# Adam Hecktman and Devindra “Dave” Sawh

DSC 478 Final Project for Winter 2020

Prediction Modelling for “Consumer Services Complaints for the City of New York”

Url to original data set : <https://data.cityofnewyork.us/Business/Consumer-Services-Mediated-Complaints/nre2-6m2s>

URL to video presentation: <https://vimeo.com/398316334>

The data includes about 2600 hundred records and this may cause us to lean towards a cross validation approach but will still try other methods for training.

Attributes include the following:

|  |  |  |
| --- | --- | --- |
| **Business Name** | Name of the business against which a complaint has been made. |  |
| **Industry** | The business category of the business against which a complaint has been made. |  |
| **Complaint Type** | Indicates the type of complaint made. |  |
| **Mediation Start Date** | Date mediation started. During the mediation process, DCA sends a copy of complaint to the business for written response. Then, mostly over the phone, a DCA mediator speaks with both consumer and the business to reach an agreement and settle the matter. |  |
| **Mediation Close Date** | Date mediation ended. |  |
| **Complaint Result** | Outcome of mediation efforts. See Appendix A for further details about complaint results. |  |
| **Satisfaction** | This section indicates whether the complaint was mediated to the satisfaction of both the business and consumer. See Appendix A for Yes, No, and NA assignments. |  |
| **Restitution** | Total amount of consumer restitution secured through mediation. |  |
| **Business Building** | The building number of the business’s address. |  |
| **Business Street** | The street name of the business’s address. |  |
| **Building Address Unit** | The unit number of the business’s address (e.g., Apartment/Suite/Other). |  |
| **Business City** | The city where the business is located. |  |
| **Business State** | The state where the business is located. |  |
| **Business Zip** | The zip code where the business is located. |  |
| **Complainant Zip** | The zip code where the individual who filed the complaint is located. |  |
| **Longitude** | Geo coordinates |  |
| **Latitude** | Geo coordinates |  |

# Objective:

The City of New York tracks consumer complaints against businesses that were mediated by the Department of Consumer Affairs (DCA). The dataset contains data for last and current calendar years about the business, the complaint, and the outcome. It also includes a “satisfaction” field, indicated whether or not both parties were satisfied with the outcome. It excludes complaints that may have ongoing legal investigations.

There are many use cases for this dataset. For example, the DCA Consumer Services Division can use the data to track satisfaction with the mediation services provided. It can also leverage the data to track restitution amounts, as well as businesses with a large number of complaints.

For this analysis, we used basic statistical analysis, data exploration and visualization to gain an understanding of the data, and then used multiple techniques to predict if a new complaint will be satisfied or not based on other attributes, as well as a regression analysis to predict restitution amounts.   The data includes about 2600 hundred records and this may cause us to lean towards a cross validation approach but will still try other methods for training.

The overall objective is to create a model to help predict if a complaint with both parties being satisfied or not.

Note: This data is fairly new which is one of the reasons we selected so that we are not influenced by other models that may exist for older data sets.

# Data Cleansing:

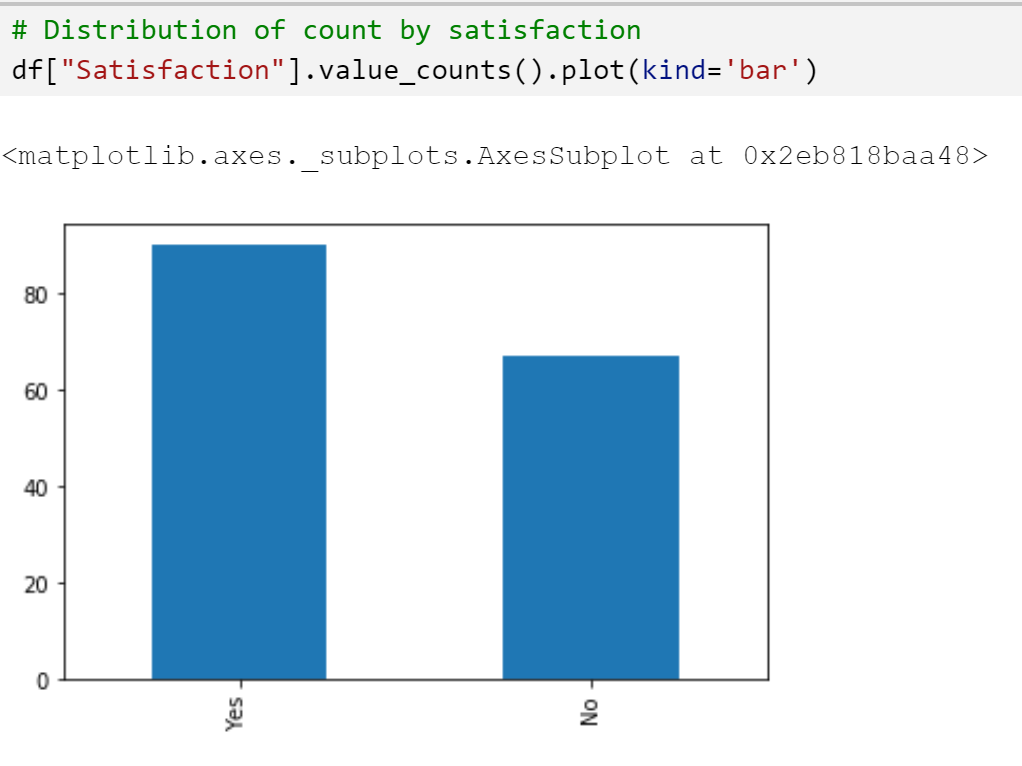
A number of data cleansing steps were required before moving into data exploration and machine learning tasks. The name of the businesses cited in the complaints had sometimes appeared differently for the same business entity. We leveraged the open-source tool [OpenRefine](https://openrefine.org/) to identify those situations and where we were certain that different versions of the business name referred to the same business, we merged them. The same process was taken for the Industry column as well as the Complaint Type.

