



Las Vegas Restaurant Inspections

Zizhong Liu
Springboard Capstone 3

1. Introduction

❖ Problem Statement

- The main goal of this Capstone project is to explore the possibility of building a classification model to predict the outcome of a restaurant's next inspection based on the provided data of the previous inspection in Las Vegas. The subtasks consist of 1) to analyze the provided information and interpret all the information; 2) to select important features and perform data cleaning and preprocessing; 3) to find the best classifier to create a model and predict the outcomes.

❖ Key Stakeholders

- A stakeholder can be defined generally as any individuals, end-users, or organizations who have an interest in the food regulation, in which restaurant inspection is one of the key components. It is important to take into consideration the views and concerns of relevant stakeholders in policymaking. The stakeholders can come from consumers, enterprises of food business, and food regulatory and enforcement agencies



2. Data Analysis

❖ Data Overview

- Info (10): RESTAURANT_SERIAL_NUMBER, RESTAURANT_PERMIT_NUMBER, RESTAURANT_NAME, RESTAURANT_LOCATION, ADDRESS, CITY, STATE, ZIP, LAT_LONG_RAW, VIOLATIONS_RAW
- Time (2): INSPECTION_TIME, RECORD_UPDATED.
- Categorical (6): RESTAURANT_CATEGORY, CURRENT_GRADE, INSPECTION_TYPE, FIRST_VIOLATION_TYPE, SECOND_VIOLATION_TYPE, THIRD_VIOLATION_TYPE.
- Target (1): NEXT_INSPECTION_GRADE_C_OR_BELOW (Have noise).
- Continuous(9): CURRENT_DEMERITS, EMPLOYEE_COUNT, MEDIAN_EMPLOYEE_AGE, MEDIAN_EMPLOYEE_TENURE, INSPECTION_DEMERITS(NA inside), FIRST_VIOLATION, SECOND_VIOLATION, THIRD_VIOLATION, NUMBER_OF_VIOLATIONS (NA inside).

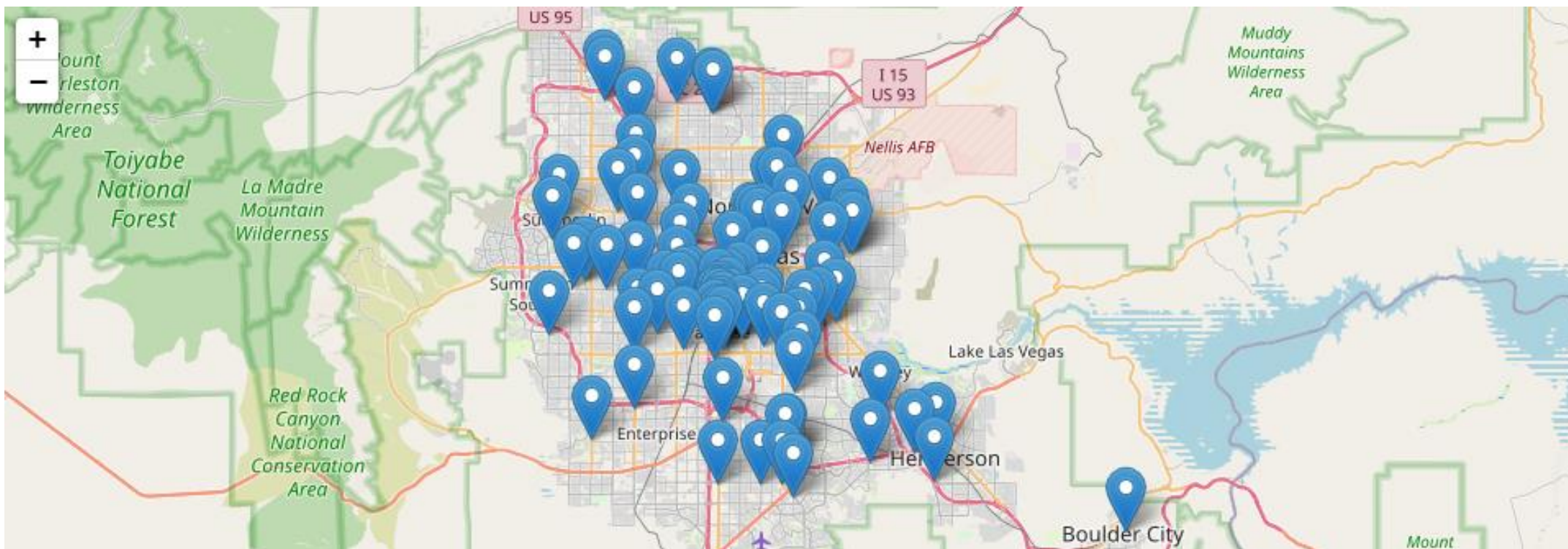
Data.dtypes

RESTAURANT_SERIAL_NUMBER	object
RESTAURANT_PERMIT_NUMBER	object
RESTAURANT_NAME	object
RESTAURANT_LOCATION	object
RESTAURANT_CATEGORY	object
ADDRESS	object
CITY	object
STATE	object
ZIP	object
CURRENT_DEMERITS	float64
CURRENT_GRADE	object
EMPLOYEE_COUNT	float64
MEDIAN_EMPLOYEE_AGE	float64
MEDIAN_EMPLOYEE_TENURE	float64
INSPECTION_TIME	object
INSPECTION_TYPE	object
INSPECTION_DEMERITS	object
VIOLATIONS_RAW	object
RECORD_UPDATED	object
LAT_LONG_RAW	object
FIRST_VIOLATION	float64
SECOND_VIOLATION	float64
THIRD_VIOLATION	float64
FIRST_VIOLATION_TYPE	object
SECOND_VIOLATION_TYPE	object
THIRD_VIOLATION_TYPE	object
NUMBER_OF_VIOLATIONS	object
NEXT_INSPECTION_GRADE_C_OR_BELOW	object
dtype:	object

