CAPSTONE REPORT

Analysis of Las Vegas Restaurant Inspections

Yash Nikhare

INTRODUCTION

This dataset explores recent restaurant inspections in Las Vegas, Nevada, offering a comprehensive assessment of the city's dining establishments. It provides inspection results, key information such as inspection dates, and the types of inspections conducted by the Southern Nevada Health District. The project commences with a thorough understanding of the data's nuances, followed by an investigation into the predictability of the next inspection grade using a variety of numerical and categorical variables through classification models.

DATA

There are **17196 rows** in the dataset, each of which represents a unique listing. There are **29 columns**, each of which represents a unique feature, which has been described in the table below:

Column Name	Description	Type
RESTAURANT_SERIAL_ NUMBER	Restaurant's unique identifier for the listing.	object
RESTAURANT_PERMIT _NUMBER	Restaurant's permit number.	object
RESTAURANT_NAME	Name of the Restaurant.	object
RESTAURANT_LOCATION	Restaurant's location	object
RESTAURANT_CATEGORY	Categories of restaurant	object
ADDRESS	Restaurant's Address	object
CITY	Name of a city where restaurants are situated	object

STATE	Name of the state where restaurants are situated	object
ZIP	Zip code basically it's a system of postal codes	object
CURRENT_DEMERITS	Represent the demerit score assigned to a restaurant during inspections, reflecting its compliance with health and safety regulations	float64
CURRENT_GRADE	Represents the current inspection grade assigned to a restaurant,	object
EMPLOYEE_COUNT	The number of total employees in particular restaurant category.	float64
MEDIAN_EMPLOYEE_AGE	The median age of employees in particular restaurant category	object
MEDIAN_EMPLOYEE_TENURE	The median employees tenure in particular restaurant category.	float64
INSPECTION_TIME	Time of the inspection	object
INSPECTION_TYPE	Type of inspection conducted.	object
INSPECTION_DEMERITS	Demerits assigned during inspection.	object
VIOLATIONS_RAW	Raw violation data.	object
RECORD_UPDATED	Timestamp of record update.	object
LAT_LONG_RAW	latitude and longitude data.	object
FIRST_VIOLATION	First recorded violation.	float64
SECOND_VIOLATION	Second recorded violation.	float64
THIRD_VIOLATION	Third recorded violation.	float64
FIRST_VIOLATION_TYPE	Type of the first violation.	object
SECOND_VIOLATION_TYPE	Type of the second violation.	object
THIRD_VIOLATION_TYPE	Type of the third violation.	object
NUMBER_OF_VIOLATIONS	Total number of violations.	object
NEXT_INSPECTION_GRADE_C_OR_BELOW	Indicates if the next inspection grade is C or below.	float64
INPSECTION_FORMAT	Format of the inspection data.	object

DATA CLEANING

1: Evaluated missing values in each feature, removed the column with the highest percentage of null values, and dropped irrelevant columns like 'INPSECTION_FORMAT,' 'RESTAURANT_SERIAL_NUMBER,' 'RESTAURANT_PERMIT_NUMBER,' 'RESTAURANT_NAME,' 'RESTAURANT_LOCATION,' and 'ZIP' to enhance data quality and relevance for analysis. By doing this, I made the dataset cleaner and more focused on what truly matters for our analysis.

2: Feature Preprocessing: The features **'NEXT_INSPECTION_GRADE_C_OR_BELOW,' 'EMPLOYEE_COUNT,' 'MEDIAN_EMPLOYEE_AGE,' 'CURRENT_DEMERITS,' 'RESTAURANT_CATEGORY,'** and others undergo a series of cleaning and preprocessing steps. This process involves addressing outliers, applying mapping, and converting them to suitable data types. Each feature is preprocessed based on its unique characteristics and specific requirements. This ensures that the data is prepared appropriately for analysis in a data scientist's context.