We also needed to change the format of the columns to reflect the proper datatypes (all datatypes initially we of type ‘object’). We dropped columns that were not relevant to our analysis. Finally, using Pandas getdummies, we put the dataframe in standard spreadsheet format to be able to analyze categorical variables properly.

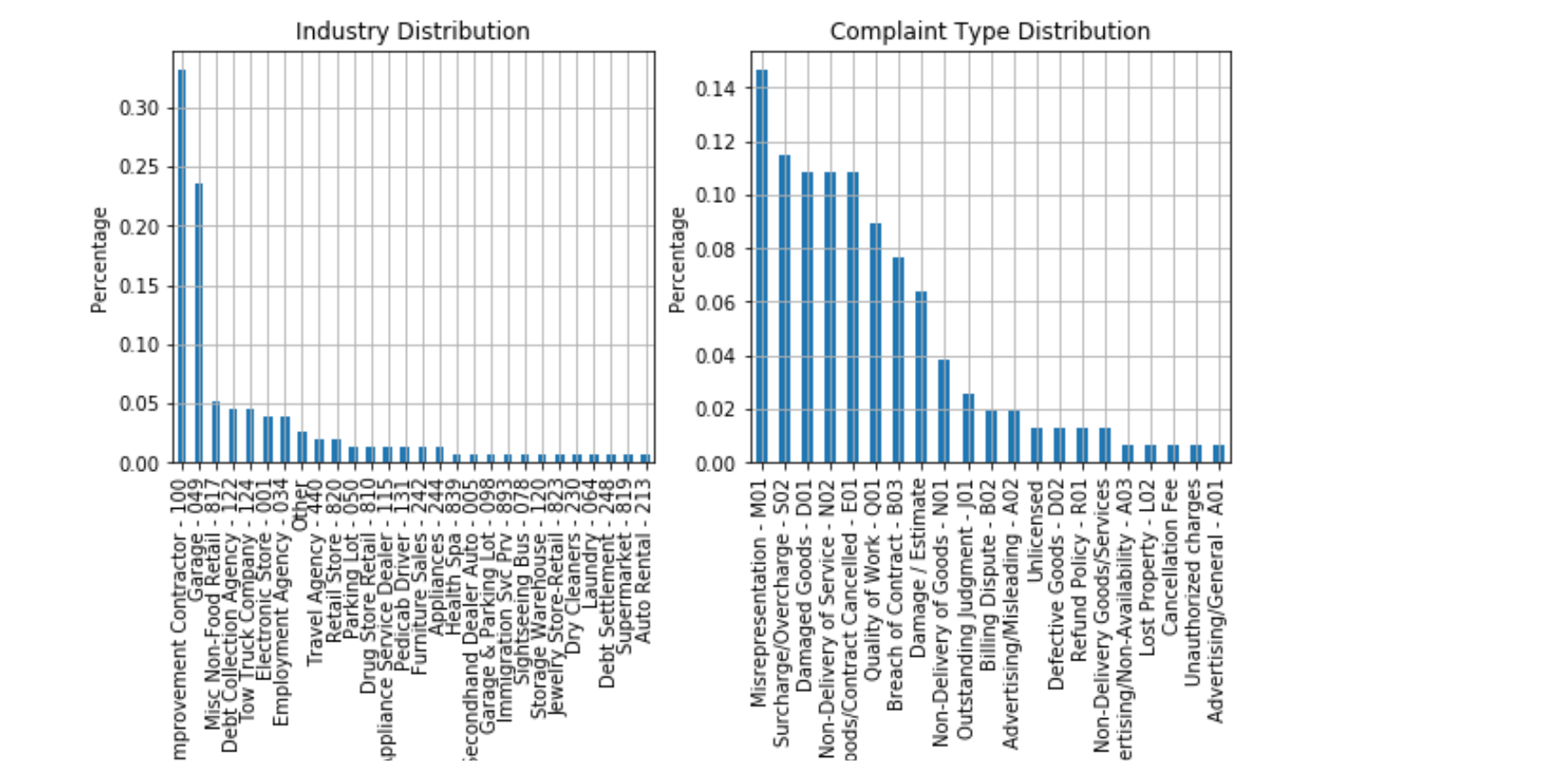
We then leveraged the start date and close date of the mediation process to create a new calculated column for duration. This new column was then changed from a datetime to a float as to appended to of the standard sheet format dataframe. We then leveraged the duration variable to get better results with our clustering algorithm (see below).