2. Data Analysis

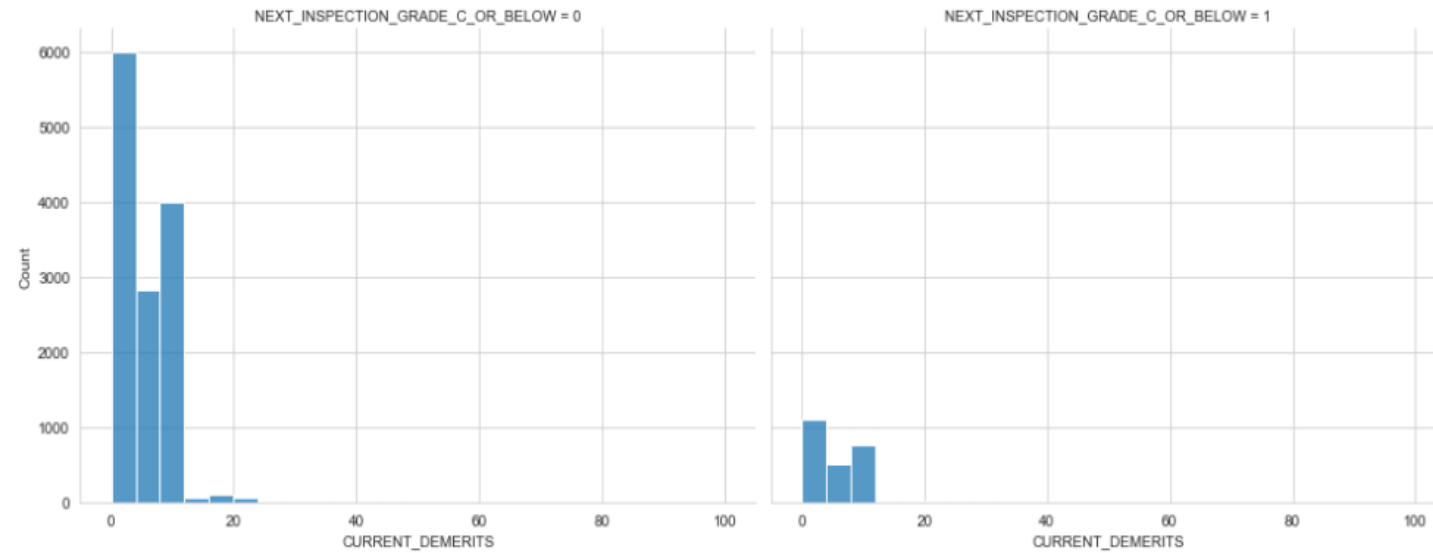
❖ Continuous Features

RESTAURANT_LOCATION

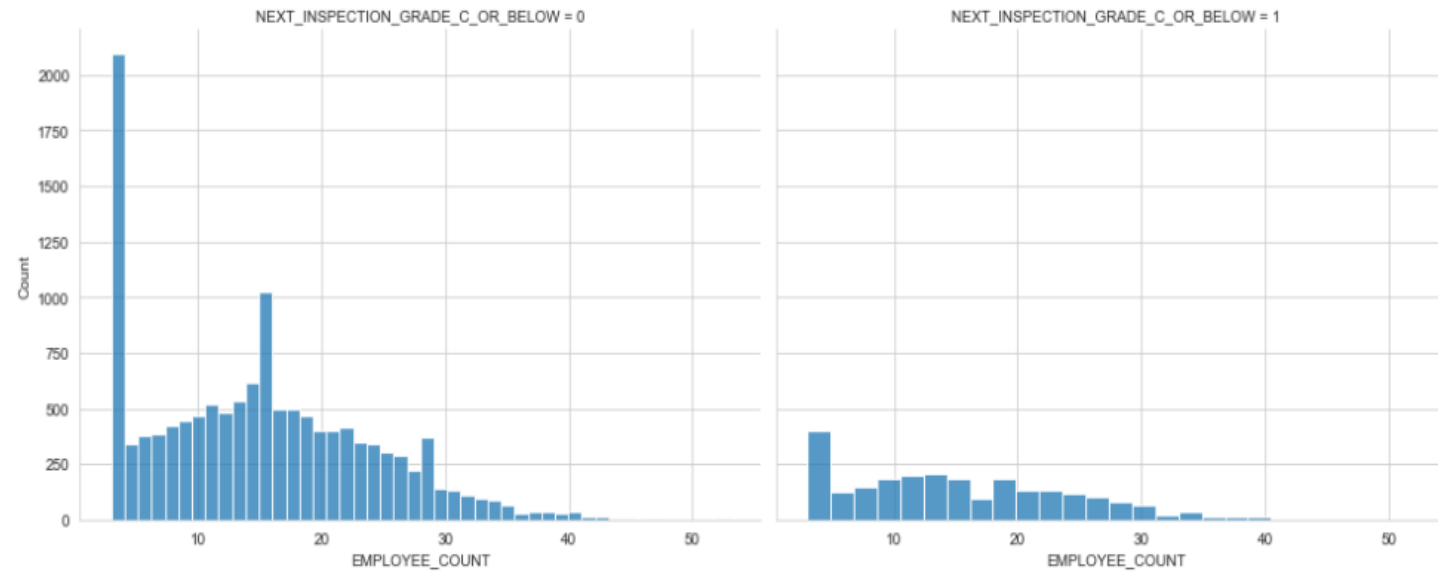


2. Data Analysis ❖ Continuous Features

'CURRENT_DEMERITS'