VISUALIZATION

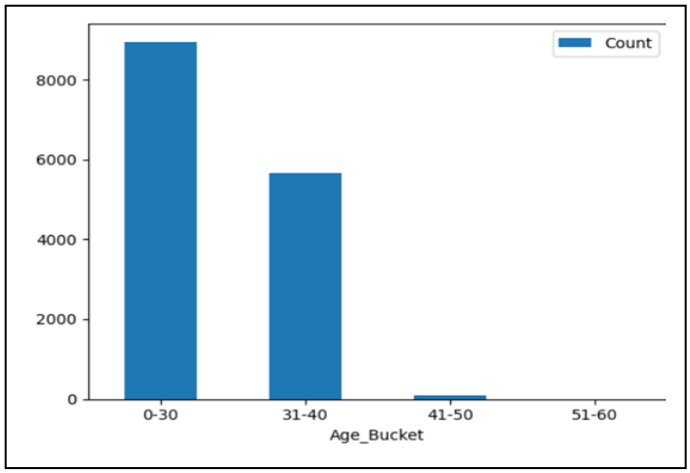

Figure 1 - Distribution of Employee's Age following outlier removal

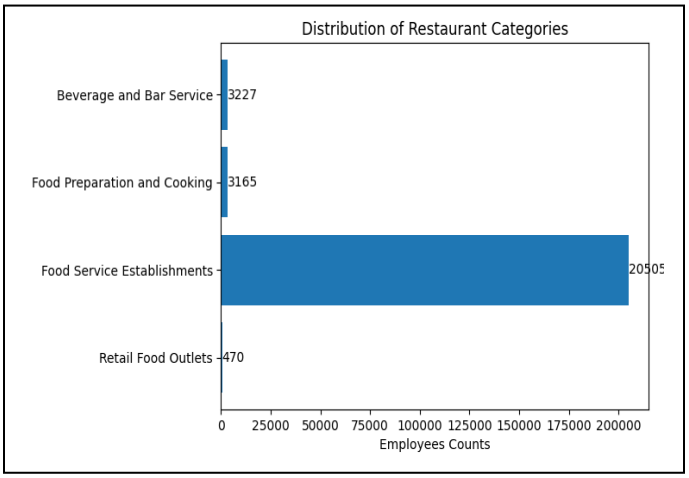

Figure 2 - Distribution of Employee's Count across Restaurant categories

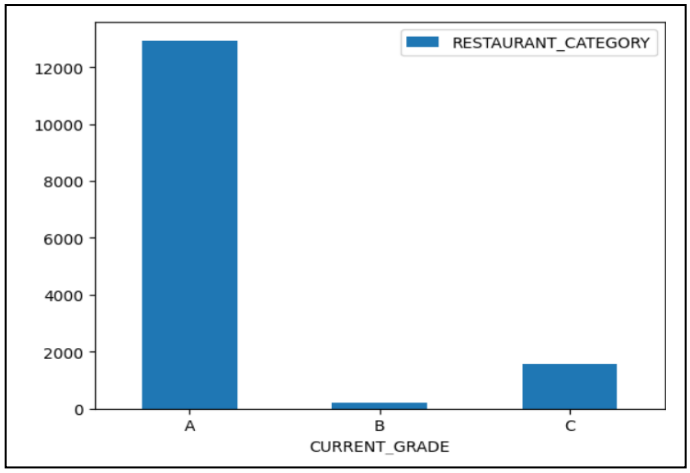

Figure 3 - Distribution of Current grade

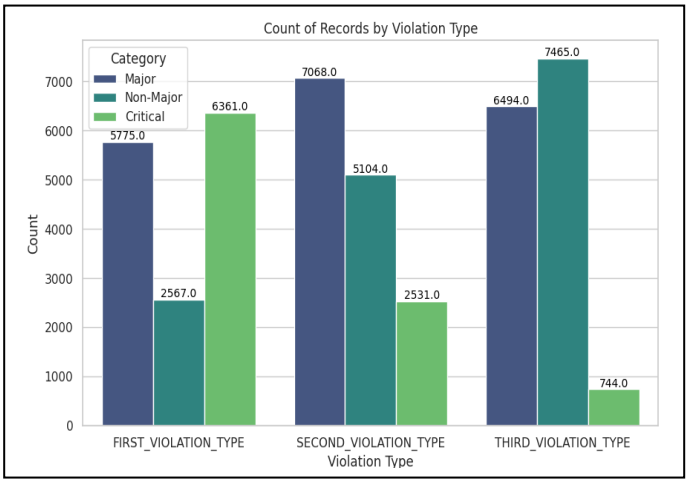

Figure 4 - Distribution of Current grade

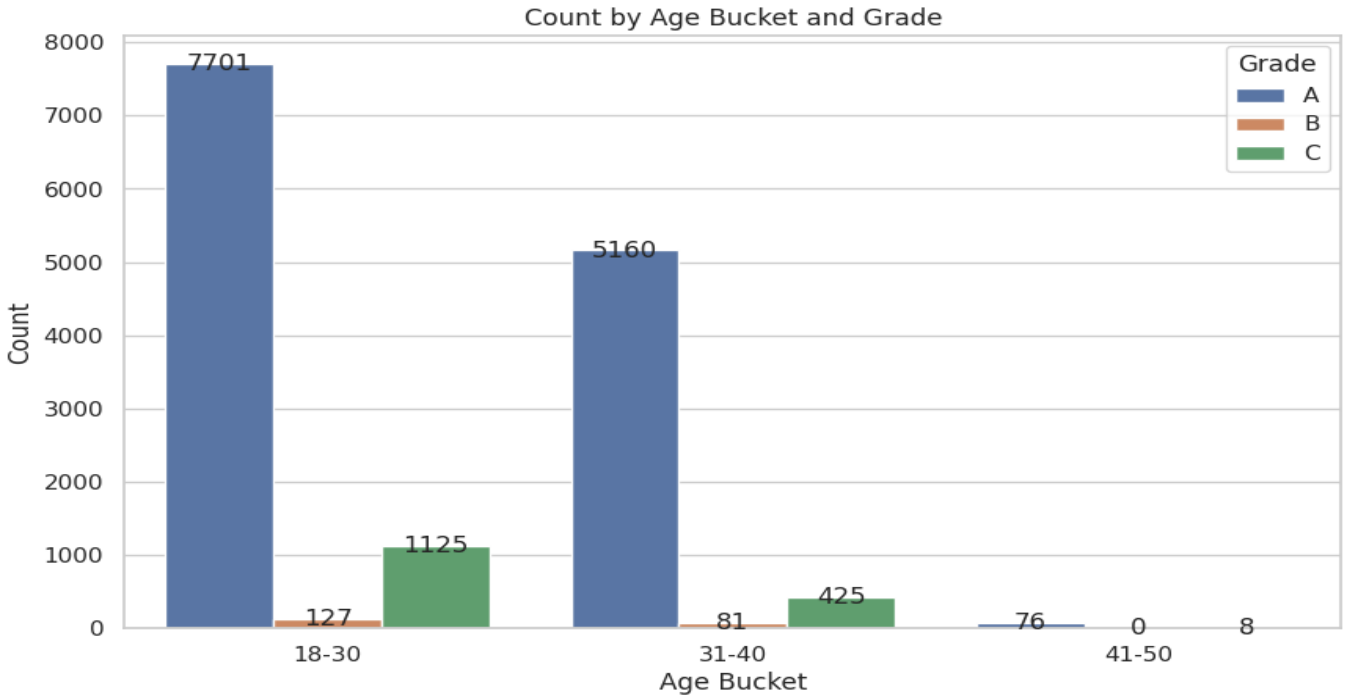

Figure 5 - Distribution of Employee's Age and Current Grade

HYPOTHESIS TESTING

We were mostly interested in analyzing the relationship between the Next Inspection Grade with factors such as the CURRENT_GRADE, CURRENT_DEMERITS, INSPECTION_DEMERITS, RESTAURANT_CATEGORY .This analysis not only enhances our comprehension of dataset dynamics but also lays the groundwork for subsequent exploratory investigations, specifically in the context of the restaurant dataset.