# Data Exploration:

Given that the focus of the analysis is making predictions of the satisfaction of both sides of a complaint, the first analysis was of the general distribution of the satisfaction variable (yes/no). Looking at raw counts, it is obvious that both parties are satisfied more often than not:



Next, we looked at the distribution (as a percentage) of the complaints both by industry of the business, as well as the type of the complaint. Home improvement contractors and garage services dominated the industries receiving complaints to the department, heavily skewing the distribution. The top 5 complaint types were, in order: misrepresentation, surcharge/overcharging, damaged goods, non-delivery of service, and quality of work.



We further performed a visualization on the actual counts by industry and complaint type.

An analysis of the ‘Restitution’ variable shows that a vast number (75%) of cases resulted in zero-dollar payouts:

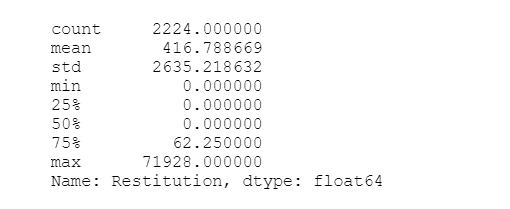


Figure 1 5 -number summary of restitution

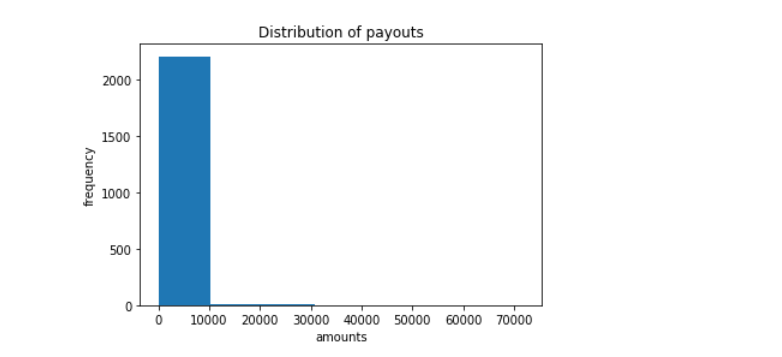


Figure 2 Distribution of restitution (raw)

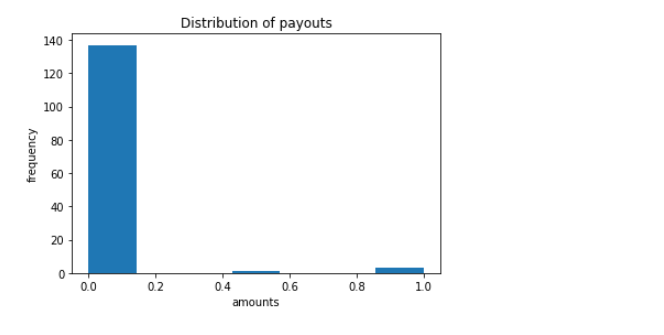


Figure 3 Distribution of restitution scaled

Because of the skewness of the restitution variable, we did not pursue a linear regression analysis.