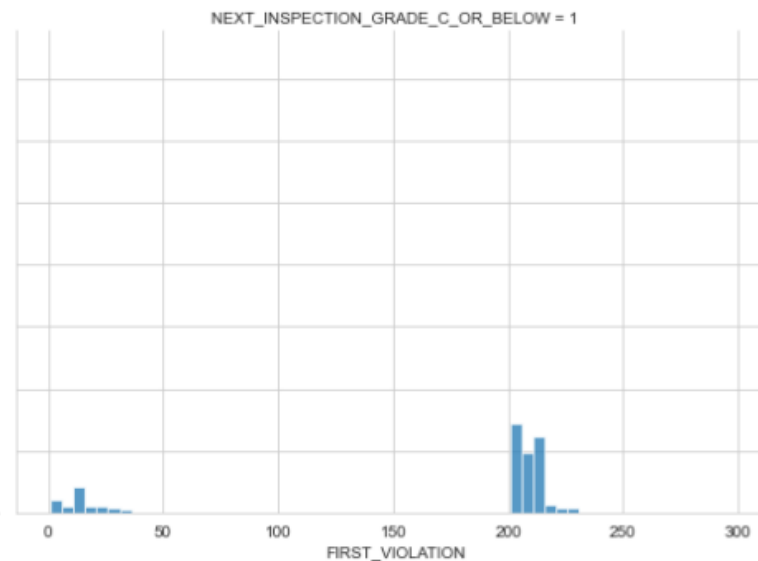
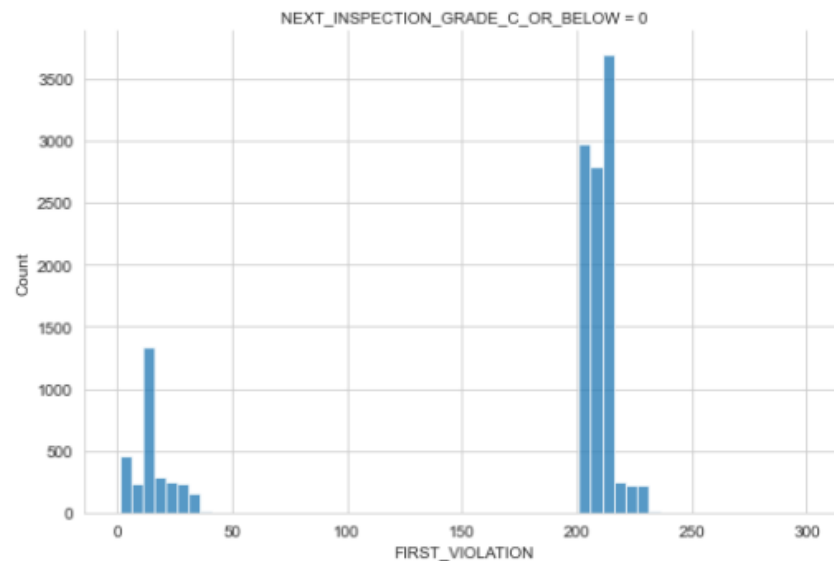


'EMPLOYEE_COUNT'

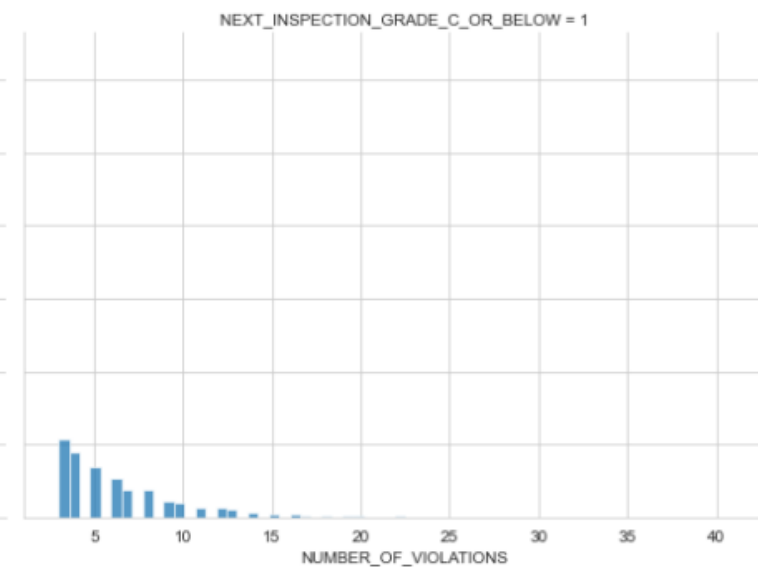
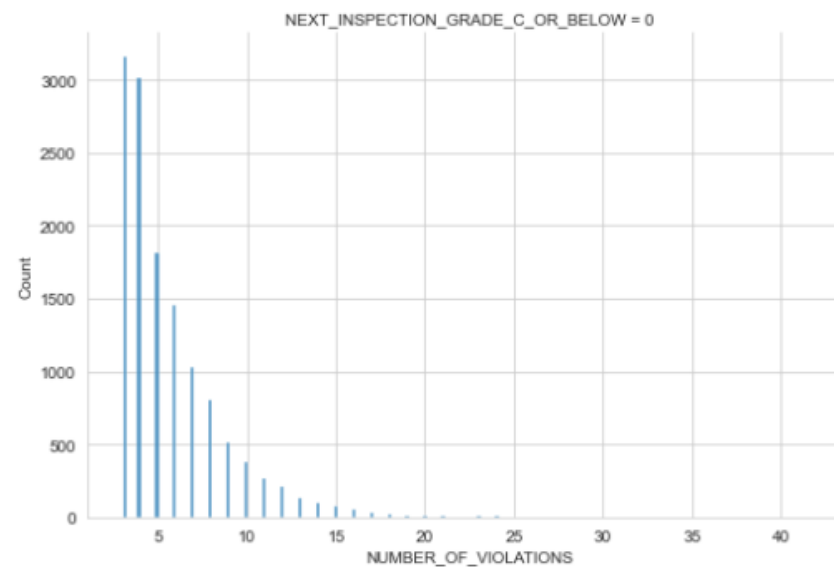


2. Data Analysis ❖ Continuous Features

'FIRST_VIOLATION'