The reason why to implement Kruskal-Wallis test:

The Kruskal-Wallis (KW) test is a non-parametric statistical tool ideal for comparing more than two independent groups and assessing significant differences among them. Unlike one-way ANOVA, the KW test doesn't rely on data conforming to a specific distribution, making it suitable for non-normally distributed datasets. Its robustness against outliers and varying group sample sizes ensures it can handle unevenly sized groups and extreme values effectively. Whether dealing with continuous or ordinal data, the KW test's versatility is evident. The p-value obtained from the KW test aids in determining whether statistically significant differences exist among groups, guiding further investigation and decision-making. With wide-ranging applications across fields like healthcare, social sciences, and quality control, the KW test is an essential tool for group or treatment comparisons.

***For all the tests conducted, level of significance (α) is set to 0.05**

Hypothesis Test 1: Does the Current Demerits of the listings vary by Current Grade?

First, we divided the listings current grade wise and removed listings where the grade was not mentioned or was zero. Our **null hypothesis** is that the **current grade does not vary across the current demerit**. To answer this question, implemented a two-sided Kruskal Wallis test. Given our group standard deviations were not equal, we decided it was not appropriate to use an ANOVA. Therefore, we instead chose to implement the non-parametric version of ANOVA, known as the Kruskal-Wallis test. It is worth noting that the KW test's non-parametric nature results in a loss of power and that the number of samples in our current grade groups is not equally distributed. However,we felt the collective dataset size. In running the test, we obtain a p-value that is significant atthe alpha level of 0.05. Hence, we **reject the null hypothesis**. Therefore, there is evidence that suggests **the current grade varies by the current demerit** and this result is not surprising.

Hypothesis Test 2: Do Median Employee Age vary by Current Grade?

Our null **hypothesis is** that median **employee age do not have an affect by current grade**. Our alternative hypothesisis that median employee age has an effect by current grade. The Kruskal-Wallis test was chosen due to its robustness and suitability for comparing groups of non-normally distributed, ordinal data. Our null hypothesis (H0) assumes that there is no substantial impact of median employee age on current grades, while the alternative hypothesis (H1) proposes that there is a significant influence of median employee age on these grades. In running the test, we obtain a p-value that is significant atthe alpha level of 0.05. Therefore, we reject **the null hypothesis**. This analysis will shed light on the correlation between employee age and the assigned current grades, which is of relevance to our understanding of restaurant inspections and performance evaluations.

Hypothesis Test 3: Do Current Grades affect the Next Inspection Grade C or Below?

Again, we applied a split to our data to create three groups grade A, grade B, grade C. We The **null hypothesis (H0) posits that there is no substantial effect** of current grades on the likelihood of receiving such a grade in the next inspection, while the **alternative hypothesis (H1) suggests that there is a significant impact of current grades on future inspection outcomes**. This analysis aims to provide insights into whether current grades serve as a reliable predictor for future inspection results, a critical aspect of restaurant evaluation and compliance. The **rejection of the null hypothesis** indicates a noteworthy connection between current grades and subsequent inspection performance.

Hypothesis Test 4 : Do Current Demerit affect the Next Inspection Grade C or Below?

First, we performed a split of the data, dividing into two groups: those with received grade C and those who not received grade C . The null hypothesis (H0) posits that current demerits have no substantial impact on the likelihood of receiving a "Next Inspection Grade C or Below," while the alternative hypothesis (H1) suggests that there is a notable effect. We ran a KW test on the groups and obtained a p-value < 0.05, leading us to **reject the null hypothesis.** The rejection of the null hypothesis implies a significant connection between current demerits and future inspection grades. It implies that restaurants with higher demerit points tend to exhibit a higher likelihood of non-compliance with

safety and regulatory standards. This statistical finding provides important insights for restaurant owners and regulatory authorities. For restaurant owners, it signals that addressing and reducing current demerits may be pivotal in avoiding lower inspection grades.

Hypothesis Test 5: Restaurant Categories affect on Next Inspection Grade C or Below?

First, we performed a split of the data, dividing into four groups: those with Food Service Establishments, Food Preparation and Cooking, Beverage and Bar Service, Retail Food Outlets. The null hypothesis (H0) posits that Restaurant Categories have no substantial impact on the likelihood of receiving a "Next Inspection Grade C or Below," while the alternative hypothesis (H1) suggests that there is a notable effect. Running KW test we got a p-value < 0.05 and hence we **reject the null hypothesis**.

