**Data Science & Business Intelligence Project**

**NYC Restaurant Inspections**

**Section: T1**

**Donald Akpovwa, Zack He, Ji Yoon Kim, Monica Sarphare, Janvi Shah**

**Business Understanding**

In a world haggled by division, our business problem focuses on something that we all share a common need for, food. Since the Covid-19 pandemic shook our world, all industries have been affected and many businesses, big and small, have been closed, especially in the restaurant industry. When you think of a place to find all the sweet delicacies of life, what place comes to mind, other than the legendary city that doesn’t sleep, New York City (NYC). The world is slowing coming back to a new norm, and with previously shutdown restaurants looking to open back their doors again, investors and entrepreneurs are increasingly looking for new restaurants to invest or establish in the great city. To help with these endeavors, Data science will play a crucial role in taking a deep dive into analyzing the historical pattern and relationship between the restaurants and the Department of Health (DOH).

We prepared a guide to what cuisines, zip codes, and types of restaurants have been cited or closed by the DOH. Our report analyzed how the cuisine type and location, among other factors, affect the DOH’s assessment of restaurants. Ultimately, we identified if being in a particular location, influences the DOH to close or negatively rate a particular cuisine more than others.

**Data Understanding & Preparation**

The data used to address our business problem is called “NYC Restaurant Inspection” and was retrieved from Kaggle at https://www.kaggle.com/new-york-city/nyc-inspections. The fields contained in the dataset are shown in Exhibit 1. The original dataset consisted of 399,918 rows (restaurants) and 18 columns (variables). After an initial analysis, it was obvious that our target variable was Critical Flag, indicating whether a restaurant was cited with a critical violation or not.

The business problem can be categorized as a classification problem: predicting whether a store would be flagged with a critical or non-critical violation. The original dataset has a lot of categorical data columns including Dates, Street, Phone, Building, etc., having hundreds of unique values. Upon converting all the columns in the original dataset into factors, there were around 450 columns generated. To make classification easier, we decided to trim the dataset by deleting columns such as Building, Street, Date, Violation Description, etc., because they were either repetitive or had too many unique values for us to work with. With that, we were down to 6 rows as seen in Exhibit 2.

Because we wanted to predict whether a restaurant would be cited with a critical flag or not, the restaurant’s Grade would be an important predictor. There were about 210,000 rows where the Grades were P, Z, Not Graded, or Blanks. Upon more research, we discovered that the grade assigned by the NYC inspections were based on the Score received. So, to make Grade more meaningful, we converted all the grades into A, B, C, or Z based on the score. A score of under 13 warranted A, under 28 warranted B, under 35 warranted C, and the rest were Z. Since the Grade column reflected the Score received, we deleted the Score column. The column of Cuisine had over 87 distinct values. Some were repetitive values, for e.g., pizza, Italian, and pasta. We reduced the number of unique values by using continents as Cuisine Origin instead of Cuisine.

We realized that there were a significant number of rows that were missing Zip codes, so we removed those for expediency. There were also several rows with blank actions. We removed those as well. Then were rows where there was a violation was recorded in the “Action” column, but no corresponding violation codes were present in the “Violation Code” column. So, we removed the rows where there were actions, but no violations recorded. However, we wanted to preserve rows where the violations in “Actions” were “No violations recorded,” so we added another code “AAA” to indicate that even though the “Violation Code” column was blank, the blank was meaningful, in that it indicated a positive scenario where no violations were recorded. This preprocessing left us with a final clean dataset of 6 columns and over 399,000 rows. Using this cleaned dataset, we performed attempted to understand the data using visualizations as described below.

*Data Format Cleaning During Modelling*

During modelling, we saw that many of the variables are not generating reasonable output, thus, we used the “mutate” function to change the types of these variables as a factor (**Exhibit 3**). Since the target variable, CRITICAL.FLAG has three values, we also transformed it into a simple “yes” or “no” question, where if there was a critical flag, we used “yes,” and no if there was non-critical or not applicable. (**Exhibit 4**).

