Machine Learning Component

1. Using Logistic Regression technique to predict a restaurant's health violation based on other factors

We used the following factors to predict whether a restaurant will be flagged for a health violation. The "CRITICAL_FLAG" field from the violations dataset was used to determine whether a restaurant had a health violation. The other factors used were Burough, Cuisine Description, Community board, Council District, Census tract, Burough's Median Income.

Since the task is a binary classification problem, to predict whether a restaurant had a health violation or not, logistic regression seemed to be the most suitable. While there were other models that could've also been used such as Random Forests and Support vector Machines (SVMs), we decided to go with logistic regression as we believed it was adequate enough to capture underlying patterns which connected neighborhood affluence, cuisine type, local socio-political and governmental structures (represented by community board, council district, census tract) with the likelihood of a health violation.

We had to clean and restructure the data. We first had to convert the "CRITICAL_FLAG" field into a binary field, as previously it contained two unique text values. We then merged the health violations dataset with our median income dataset using the Burough name as the field to perform the inner join on. We dropped the second borough column that came with the join. We then had to one-hot-encode the following columns: Borough, Cuisine Description, Community Board, Council District, Census tract. This is because they were all categorical variables.

We used a logistic regression model with a max of 1000 iterations and a C value of 0.1, as We was using Sklearn's LogisticRegression Model. The low C value means We used strong regularization. This is because after experimenting with a couple of different values, We found this to be a value that resulted in balanced metrics. We used K-fold cross validation with 10 splits and made sure to shuffle the dataset for it.

For metrics, we used 4:

a) Accuracy - representing the portion of correct predictions out of the total number of predictions
b) Precision - This is the True positive divided by the sum of the True Positive and False Positive. It is the portion of correct positive predictions over all positive predictions. High precision means positive predictions are likely to be correct
c) Recall - This is the True Positive over (True Positive + False Negative). Portion of correct positive predictions over all actual positives. High recall means the model correctly predicts positives for most actual positives.
d) F1 score - This balances Precision and Recall. High F1 means the model does a good job at avoiding false positives and false negatives.

These were the following results of the metrics for our Logistic Regression Model:

K-Fold Validation Results:
Accuracy: 0.5466 ± 0.0246
Precision: 0.5630 ± 0.0264
Recall: 0.8042 ± 0.0347
F1: 0.6619 ± 0.0249

As you can see our accuracy was fairly low, around 55%, which is only slightly better than randomly guessing. We thought that our feature selection was good, and that they held strong relationships that our model could learn and thus use to predect with some good accuracy the chance of a restaurant having a health violation. Our Precision was also pretty low meaning a positive prediction was not likely to be a true positive. Our recall was high suggesting that our positive predictions overlapped a good amount with the actual positives in the dataset. This implies a high false positive rate. Our F1 score is around 67% which suggests we do an okay job in reducing our false positivity and negativity rates, but they are still higher than we would like.

We potentially got these results because features we thought were correlated with health violations aren't actually or rather are introducing noise to the model. This can be for example from the community board, council district and census tract fields we added as features. This was done thinking they were some reflection of local governance and socio economic structure that might be connected with health laws and restaurant safety checks. Perhaps there are other features which would be important in this such as rating or price category that might capture public perception and economic standing of the restaurant. We might have gotten these results because the underlying relationships aren't as strong as we expected and we need different data to improve our predictive prowess.

The results did not correspond with our initial belief that the median income and cuisine type were connected with the likelihood of a restaurant having a health violation. This might be because median family income of the neighboourhood is not a good proxy for the financial standing of the restaurant and its ability to maintain a safe and clean cooking environment. It might also be that we are unable to capture local, community level governance and health perceptions that probably factor in the upkeep and standards of running a restaurant.

We think logistic regression was an appropriate model to use, but perhaps we could instead use RandomForest, GradientBoosting, to see whether the logistic regression model is at fault for not being complex enough to capture the underlying patterns that exist in the data.

The data was adequate for our analysis a s it involved the main features we thought would be responsible for health violations. However, it might be the case that our data was not sufficient to enable high predictive prowess and thus a better selection might be needed after all to achieve the task of successful prediction of health violations.

2. Using KNN technique to predict a restaurant's rating based on other factors.

1. Why did you use this statistical test or ML algorithm? Which other tests did you consider or evaluate?

Since we were trying to predict a rating which is a continuous value, we thought that KNN would be the most appropriate model for three main reasons. The first being is the non parametric nature of KNN, which means that there are no fundamental assumptions about the data set. This is suitable, especially in our case, since there are a multitude of variables/ features that we are using for the regression task which obviously do not conform to underlying assumptions like linearity, normality, or independence. Secondly, KNN was a good choice because of the flexibility it grants us since we can tune the hyper-parameter K according to our needs. This was especially useful as we noticed that lower values for K hurt our accuracy. We then pivoted bigger values (eventually settling on K=10), which helped us reach our eventual accuracy. Lastly, KNN is a favorable choice because it is robust to a diverse range of variables that might have completely different relationships to the target variable rating. For example, price category might have a more simplistic linear relationship with rating, while the feature median income might indicate the spending power of the local population, which could correlate with the type of restaurants (e.g., more upscale restaurants in wealthier areas) and thus their ratings. Furthermore, the geography of the restaurant (features like zipcode, borough, etc) may have an extremely complex relationship with rating.

What metric(s) did you use to measure success or failure, and why did you use it?

a) Accuracy - The portion of correct predictions out of total predictions. Used because it gives us a straightforward measure of the model's correctness.
b) Cross-Validation Scores - This is the average score from multiple training and testing on different subsets of the data, in our case we used 10-fold meaning our data was split into 10 subsets with 1 being used for testing and the others for training. This gives us a measure of the model's generalizability and reduces overfitting by averaging scores over different subsets.

What challenges did you face evaluating the model? Did you have to clean or restructure your data?

One major deliberation and challenge we faced was deciding how to handle the prediction of restaurant ratings. The ratings range from 0 to 5 and are discrete integer values. Initially, we faced a dilemma in categorizing these ratings either as categorical or ordinal data.

> Categorical Approach: We initially thought of just treating each discrete value as a separate category. This allows us to use KNeighborsClassifier (instead of the regression version of the function). Furthermore, we were able to apply classification metrics such as precision, recall, F1-score, and accuracy to evaluate our model as previously

mentioned. These metrics helped us understand not only the overall effectiveness of the model but also how well it predicted each specific rating.

Ordinal Approach: However, we also recognized that ratings have a natural order—the difference between a rating of 4 and 5 should be less significant than between 4 and 1. This understanding led us to consider ordinal regression techniques, which could potentially leverage the ordered nature of our target variable to improve model performance. By treating the problem as a regression or an ordered classification, we hoped to capture the inherent order in the ratings, offering a more nuanced understanding and potentially more accurate predictions.

We didn't arrive at a decision immediately, as there were definitely benefits to both methods. Treating ratings as categorical simplifies the problem but ignores the inherent order between ratings. On the other hand, treating them as ordinal introduces complexity into the model but respects the natural ranking of the data.

After extensive testing and evaluation, we chose to proceed with the categorical approach using KNN classification. The categorical approach is also fundamentally superior in this situation because these types of models do not assume any specific mathematical relationship (like a linear or ordinal progression) between the feature set and the target variable. For example, we don't want the model to learn an ordinal progression/ linear regression between rating and geographical location, as the relationship is much more complicated than that. This is especially important since we also have relatively sparse.

This challenge reminds us of always understanding the nature of your data and how that not only informs the model we choose but the representation of our data.

What is your interpretation of the results?

Interpretation of Results:

The results from our KNN classification model are both encouraging and indicative of robust predictive performance. The cross-validation scores consistently hovered around 0.8, with a mean cross-validation score of 0.79. This consistency across different subsets of the data suggests that our model is stable and reliable, showing little variance in performance due to random fluctuations in the training data. The accuracy on the test set reached 0.85, which is notably higher than the mean cross-validation score, further underscoring the model's effectiveness in handling new, unseen data.

Evaluation of Hypothesis and Prediction Accuracy:

With these results, we can confidently claim that our model meets expectations. The hypothesis that our categorical approach using KNN would adequately capture complex relationships in our data without needing to impose a specific mathematical relationship between features and the target variable is supported by the data. In other words, we were right in our hypothesis that not assuming any specific mathematical relationship between the feature set and the target variable would result in a meaningful model. The high accuracy rate, particularly in the test set, validates our choice of model and approach, suggesting that treating ratings as distinct categories effectively captures the nuances necessary for accurate predictions.

Intuitively, the results are highly satisfactory. Achieving an accuracy of 0.85 on the test set is particularly commendable and exceeds typical benchmarks for models in similar contexts. This success not only demonstrates the model's capability to generalize well but also reassures us of its practical applicability. The consistency in cross-validation scores further bolsters our confidence, indicating that the model is not overfitted to the peculiarities of the training data but is genuinely capturing underlying patterns that are predictive of the outcomes.

Given the complexity of the features involved and the potential challenges of sparse data, the results are not just good—they are exceptional. The robustness indicated by cross-validation and the high accuracy on the test set make us confident in the model’s reliability and in our methodological choices.

Results:

```
Cross-Validation Scores: [0.8        0.8        0.8        0.76666667 0.76666667 0.76666667
 0.8        0.8        0.8        0.8       ]
Mean Cross-Validation Score: 0.79
Accuracy on Test Set: 0.85
```

3. Did you find the results corresponded with your initial belief in the data? If yes/no, why do you think this was the case?

Yes, our results show a high accuracy of around 85% on the testing dataset. This shows that the features we used were relevant to determining the rating a restaurant is given. The price category which represents how expensive the food is and the median income of the neighborhood it is in both contribute to determining the rating people give it. We know people’s perception of quality is often inextricably linked with cost and we suspected the affluence of a neighborhood would have an impact on the rating.

4. Do you believe the tools for analysis that you chose were appropriate? If yes/no, why or what method could have been used?