This finding has substantial implications for restaurant owners and regulatory agencies. Specifically, certain categories, such as "Food Service Establishments," "Food Preparation and Cooking," and "Beverage and Bar Service," are more prone to inspection grades of "C or Below." Conversely, categories like "Retail Food Outlets" show a significantly reduced likelihood of receiving such grades.

Restaurant owners can use this information to tailor their compliance efforts and address category-specific compliance issues. Regulatory authorities can focus their inspection and enforcement resources more efficiently by recognizing the varying risk profiles across different restaurant categories. The statistical evidence supports the idea that restaurant categories do matter in predicting inspection outcomes.

Hypothesis Test 6: Does Inspection type affect the Next Inspection Grade C or Below?

First, we performed a split of the data, dividing into two groups: those with received grade C and those who not received grade C . The null hypothesis (H0) posits that Inspection type have no substantial impact on the likelihood of receiving a "Next Inspection Grade C or Below," while the alternative hypothesis (H1) suggests that there is a notable effect. We ran a KW test on the two groups and obtained a p-value > 0.05, hence leading us to **fail to reject the null hypothesis**.

The failure to reject the null hypothesis suggests that, based on the data available, we cannot confidently assert that the type of inspection significantly affects the future inspection grades assigned to restaurants. This finding may imply that other factors not considered in this analysis, such as specific compliance issues or regional variations, might play a more crucial role in predicting inspection outcomes. Further investigation or a larger dataset may be necessary to draw more definitive conclusions regarding the relationship between Inspection Type and inspection grades.

CLASSIFICATION

For all of our models I used a standard train:test split of 0.8:0.2.

Following our hypothesis testing, we were interested in implementing a **binary classification** model to classify the whether the restaurant would receive C grade or not of listings based on the following 9 parameters:' **RESTAURANT_CATEGORY, 'CURRENT_DEMERITS', 'CURRENT_GRADE',**

'EMPLOYEE_COUNT', MEDIAN_EMPLOYEE_AGE', 'MEDIAN_EMPLOYEE_TENURE', ' FIRST_VIOLATION', 'SECOND_VIOLATION',' NUMBER_OF_VIOLATIONS'.

THIRD_VIOLATION and INSPECTION_DEMERITS are dropped because of having strong correlation with SECOND_VIOLATION' and NUMBER_OF_VIOLATIONS' respectively.