**Visualizations**

*Cuisine Origin by Borough*

In our first section, we wanted to understand the landscape of cuisine type by borough to see if certain cuisines were predominant in certain parts of NYC. Based on the bar chart in **Exhibit 5**, we concluded that North American was the predominant cuisine type in 3 of out 5 boroughs: Manhattan (35%), Brooklyn (27%), and Staten Island (32%). In the Bronx, there was a higher percentage of Central American cuisine (~17%) compared to all other boroughs. In Queens, there was a larger percentage of East Asian cuisine (25%) compared to North American (21%) and all other cuisine types. In each borough, except for the Bronx and Staten Island, the proportion of European cuisine was commensurate with North American cuisine.

*Count of Critical Flag vs. Grade*

Our team wanted to understand how the type of critical flag (critical, not critical, not applicable) was distributed within each grade (A, B, C, Z). We graphed each grade on the x-axis and the count of critical flag on the y-axis. Based on the bar chart in **Exhibit 6**, it is evident that each grade is divided evenly between critical and non-critical flags. For example, of all the restaurants that received a grade of A, ~50% were flagged as critical and the other ~50% were flagged as non-critical. This was an interesting observation as we expected the proportion of critical flags to be much lower for higher grades like A and B.

*Count of Critical Flag vs. Cuisine*

We also wanted to explore how cuisine type affected critical flag. We expected western cuisines such as those from North American and Europe to have a lower proportion of non-critical flags to critical flags, and we expected ethnic cuisines such as those from Asia and Central America to have a higher proportion of critical flags to non-critical flags. Based on the bar chart in **Exhibit 7**, it is evident that across all cuisines, the proportion of critical flags slightly outweighs the proportion of non-critical flags in a 55/45 ratio. This indicated to us that the type of cuisine did not have a significant impact on flag type.

*Count of Critical Flag vs. Borough*

Since cuisine type did not seem to be a strong predictor of critical flag, we wanted to test whether the borough within which a restaurant was located had a significant effect on flag type. In other words, if a specific borough received more critical flags than others. We graphed each borough (boro) on the x-axis and critical flag (critical, not critical, not applicable) on the y-axis. Given the sheer size and density of the borough of Manhattan, we expected it to have a higher proportion of critical to non-critical flags. Based on the bar chart in **Exhibit 8**, however, it is evident that within each borough, there a relatively even proportion of critical to non-critical flags. Critical flags slightly outnumber non-critical flags. Based on these results, we concluded that the type of cuisine did not have a significant effect on whether a restaurant received a critical flag vs a non-critical flag.

*Cuisine vs. Grade*

Setting aside our critical flag target variable, we wanted to observe patterns between type of cuisine and grade. Similar to our logic when comparing cuisine to critical flag, we hypothesized that North American and other western cuisines would receive a higher proportion of favorable grades compared to non-western cuisines. Based on the results from our bar chart in **Exhibit 9**, it is evident that each cuisine received roughly the proportion of grades A, B, C, and Z. That is, each cuisine type received about 50% As, 30% Bs, 16% Cs, and 6% Zs. This indicated to us that there was no notable pattern among cuisines type and grade.

*Borough vs. Grade*

We also sought to observe patterns between borough and grade. Specifically, whether certain boroughs were more likely to receive higher grades than others. We graphed each borough on the x-axis and the count of each grade on the y-axis. Based on the bar chart in **Exhibit 10**, each borough had a relatively even proportion of each grade: 45% As, 32% Bs, 17% Cs, and 5% Zs. Based on these results, it was clear that there was no distinct pattern between borough and grade. Each borough was equally as likely to receive a particular grade.

**Modeling**

*Linear Classifier – SVM*

We used four variables for the Support Vector Machine (SVM) model, including BORO, GRADE, ACTION, and VIOLATION. Before we conduct the train and test split, we included all the dataset to check how the model performs on the entire dataset. As a result, with 100 percent accuracy, we were able to find that 220,082 entries were flagged critical and 172,857 entries were flagged not critical.

After this initial step, we performed train and test split with 80% of the data for training and 20% for testing. In addition to the train and test split, we applied cost to tune the SVM model. To minimize the misclassifications, a cost of 0.1 was applied to the model. As it can be seen in **Exhibit 11**, out of 78,587 test entries, we got 34,578 true negative and 44,005 true positive meaning that only 4 entries are misclassified. By comparing the predicted value with the actual value, we found the SVM model has approximately 99.999% accuracy.

Through the SVM model, we found out that ACTION, VIOLATON, and GRADE are positively correlated with each other, and do not add much of variance to the model. Since this indicates that actual data is close to where the model predicted, in other models, we decided to include only significant variables among ACTION, VIOLATION, and GRADE.

*Classification Tree Model*

As mentioned in the Business Understanding part, the main problem we are focusing on solving is, what types of restaurants have a higher chance of being flagged critical in the health code violation inspection. Therefore, naturally, it is a classification problem with techniques used in supervised learning.

The first step was to build a model using the entire dataset just to get a basic understanding of how the model would perform. To answer our business problem, we included three variables: BORO, CUISINE.ORIGIN, and GRADE. We set cp value to be 0 for a bigger tree and limit the depth to be maxed at 5 for better presentation. We can tell from the final plot (**Exhibit 13**) that BORO and GRADE are, comparatively speaking, more important variables in this specific model. About 69% of the entries in the entire dataset were flagged critical, and there are three ways to predict if a certain restaurant would be safe from the inspection.

1. They were given a grade of Z
2. They were given a grade of A, AND they are specialized in Africa, American, Middle East, and Other cuisine, but 47% of those restaurants were flagged critical
3. They were given a grade of A, AND they are specialized in Australia, Central America, and North America cuisine, but 49% of those restaurants were flagged critical

Next, to further examine the model, we decided to split the dataset into training and testing data subsets, with 20% of them being the training and the rest being testing. We then ran the model on the training dataset and the model can be seen in **Exhibit 14**. This time, about 63% of the restaurants were marked critical and BORO became one of the deciding factors. There are three ways to predict if a certain restaurant would be safe from the inspection as well:

1. They were given a grade of Z
2. They were given a grade of A AND they are specialized in Africa, American, Middle East, North America, and Other cuisine, but 48% of those restaurants were flagged as critical
3. They were given a grade of A, AND they are specialized in Central America, Europe, and South America Cuisine, AND they are in Bronx, Brooklyn, or Queens, but 50% of them were flagged as critical

In both models, GRADE is at the root of the tree, which makes it the most important variable. After which, we pruned the tree using the cp value with the minimum xerror and applied this model to the test dataset to get the predicted probabilities. Using a 60% cut-off, we generated class prediction because it is the cut-off point where we got the lowest error rate, which was at around 38.6%. We then created a confusion table and ROC curve of this model, and the AUC was 0.636 (**Exhibit 15 and Exhibit 16**).

*Logistic Regression*

For ease of use, we renamed all the columns “zipcode,” “boro,” “cuisine\_origin,” “action,” “violation,” “critical\_flag,” and “grade.” For logistic regression, since these are all categorical variables, we converted them to factors using the mutate function. Additionally, we also created a column called “flag” that indicated whether the “critical\_flag” was critical or not with a “yes” to indicate critical and “no” to indicate all other values. Then, we ran logistic regression on all the data without the train/test split using predictors “zipcode,” “boro,” “cuisine\_origin,” and “grade.” The accuracy of the first logistic regression was 61.43% (**Exhibit 17**) using a cutoff of 0.5 with the important variables being “zipcode,” “boroBROOKLYN,” “boroQueens,” and “gradeB,” and “gradeC.” The AUC is 0.642 as seen in **Exhibit 18**.

Next, to further examine the model, we decided to perform a train/test split with 80% of data being training data and the rest 20% data being testing data. Then we ran the model on the training data using the same predictors as the ones before: Based on the results of the “zipcode,” “boro,” “cuisine\_origin,” and “grade.” We received similar results, the accuracy of the logistic regression model on the test dataset was 61.32% (**Exhibit 19**) using a cutoff of 0.5 with the important variables being only the “grade” as “gradeB,” and “gradeC.” The AUC is 0.641 as seen in **Exhibit 20**. Based on these models, we can determine that “gradeB” and “gradeC” are likely the most important predictors of whether a restaurant will receive a critical flag.

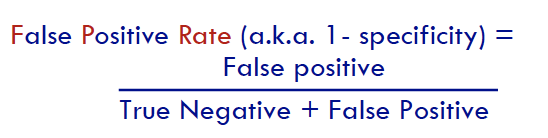
**Evaluation**

*Linear Classifier – SVM*

The SVM model automatically generated a K-fold validation score, and we observed that an error estimate of SVM using 10-fold cross validation was 3.181157e-5 (0.00003181157). The error rate of the SVM model was very small.

In addition, the false positive rate of the SVM model was around 1.15667e-4 (0.0000115667). As it can be seen in **Exhibit 12**, the AUC was 1.000.

*Classification Tree Model*

Even after pruning, the classification tree model has an error rate of around 38.6%, which is not ideal to start with. Furthermore, we constructed a confusion matrix (**Exhibit 15**) for a more detailed analysis. Using the formula:  
 

We observed a false positive rate of around 37.2% from the confusion matrix (**Exhibit 15**), as a result, the AUC was low as well at 0.636.

*Logistic Regression*

The accuracy of the first logistic regression was 61.43% (**Exhibit 17)** using a cutoff of 0.5 with the important variables being “zipcode,” “boroBROOKLYN,” “boroQueens,” and “gradeB,” and “gradeC.” The AUC is 0.642 as seen in **Exhibit 18**.

We received similar results, the accuracy of the ;ogistic regression model on the test dataset was 61.32% (**Exhibit 19**) using a cutoff of 0.5 with the important variables being only the “grade” as “gradeB,” and “gradeC.” The AUC is 0.641 as seen in **Exhibit 20**. Based on these models, we can determine that “gradeB” and “gradeC” are likely the most important predictors of whether a restaurant will receive a critical flag but they are not very predictive of the flag at 61% accuracy or 39% error rate.

**Deployment**

Based on the models, both logistic regression and classification model, indicate that grade is the most significant predictor of whether a restaurant will be cited for a critical or non-critical violation. Furthermore, since the accuracy of both the models is around 61% and error rate is around 39%, neither model is likely reliable to predict whether a restaurant will be cited for a critical or non-critical flag.

In our business problem, we aimed to analyze whether cuisine type or restaurant location had any impact on whether a restaurant was cited for a critical or non-critical violation. However, based on the models, we find that it is unlikely that a restaurant’s cuisine or location impacts whether a restaurant was cited for a critical or non-critical violation. This conclusion is also supported by visualizations earlier in the paper that indicate a proportional distribution of critical vs non-critical flags for cuisine and location (zipcode).

**Appendix**

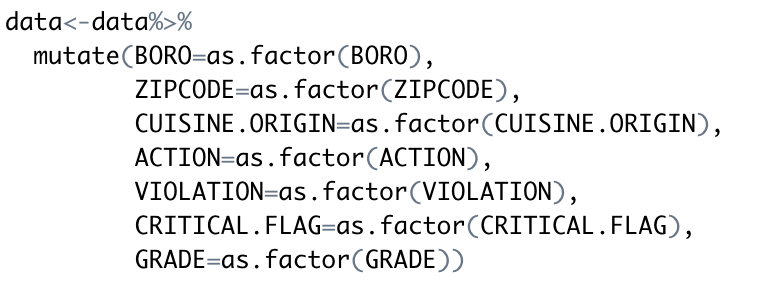
**Exhibit 1.** Original Dataset Columns

|  |  |  |
| --- | --- | --- |
| **Data Field Name** | **Data Type** | **Description** |
| CAMIS | Varchar | This is a unique identifier for the entity (restaurant) |
| DBA | varchar | This field represents the name (doing business as) of the entity (restaurant) |
| BORO | Varchar | Borough in which the entity (restaurant) is located. NOTE: There may be discrepancies between zip code and listed boro due to differences in an establishment's mailing address and physical location |
| BUILDING | Varchar | This field represents the building number for the entity (restaurant) |
| STREET | Varchar | This field represents the street name at which the entity (restaurant) is located. |
| ZIPCODE | Varchar | Zip code as per the address of the entity (restaurant) |
| PHONE | Varchar | Phone number |
| CUISINE DESCRIPTION | Varchar | This field describes the entity (restaurant) cuisine. |
| INSPECTION DATE | Datetime | This field represents the date of inspection. NOTE: Inspection dates of 1/1/1900 mean an establishment has not yet had an inspection |
| ACTION | Varchar | This field represents the action that is associated with each restaurant inspection. |
| VIOLATION CODE | Varchar | This field represents each violation associated with a restaurant inspection. |
| VIOLATION DESCRIPTION | Varchar | This field describes the violation codes |
| CRITICAL FLAG | Varchar | Critical violations are those most likely to contribute to foodborne illness. |
| SCORE | Varchar | Total score for a particular inspection; updated based on adjudication results. |
| GRADE | Varchar | This field represents the grade associated with this inspection. Grades given during a reopening inspection are derived from the previous re-inspection. |
| GRADE DATE | Datetime | The date when the grade was issued to the entity (restaurant) |
| RECORD DATE | Datetime | The date when the webextract was run to produce this data set |

**Exhibit 2.** Final/Trimmed Dataset Columns

|  |  |  |
| --- | --- | --- |
| BORO | Varchar | Borough in which the entity (restaurant) is located. NOTE: There may be discrepancies between zip code and listed boro due to differences in an establishment's mailing address and physical location |
| ZIPCODE | Varchar | Zip code as per the address of the entity (restaurant) |
| CUISINE ORIGIN | Varchar | This field describes the entity (restaurant) cuisine’s origin by continent. |
| ACTION | Varchar | This field represents the action that is associated with each restaurant inspection. |
| CRITICAL FLAG | Varchar | Critical violations are those most likely to contribute to foodborne illness. |
| GRADE | Varchar | This field represents the grade associated with this inspection. Grades given during a reopening inspection are derived from the previous re-inspection or scores. |

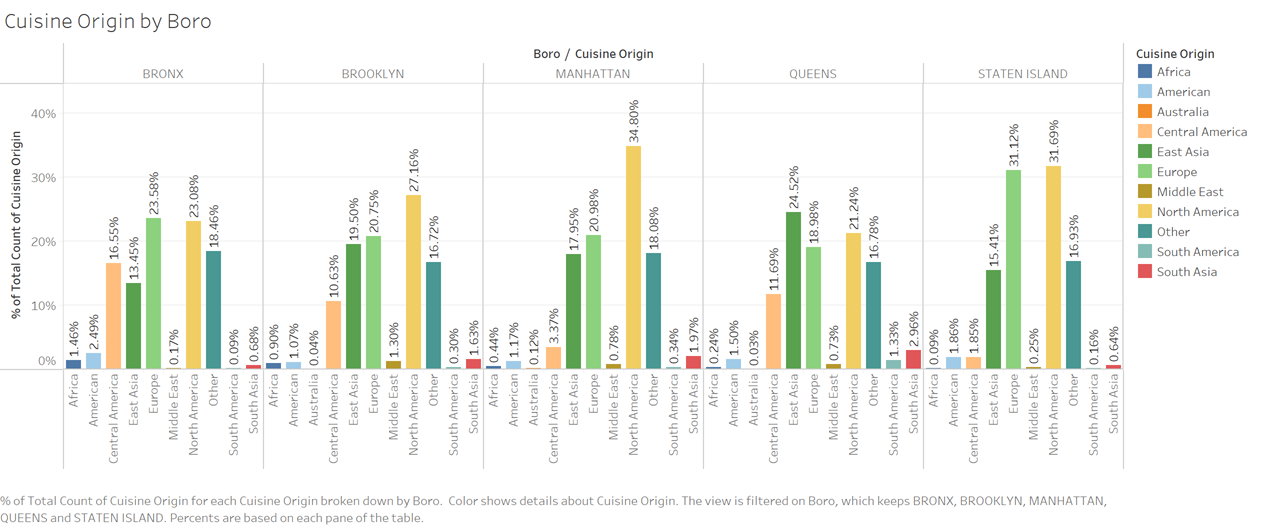
**Exhibit 3.** Variable Mutation



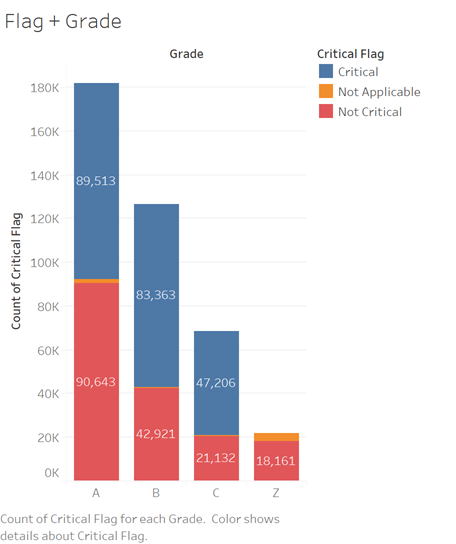
**Exhibit 4.** Critical. Filtered Mutation



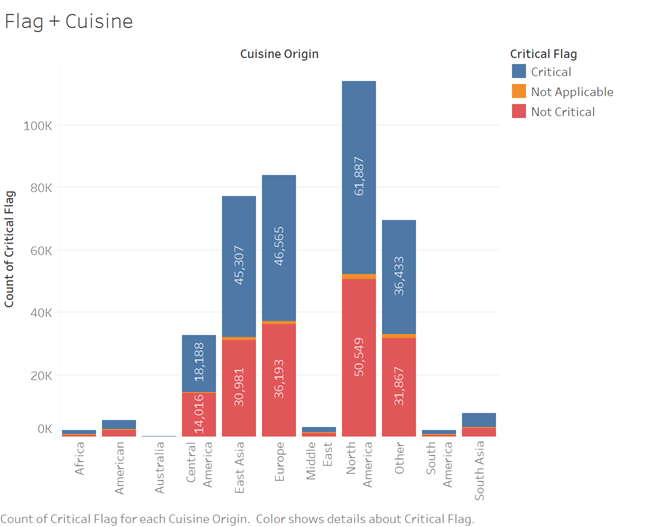
**Exhibit 5.** Cuisine Origin by Borough



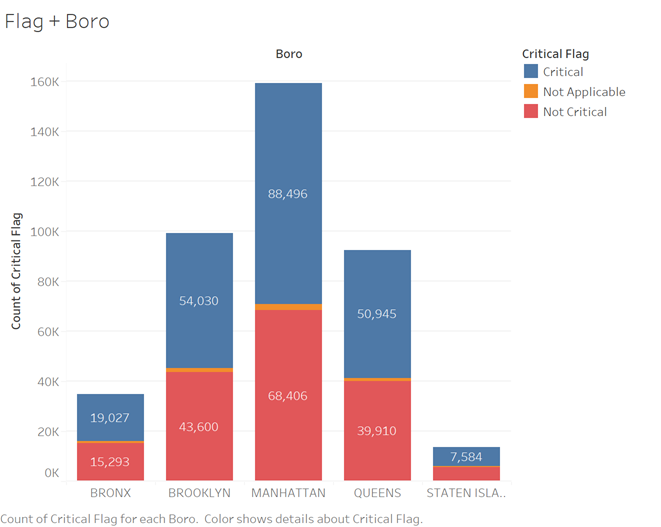
**Exhibit 6.** Count of Critical Flag vs. Grade



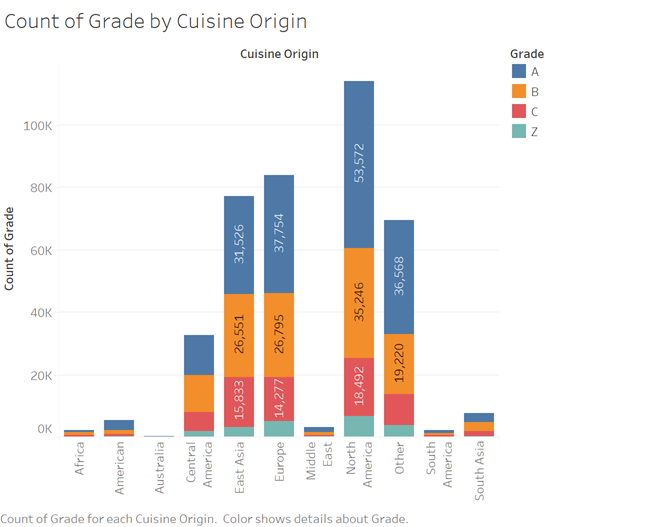
**Exhibit 7.** Count of Critical Flag vs. Cuisine



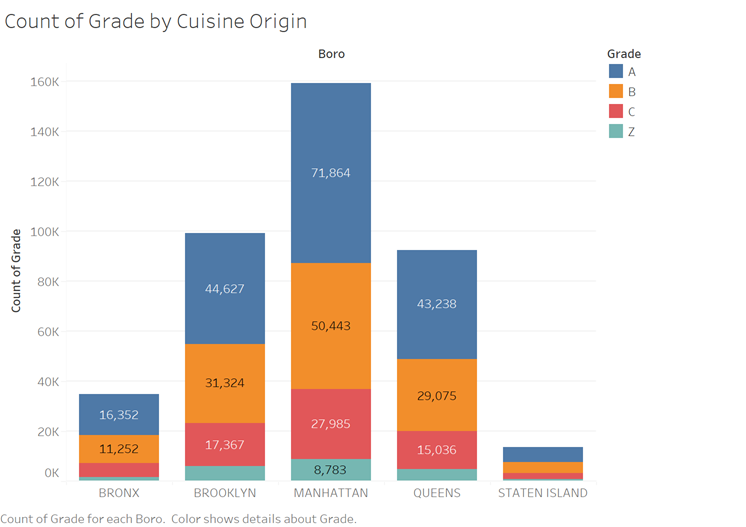
**Exhibit 8.** Count of Critical Flag vs. Borough



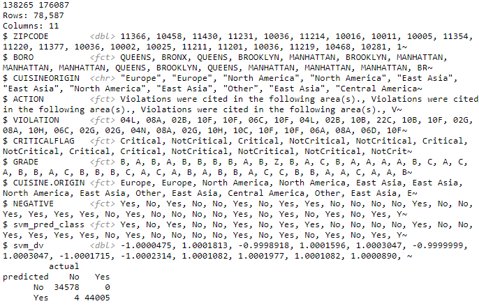
**Exhibit 9.** Cuisine vs. Grade



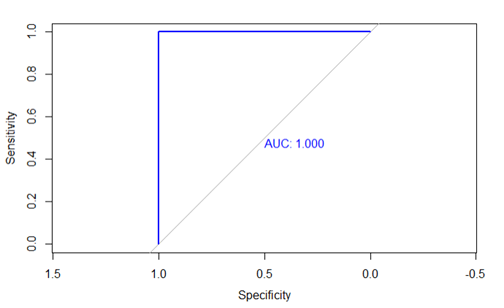
**Exhibit 10.** Borough vs. Grade



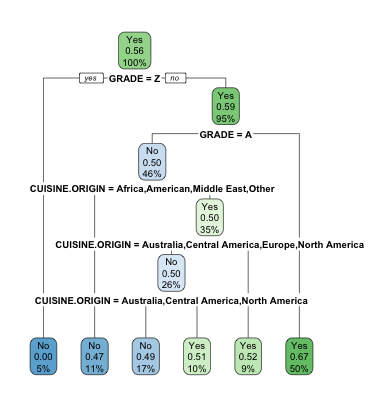
**Exhibit 11.** Linear Classifier – SVM Confusion Matrix



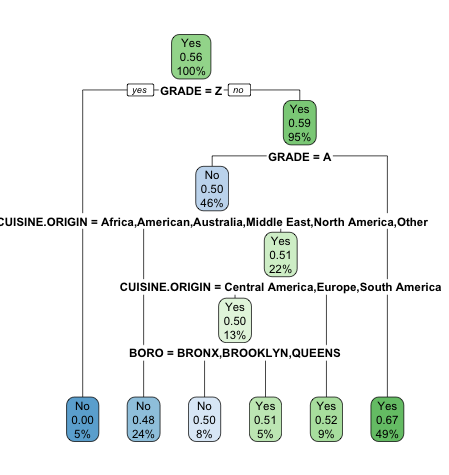
**Exhibit 12.** Linear Classifier – SVM ROC/AUC



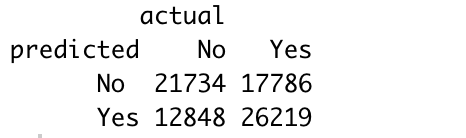
**Exhibit 13.** General Classification Tree



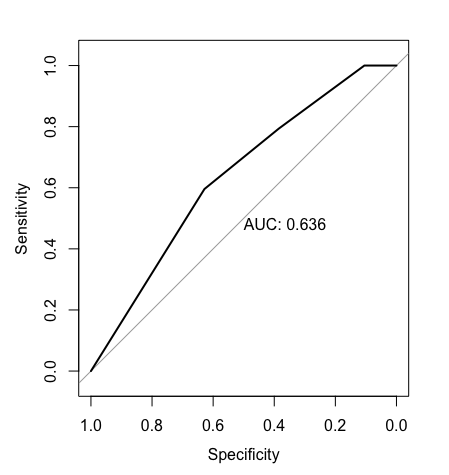
**Exhibit 14.** Split Classification Tree



**Exhibit 15.** CT Confusion Table



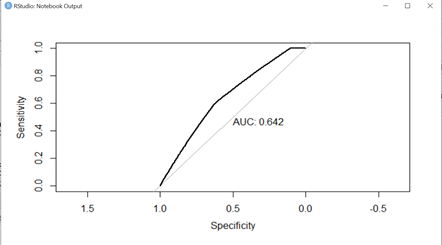
**Exhibit 16.** CT ROC/AUC



**Exhibit 17.** Confusion Matrix for Logistic Regression on Entire Dataset



**Exhibit 18.** ROC/AUC for Logistic Regression on Entire Dataset



**Exhibit 19.** Confusion Matrix for Logistic Regression on Test Dataset



**Exhibit 20.** ROC/AUC for Logistic Regression on Test Dataset