# Knowledge Discovery – Supervised Learning

## KNN

In our analysis of the targeted data (satisfaction), we pursued both supervised machine learning models along with an unsupervised model to broaden what we could determine about that variable. We started with the supervised models. We first leveraged K Nearest Neighbors to predict whether a future case would bring satisfaction to both parties. The initial model used 10 neighbors and a Minkowski distance for weights. This provided an accuracy of 78% per the classification report below.

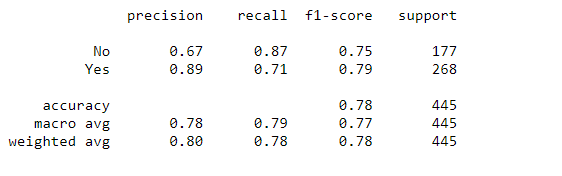


Figure 4 Classification Report Using n=10, Minkowski Distance

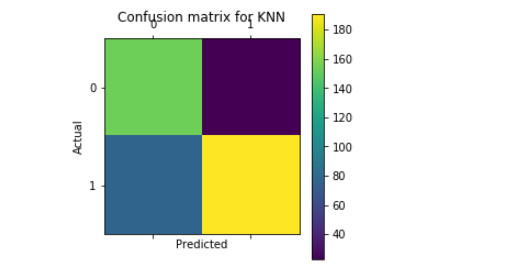


Figure 5 Confusion Matrix Using n=10, KNN n =10, no duration variable

We then calculated the duration of the mediation and incorporated the duration variable into the model. Re-running the KNN using duration, we were able to increase the accuracy by another 4% to nearly 82%:

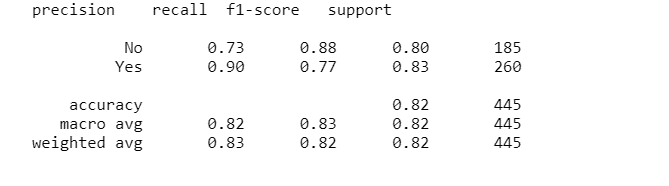


Figure 6 Classification Report for KNN incorporating duration

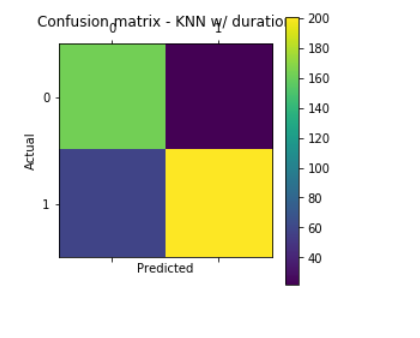


Figure 7 Confusion Matrix for KNN leveraging duration

## Decision Tree

We then tried a decision tree analysis, which provided insights as to which features were most important in determining satisfaction. Initially, we ran the decision tree using Gini as our cost function to measure impurity. This resulted in a tree that prioritizes…as values (see tree below). We set the max leaf node parameter to 10 to avoid a very large (and potentially overfitted) tree. This tree resulted an accuracy score of just over 78% (and a similar score using 10-fold cross validation):

This tree, using Gini, prioritizes whether a debt collection agency was involved in the complaint, followed by whether the complaint was filed against an in-state business vs. an out-of-state business. Business type and complaint type were also priority decision variables.