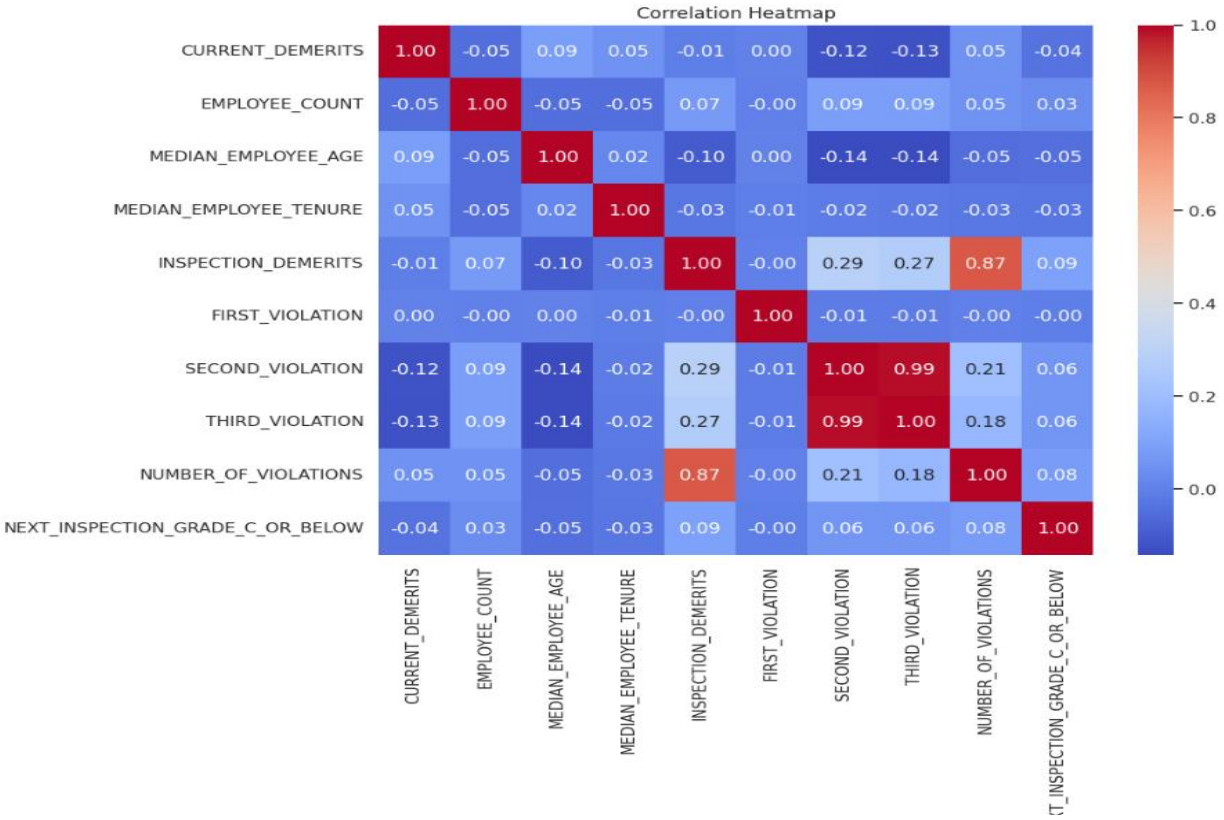

Figure 6 – Heatmap (Correlation matrix) among features

After this, one-hot encoding was applied to the 'RESTAURANT_CATEGORY' column, while Label Encoding was applied to the 'CURRENT_GRADE' column. These encoding methods are utilized to represent categorical variables as binary vectors and numerical columns, respectively, making them compatible with machine learning algorithms. At this stage also ensured that we had no missing data, and then trimmed the data for all future modeling to reduce outliers and irrelevant/ mal values.

Before training classification model we have to deal with class imbalance. To tackle this problem, I have used SMOTE Tomek

1. **Enhanced Balancing:** SMOTETomek combines SMOTE and Tomek links to generate synthetic minority class samples while eliminating ambiguous instances from the majority class, effectively balancing class distribution.
2. **Better Generalization**: SMOTETomek strategically interpolates synthetic samples between existing data points, improving the model's generalization without overfitting.
3. **Noise Reduction**: The use of Tomek links helps in identifying and removing outliers and noisy data points, leading to a cleaner and more robust dataset.
4. **Versatility**: SMOTETomek is applicable to a wide range of classification problems, particularly those with imbalanced datasets, making it a versatile and effective resampling technique.

Implement Catboost, XGboost classifier and for hyper parameter tuning implement GridSearch CV considering the hyperparameters for
Catboost : **iterations**=[500, 1000], **depth**= [6, 8, 10], learning_rate= [0.05, 0.1, 0.15] and auto_class_weights: [None, 'Balanced']

XGBoost : **max_depth**= [3, 6, 9], **learning_rate**= [0.01, 0.1, 0.2], **n_estimators**= [100, 300, 500]

RESULTS

Algorithm	F1 score on training data	F1 score on test data	ROC-AUC score on training data	ROC-AUC score on test data
Catboost	0. 8872	0.9034	0.9487	0.9108
XGBoost	0.8783	0.8978	0.9427	0.8993