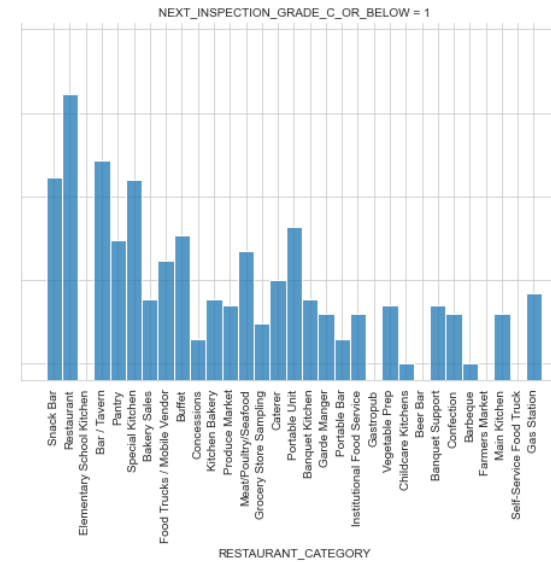
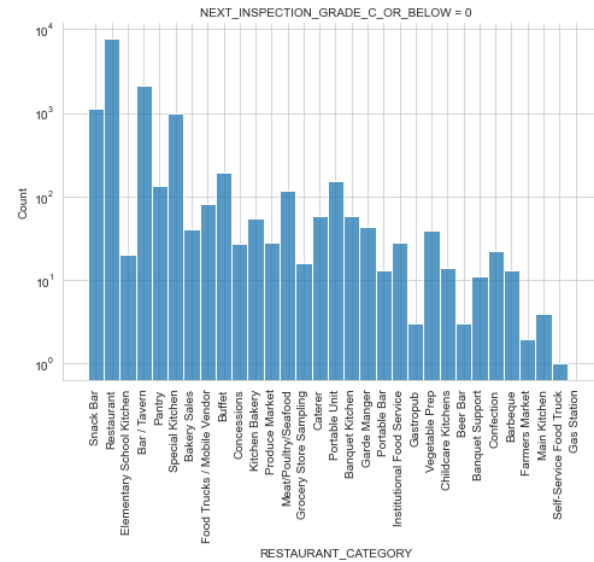


'NUMBER_OF_VIOLATIONS'

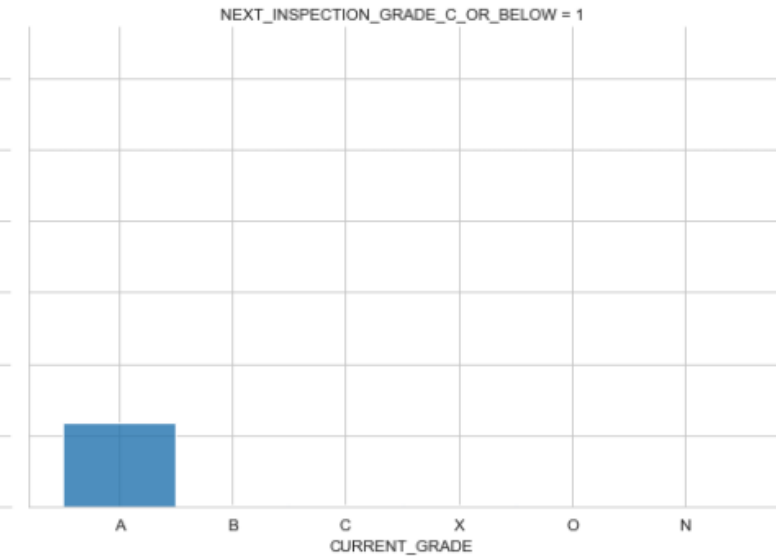
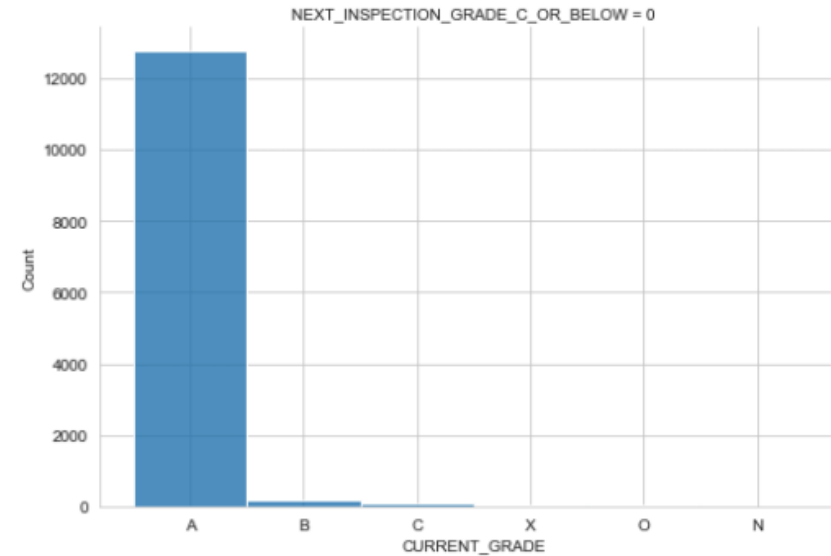


2. Data Analysis ❖ Continuous Features

'RESTAURANT_CATEGORY'



'CURRENT_GRADE'



2. Data Analysis ❖ Continuous Features

'FIRST_VIOLATION_TYPE'

RESTAURANT_SERIAL_NUMBER	
FIRST_VIOLATION_TYPE	
Critical	7307
Imminent Health Hazard	3
Major	6698
Non-Major	1580

'SECOND_VIOLATION_TYPE'

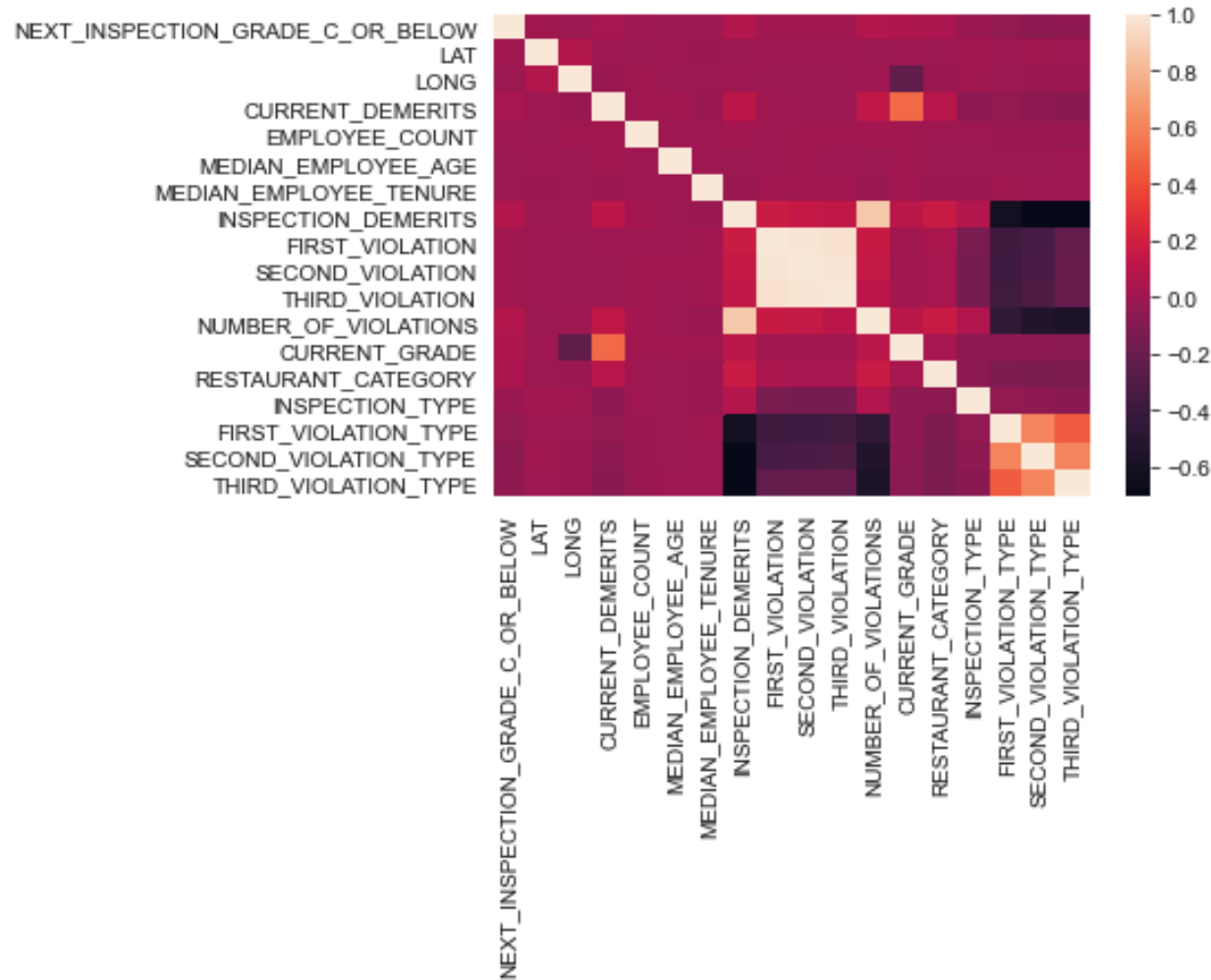
RESTAURANT_SERIAL_NUMBER	
SECOND_VIOLATION_TYPE	
Critical	2972
Imminent Health Hazard	5
Major	8134
Non-Major	4476

'THIRD_VIOLATION_TYPE'

RESTAURANT_SERIAL_NUMBER	
THIRD_VIOLATION_TYPE	
Critical	866
Imminent Health Hazard	37
Major	7444
Non-Major	7240

3. Featuring Engineering

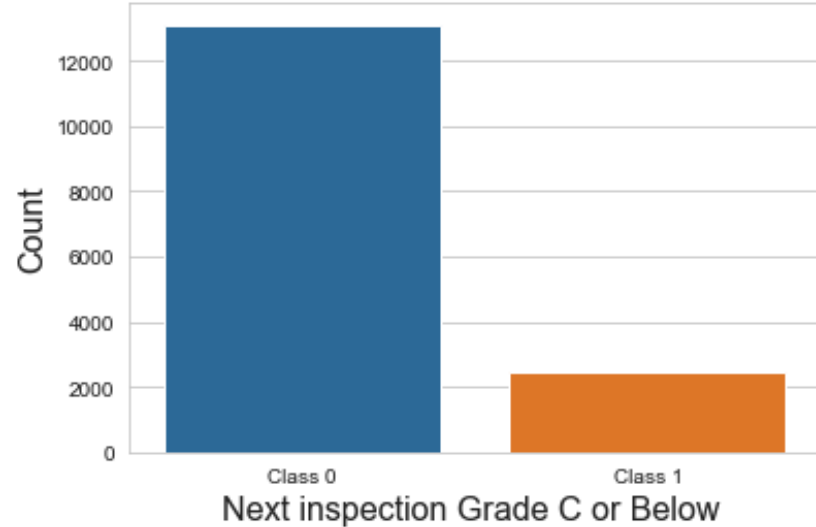
❖ Feature Correlations



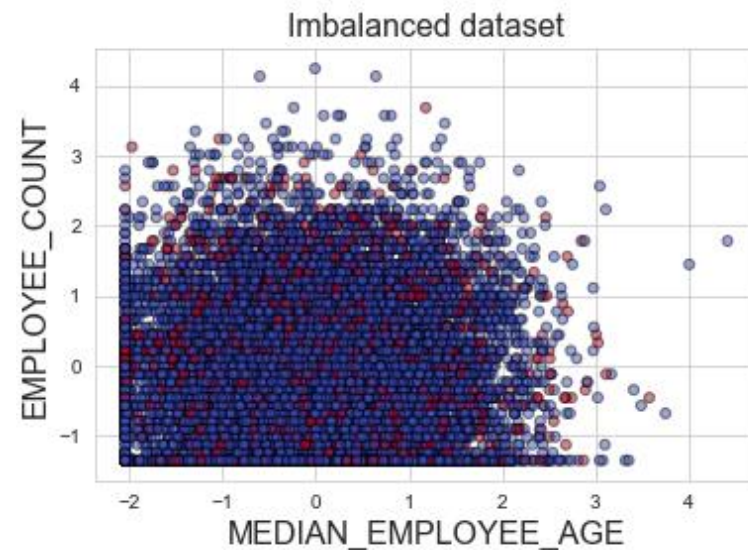
Investigating the correlation between class and each feature can help us to select the best features. The results show that there are both positive and negative correlations.

3. Featuring Engineering

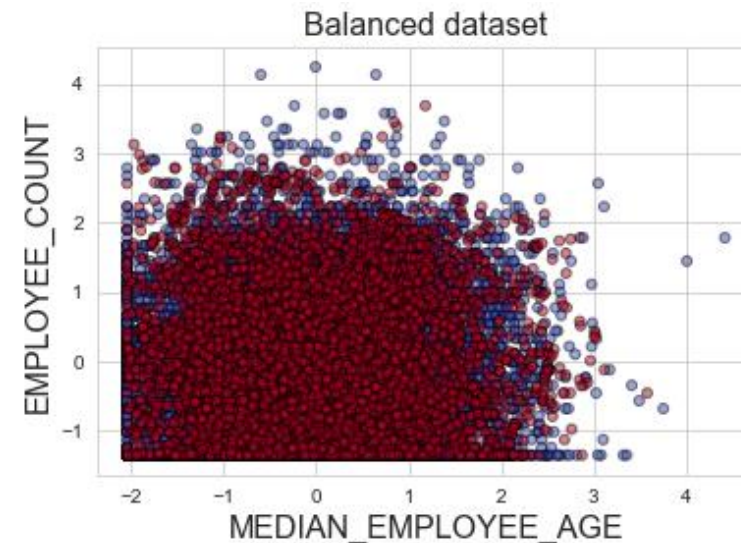
❖ Imbalanced data



Before using this data in the model, we need to pay attention to the distribution of the class NEXT_INSPECTION_GRADE_C_OR_BELOW. Counting the number of zeros and ones, we can find that we have imbalanced data.



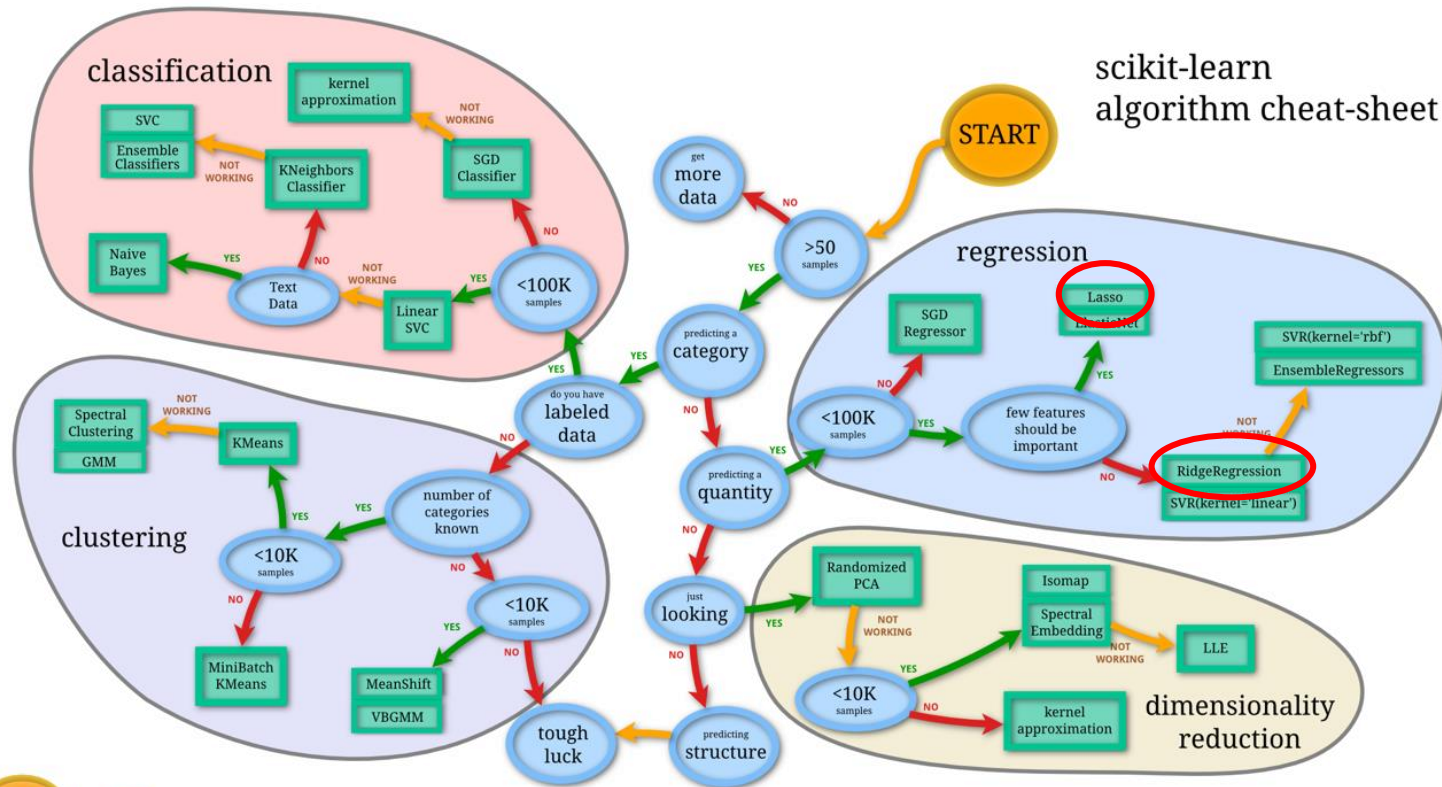
SMOTE



4. Modeling

- ❖ Model Selection

❖ Model Selection



- Logistic Regression
- Decision Tree
- Random Forest
- KNN or k-Nearest Neighbors
- Naive Bayes classifier
- Gradient Boosting Classifier
- Support Vector Machines



(https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html)

4. Modeling

❖ Model Evaluation

		CONDITION determined by "Gold Standard"			
		TOTAL POPULATION	CONDITION POS	CONDITION NEG	PREVALENCE $\frac{\text{CONDITION POS}}{\text{TOTAL POPULATION}}$
TEST OUT- COME	TEST POS	True Pos TP	Type I Error False Pos FP	Precision Pos Predictive Value $PPV = \frac{TP}{\text{TEST P}}$	False Discovery Rate $FDR = \frac{FP}{\text{TEST P}}$
	TEST NEG	Type II Error False Neg FN	True Neg TN	False Omission Rate $FOR = \frac{FN}{\text{TEST N}}$	Neg Predictive Value $NPV = \frac{TN}{\text{TEST N}}$
ACCURACY ACC $ACC = \frac{TP + TN}{\text{TOT POP}}$		Sensitivity (SN), Recall Total Pos Rate TPR = $\frac{TP}{\text{CONDITION POS}}$	Fall-Out False Pos Rate FPR = $\frac{FP}{\text{CONDITION NEG}}$	Pos Likelihood Ratio LR + = $\frac{TPR}{FPR}$	Diagnostic Odds Ratio DOR = $\frac{LR +}{LR -}$
		Miss Rate False Neg Rate FNR = $\frac{FN}{\text{CONDITION POS}}$	Specificity (SPC) True Neg Rate TNR = $\frac{TN}{\text{CONDITION NEG}}$	Neg Likelihood Ratio LR - = $\frac{TNR}{FNR}$	