Yes, we think using K nearest neighbors was appropriate as we were trying to find groups that were similar based on the input features and then we could assign new data to the rating category of the most similar group. We cluster the restaurants and ideally the clustering will be representative of how ratings are assigned to these restaurants. Based on the results, it would seem that this is what happened. Alternatively, we could've also used multi-class regression or ensemble learning to learn and predict the ratings.

5. Was the data adequate for your analysis? If not, what aspects of the data were problematic and how could you have remedied that?

The data was not adequate for our analysis. The dataset containing the ratings after cleaning it so that all the necessary fields had values was around 300 rows, so our dataset was pretty small in that sense. This is also because the ratings dataset only had zipcode data and we had to join it on our dataset of neighborhood to median income, so in finding a mapping from zip code to borough, we realized that a good proportion of the data corresponded to restaurants in new jersey and other non NYC areas, and since we only had the mapping for NYC areas, we had to drop those data points.

It was also hard to join this ratings dataset on the violations one as there was not so much overlap and simply joining on name was not good enough as there were variations in how names were written as well as addresses that made this impossible. This meant we could not utilize information we suspected would contribute to ratings such as whether the restaurant had a health violation, Cuisine description, or health grade. We also had limited category representation as far as ratings are concerned. After we changed it from a continuous variable to 5 discrete buckets by rounding the ratings to the nearest integer, we were left with a dataset that only contained ratings 3,4,5. With 4 being the most dominant. This definitely adds a big sample bias to our model as it might simply be unable to predict for ratings 1 or 2 because it has never seen it in training. We could've remedied this by scraping our own data from google, instead of relying on the pre-scraped dataset we used for ratings. This might've allowed us to collect data on much more restaurants and thus have a larger cleaned dataset to train and test our model on.

<u>Hypothesis Testing Component:</u>

For this portion of the analysis component, we had many ideas. We wanted to test how price, ratings, neighborhoods, median incomes, quality of food, nature of inspections and ethnicity of food related with each other in the NYC boroughs. We furthermore wanted to not use the same hypothesis test statistic for all of our hypothesis. We explored many kinds of test statistics including but not limited to Z-statistic, T-statistic, F-statistic, Chi-Square statistic (χ^2), correlation coefficient (r), regression coefficients. After a lot of trial and error and seeing what we wanted to learn out about more and what metrics our data was better prepared to answer, we decided to investigate into the following hypothesis below.

For our first hypothesis we wanted to know if there was some sort of correlation between the ratings garnered by restaurants online and the price they charged for providing their food and beverage services. Our initial hypothesis was that more expensive restaurants might end up subconsciously getting more higher ratings as people are primed to believe that greater quality comes at a greater price. While this may not necessarily be the case for something like food, we wanted to try and see if we could get some confirmation for this story from the data. We used the single sample t-test similar to the one we used in the section / lab to come to this conclusion. We found that there was no real correlation between price and ratings. The average ratings for price category 4 restaurants was similar to that of the entire population. We thought a a simple t-test would be sufficient for this analysis because it was clear and concise and answers the question to the point. It was also very easy to use as our dataframe was primed to have such a test performed on its data. This question was also important because it could be used to answer some considerations that may come up in our ML models as well. Our interpretation of the results is that in order to have a great food experience price is not really a defining factor in the city of new york. We agree with the hypothesis and are convinced by what it says. Intuitively it wss kind of comforting as we all are moving to new york and it helps us feel better about not having to burn a hole in our wallets to eat good food. The results did not really correspond with our initial belief because we felt the test would be significant. We thought there might be atleast a slight correlation between price and ratings. We believe that the tools chosen were appropriate and that we had enough data to perform our stats testing.

Our second hypothesis was to try and see if there was a relationship between ethnic restaurants and the number of critical violations. The reason we wanted to test this is because there are a lot of narratives that exist that say bad things about ethnic restaurants and immigrant restauranteers which we wanted to disprove. We feel like online there is a lot of xenophobia and one the things the say is about the cleanliness standards of these restaurants. We used the proportional one sample z test. We had to use this test because one of the variables (CRITICAL FLAG) is binary and we cannot calculate the variance of this. Because we can’t get the variance this was the ideal test to use. We didn’t have to clean or restructure the data to use the test. We were very happy with the interpretation of our result because it helped prove our point which is that these ethnic restaurants are not more likely to be in violation of health codes. We like this result and was what we believed in initially. I think the tools chosen were appropriate and the method was apt. The data was adequate for the analysis as well and the data did not need to be remedied.

Our last hypothesis was the most interesting one. We looked at so many different types of tests to try and model the last aspect we wanted to take a look into. We wanted to see if there was a correlation between the boroughs and the number of critical violations. The broad idea was that different borough have different median incomes and thereby levels of affluence. As a result it may not be highly outlandish to make an inquiry into this hypothesis. We found that since it was a relationship between two categorical variables the chi square test for independence was ideal. We did not have to change any aspect of our dataframe to use this and the in built functions plugged right in. The results came back and told us that there was no correlation between borough and violations and there was indeed independence between them. We like this result and it goes to show that cleanliness and quality of food are not affected by price, and other peripheral stuff but mainly about how much the owners of the restaurant care. The tools we used were appropriate for this inquiry and the data was adequate as well.