Performance on NEW SAMPLE DATA

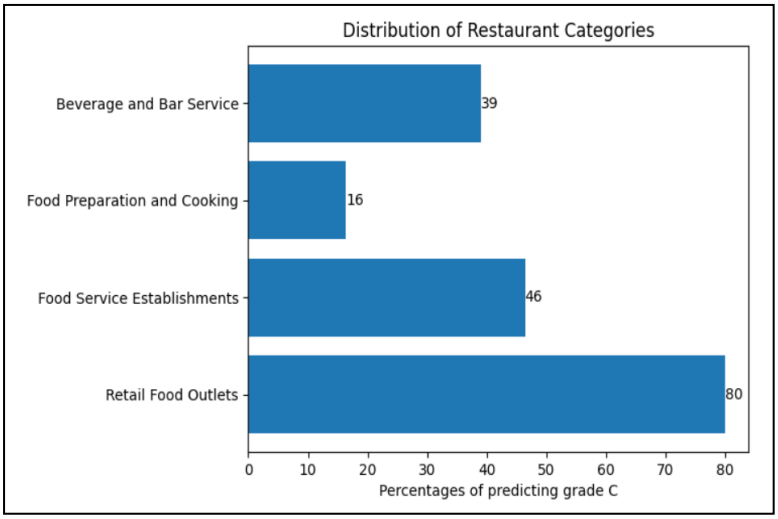

Figure 7 – Restaurant Inspection Grades: Percentage of Restaurants Predicted to Receive a Grade C, by Restaurant Category

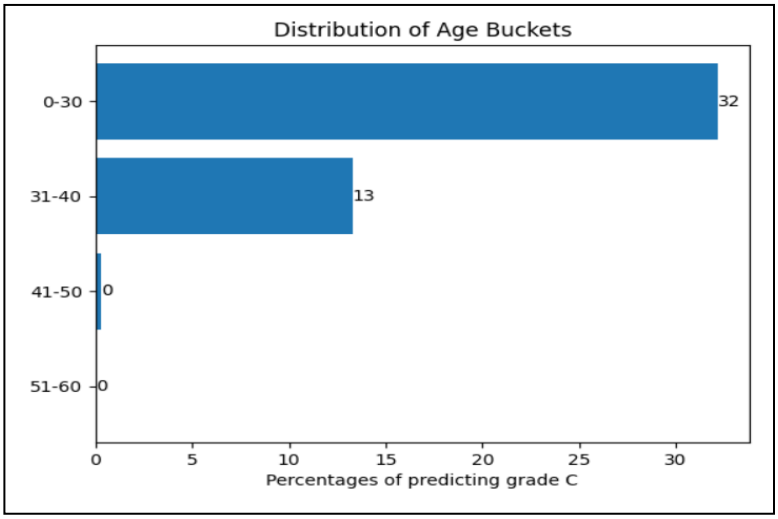

Figure 8 – Risk of next inspection grade C for restaurants by employee median age bucket

ACTION PLAN:

1:What types of insurance products can be purchased for restaurants? What is typically covered under these products? What are typical exclusions?

Insurance products for Restaurants

1. Businessowners Policy (BOP): Coverage: Damage to the restaurant's physical property.
2. Commercial Property Insurance: Coverage: Damage to the restaurant's property.
3. Business Liability Insurance: Coverage: Liability for bodily injury or property damage caused by employees, products, or operations.
4. Workers' Compensation Insurance: Coverage: Claims related to employee injuries.

5. Equipment Breakdown Insurance: Coverage: Losses due to equipment breakdown.
6 Flood and Earthquake Insurance: Coverage: Losses due to floods or earthquakes.
7. Management Liability Insurance: Coverage: Claims of wrongful termination, discrimination, or other employment-related torts.

Exclusions : Losses due to war, terrorism, nuclear accidents, mold, mildew, vermin infestation, and employee theft.

2: Suppose a restaurant employee is injured due to improper safety procedures. What types of insurance can provide restaurants with coverage for potential litigation?

Workers' compensation insurance: This is the most important type of insurance for restaurants to have, as it provides medical and wage benefits to employees who are injured or become ill at work. Workers' compensation insurance is required by law in most states.
Commercial general liability (CGL) insurance: This type of insurance protects restaurants from claims of bodily injury or property damage caused by their employees, their products, or their operations. CGL insurance can help restaurants pay for medical expenses, legal fees, and other costs if an employee is injured due to improper safety procedure.

CODE: