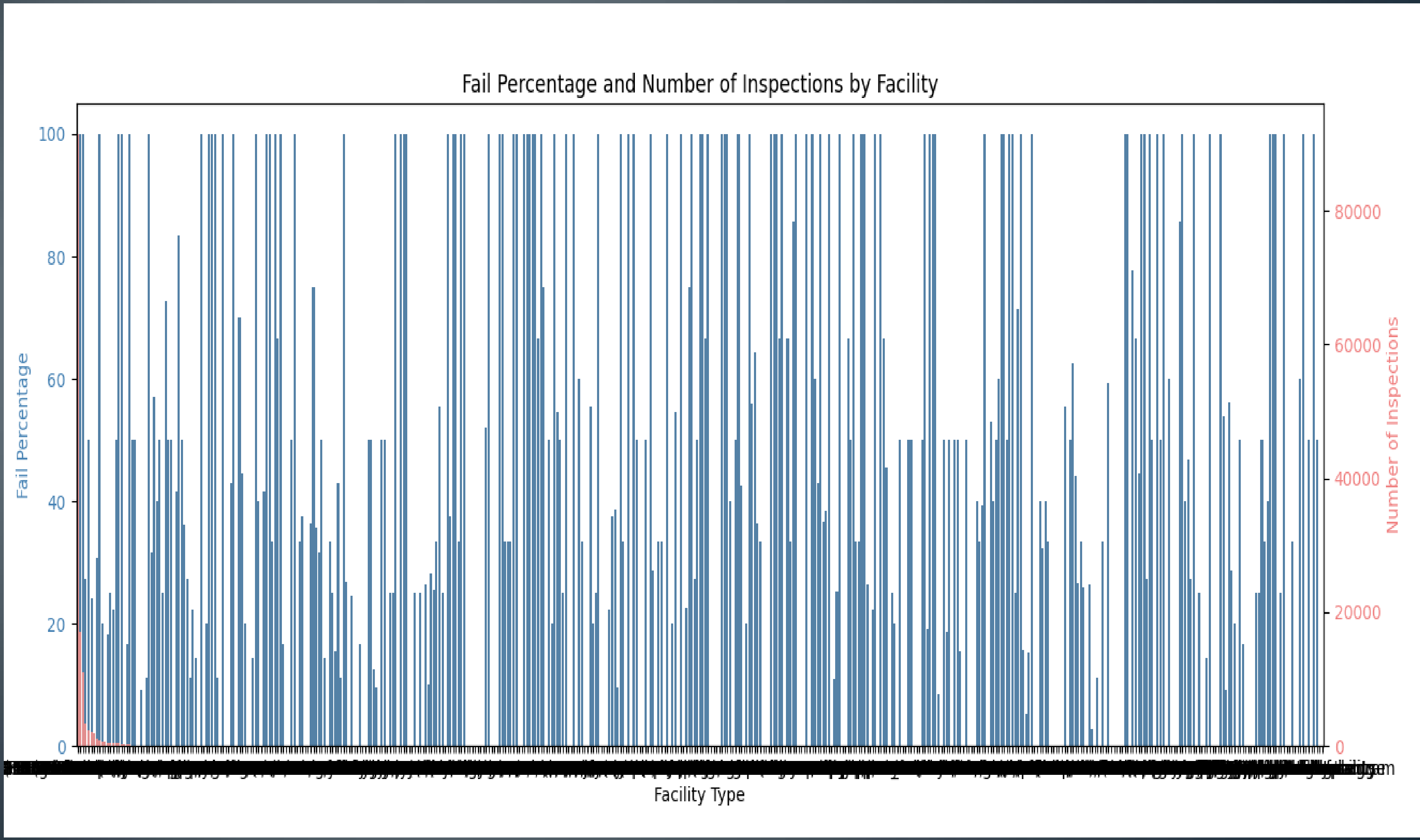
Exploratory Data Analysis



Exploratory Data Analysis



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

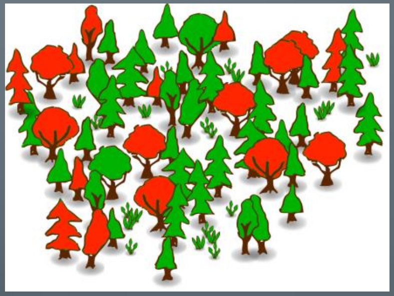
EDA

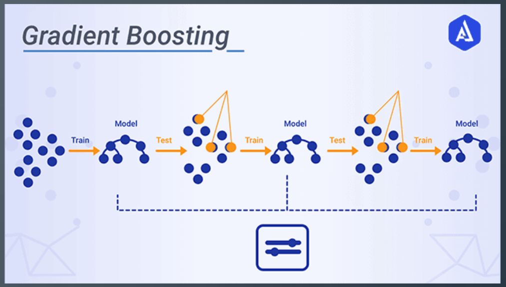
Fail rate analysis by inspection types

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

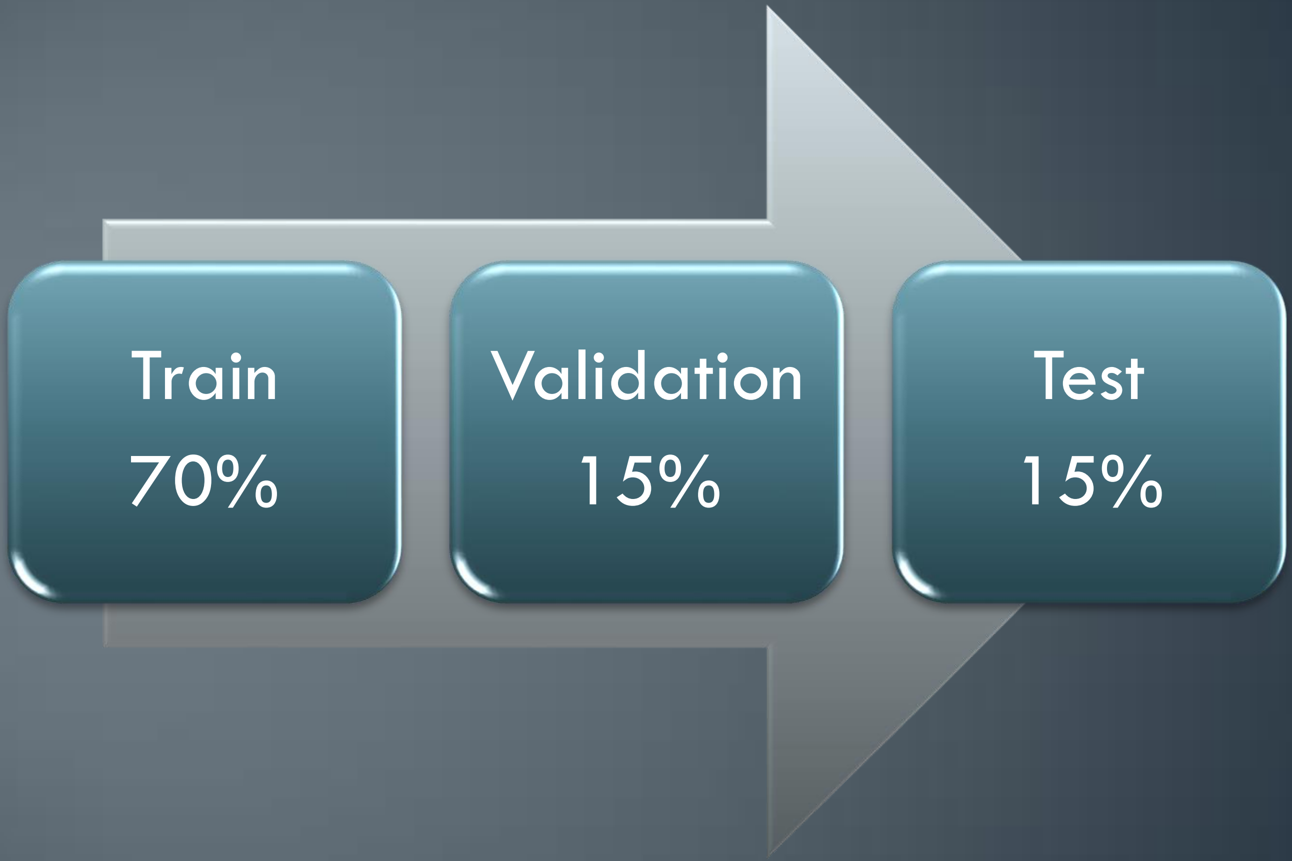
Model building





Four types of binary classification models considered.

Model Evaluation



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

No over-fit possibility

Model	Accuracy	Precision	Recall	F1
Logit Train	0.7268	0.6398	0.0142	0.0278
Logit Val	0.7279	0.6446	0.0145	0.0283
DF Train	0.7288	0.6641	0.0276	0.0530
DF Val	0.7293	0.6250	0.0284	0.0544
RF Train	0.7289	0.7060	0.0236	0.0458
RF Val	0.7293	0.6629	0.0237	0.0457
GB Train	0.7268	0.8114	0.0082	0.0162
GB Val	0.7281	0.8171	0.0091	0.0179

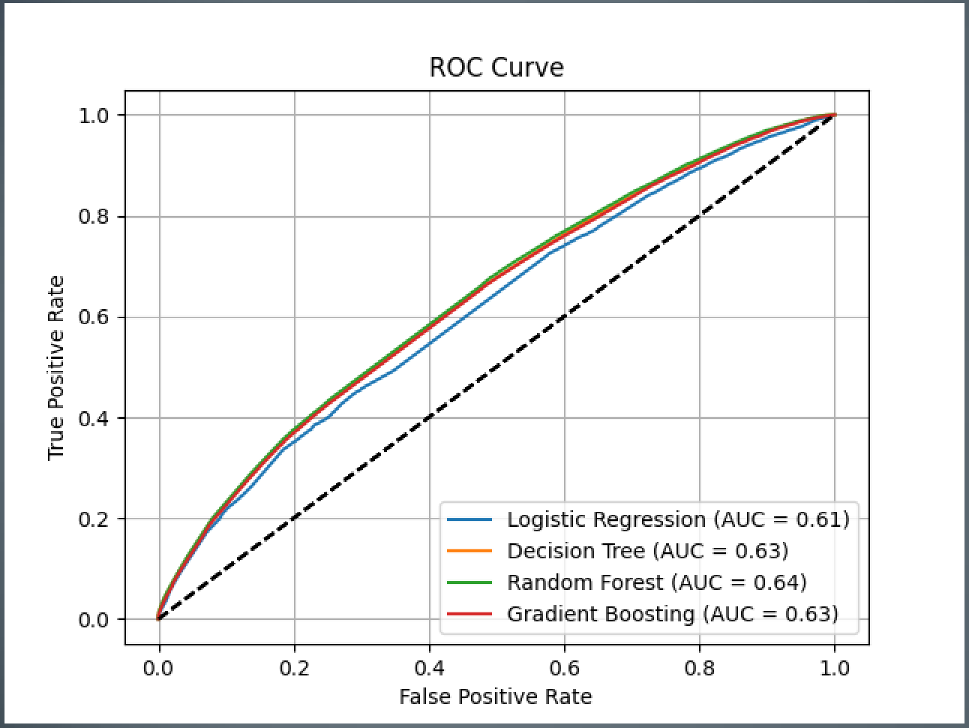
Model Results – Group #2

Over-fit possibility

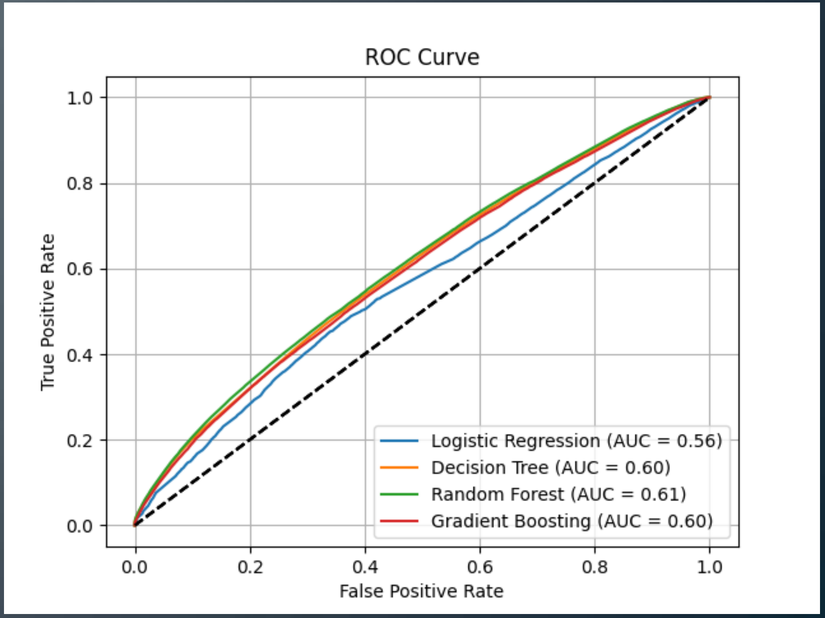
	Accuracy	Precision	Recall	F1
Logit Train	0.6846	0.6012	0.0628	0.1137
Logit Val	0.6736	0.5931	0.0599	0.1089
DF Train	0.6882	0.5757	0.1218	0.2011
DF Val	0.6771	0.5677	0.1221	0.2010
RF Train	0.6885	0.5809	0.1191	0.1977
RF Val	0.6774	0.5726	0.1182	0.1959
GB Train	0.6867	0.5817	0.0978	0.1675
GB Val	0.6755	0.5727	0.0965	0.1651

Model Evaluation

Model Group #2



Model Group #1



Hyper-parameter optimization

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

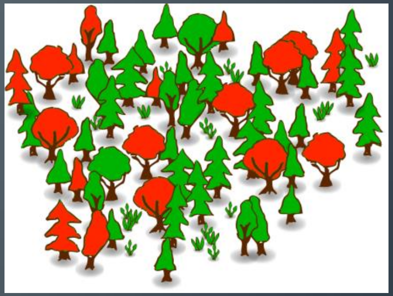
Decision Tree - Best Parameters:
{'max_depth': 10, 'min_samples_split': 10}

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	Models #2
Higher Accuracy & Precision	Lower Accuracy & Precision
Lower Recall & F1, ROC-AUC	Higher Recall & F1, ROC-AUC
No over-fit possibility	Over-fit possibility



Model Result prediction

	Logit test	DT test	RF test	GB test
Accuracy	0.6784	0.6789	0.6797	0.6776
Precision	0.5929	0.5466	0.5550	0.5490
Recall	0.0591	0.1173	0.1149	0.0921
F1	0.1074	0.1931	0.1905	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.**


KEEP
CALM
IT'S THE
LAST DAY
OF CLASS


THE
END
thank you all!