Confusion matrix (<https://www.unite.ai/what-is-a-confusion-matrix/>)

- Accuracy: Percentage of correct prediction.
- Recall: True Positive / (True Positive + False Negative)
- Precision: True Positive / (True Positive + False Positive)
- **F1-Score: The harmonic mean of precision (PRE) and recall (REC)**

4. Modeling

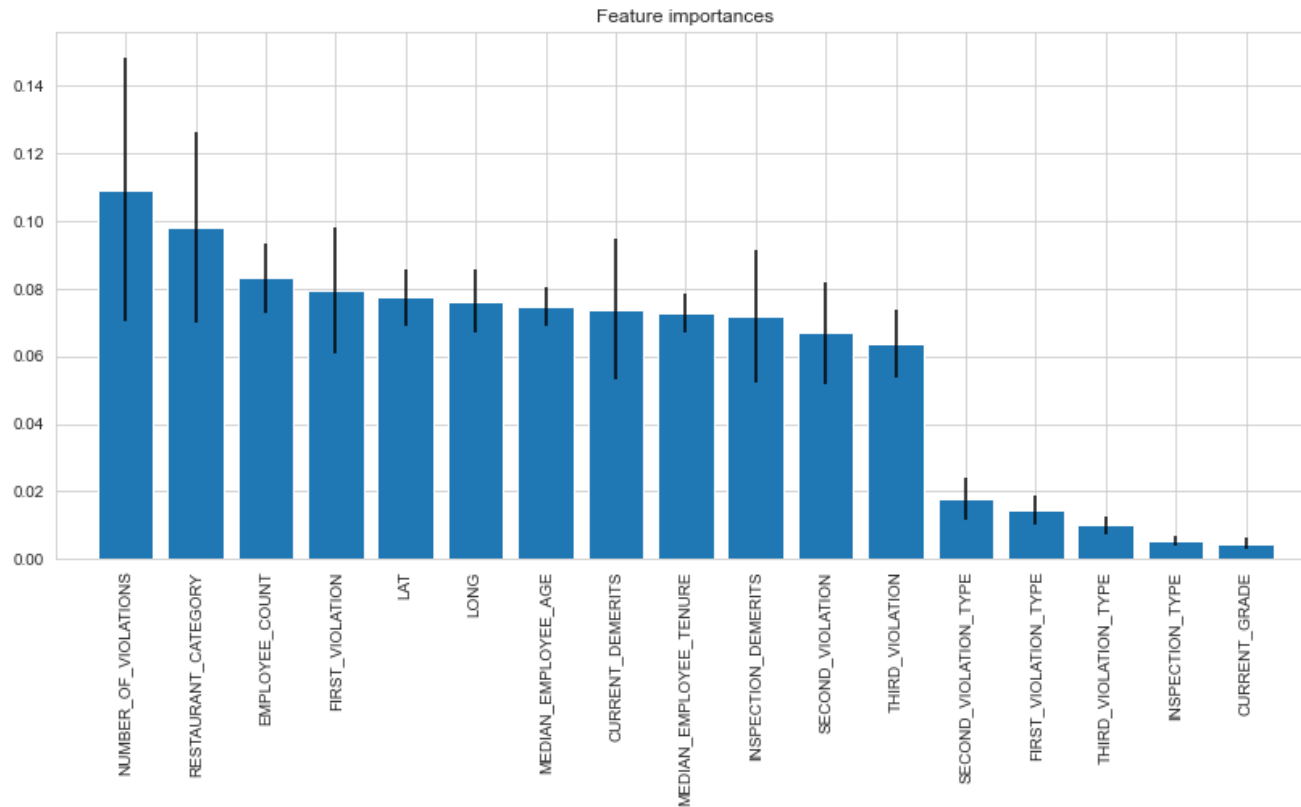
❖ Results



model	f1 train	f1 val 	Accuracy score train	Accuracy score val	best_estimator
Decision Tree Classifier	0.642973	0.843169	0.642973	0.843169	DecisionTreeClassifier(max_depth=6, min_samples_leaf=9, min_samples_split=5)
Gradient Boosting Classifier	0.905686	0.827774	0.905686	0.827774	GradientBoostingClassifier(learning_rate=0.5)
Gaussian NB	0.529243	0.765876	0.529243	0.765876	{'priors': None, 'var_smoothing': 1e-09}
Random Forest Classifier	0.826639	0.761065	0.826639	0.761065	RandomForestClassifier(max_depth=9, max_features=4, min_samples_leaf=3, min_samples_split=3)
K Neighbors Classifier	0.872913	0.615779	0.872913	0.615779	KNeighborsClassifier(leaf_size=3, p=1)
Logistic Regression	0.544032	0.58948	0.544032	0.58948	LogisticRegression(C=0.001, max_iter=1, penalty='none', solver='sag')
SVC	0.594791	0.552919	0.594791	0.552919	SVC(C=10, gamma=0.01)

4. Modeling

❖ Feature Importance

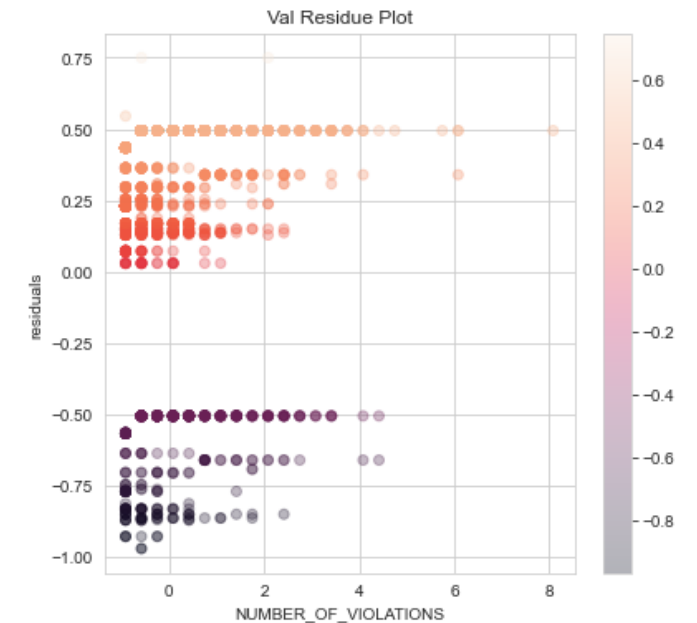
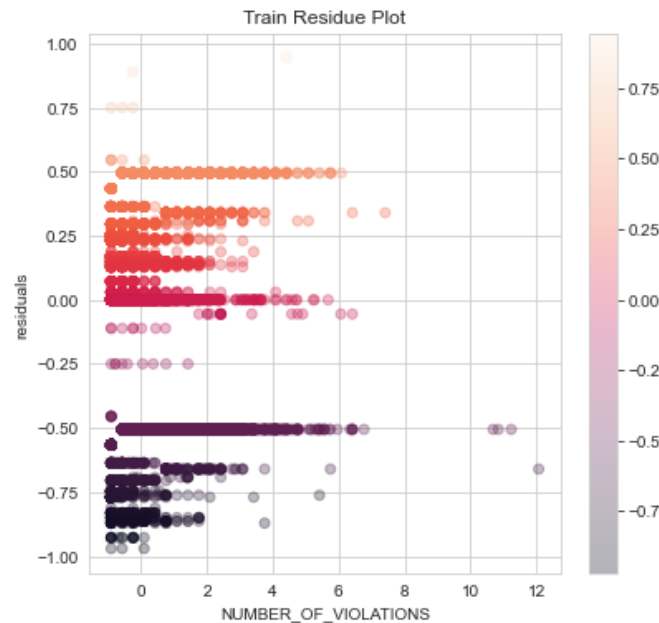
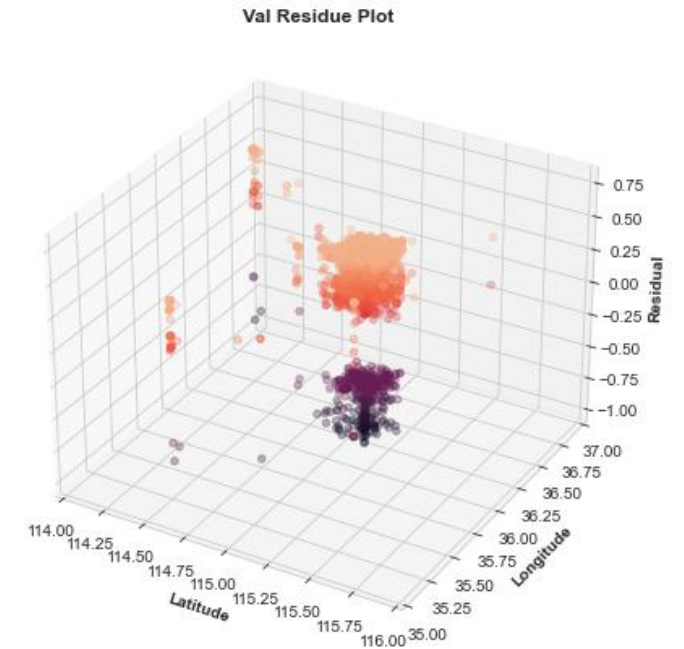
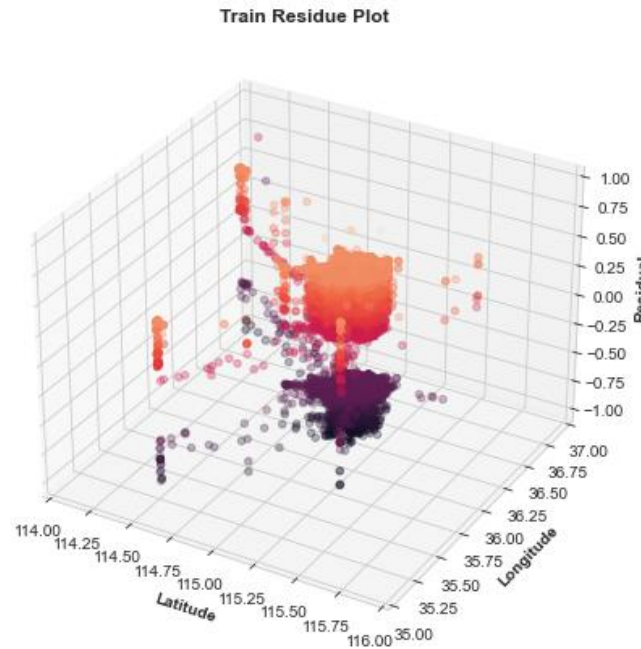


Random forest is able to provide feature importance. As shown in the figure below, the 'NUMBER_OF_VIOLATIONS' is the most important feature.

4. Modeling

❖ Residual Plots

The residuals of the probabilities ('NEXT_INSPECTION_GRADE_C_OR_BELOW' = 1) predicted by the best model, Decision Tree, are visualized in terms of the location and 'NUMBER_OF_VIOLATIONS', as shown below. The residual plot shows a random pattern, indicating a good fit for the classification model.



Conclusions

- ❖ This project investigated the data sets of restaurant inspections in Las Vegas and then built classification models to predict the restaurant's next inspection below or above C.
- ❖ Through the project, the model development cycle goes through various important stages, including data collection, data cleaning, data analysis, featuring engineering and model building.
- ❖ Based on the provided dataset, information, and results, Decision Tree is the optimum model for this dataset with hyperparameters of `max_depth = 6`, `min_samples_leaf = 9`, and `min_samples_split = 5`.

❖ Future works:

- For future work to achieve a better prediction, adding more data for class 1 to avoid imbalanced datasets is needed.
- In addition, feature importance results show that features like number of violations, restaurant category, employee count, first violation, and locations are important to predict the next inspection. Therefore, implementing further featuring engineering to create new features out of existing ones can improve the accuracy of the model.