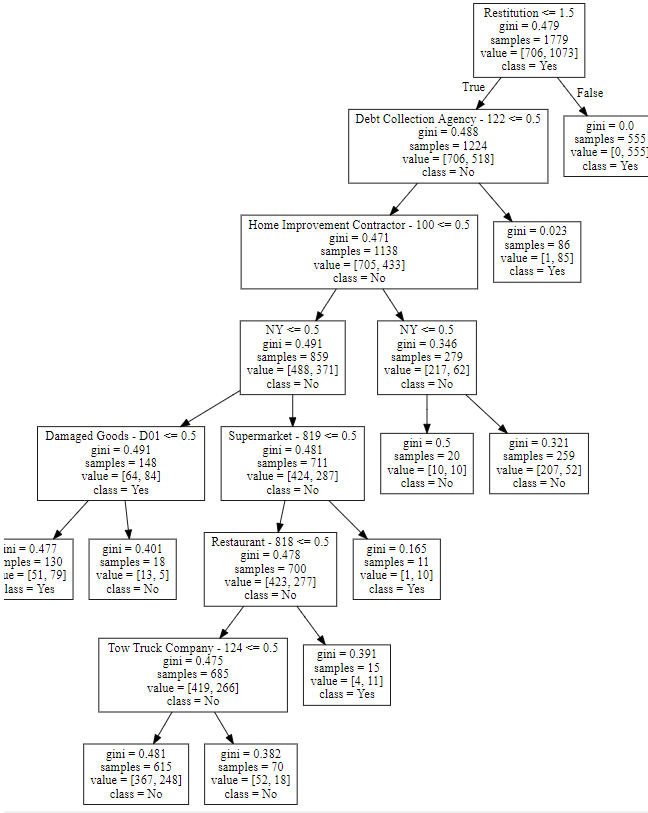


Figure 8 Decision Tree using Gini as cost function, max leaf node = 10

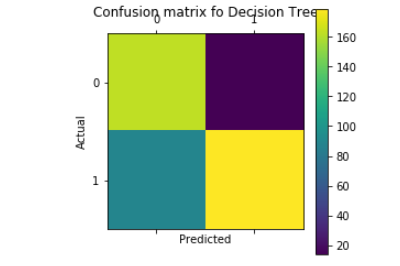
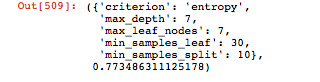


Figure 9 Confusion Matrix of Decision Tree using Gini as cost function

Next, we used a grid-search to determine the best parameters given Entropy as a cost function. The accuracy remained at a little over 78%. Debt Collection Agency, Business Type, Business State, and Complaint Type remained the top priority variables.



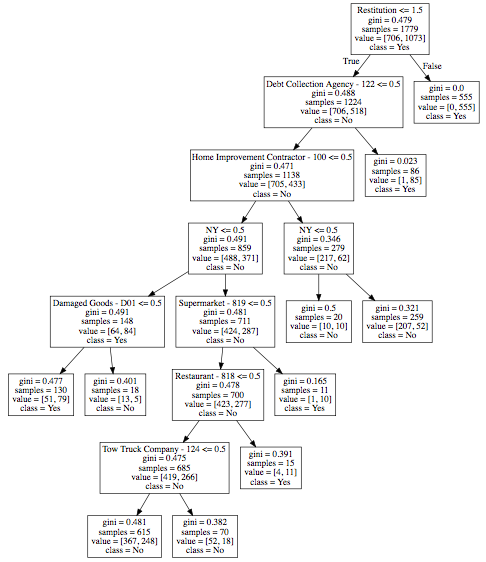


Figure 10 Decision Tree using Entropy as cost function

**Confusion Matrix**

[[163 14]

[ 89 179]]

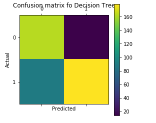
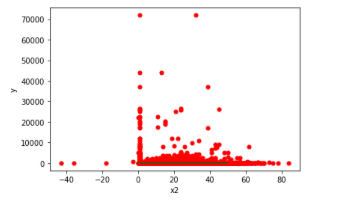


Figure 11 Confusion Matrix of Decision Tree using Entropy as cost function

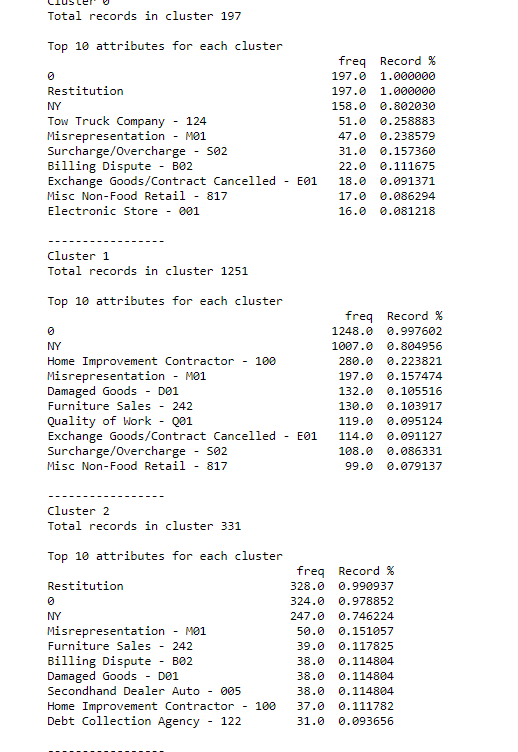
Finally, we explored the possibility of doing a linear regression on the amount of the restitution (the pay out in the case of a mediated settlement for money). However, the restitution distribution is non-linear, and very skewed to the right. We attempted to transform the data set by using logs or square roots but the distribution remained skewed.

**Scatter plot after MinMaxScaler with a feature range of 0 to 1**



# Knowledge Discovery – Unsupervised Learning

Next, we performed a Kmeans clustering algorithm (leveraging the custom algorithm from *Machine Learning in Action*, P. Harrington), and a cosine similarity function for distance, and k=3 clusters. The top attributes for the 3 clusters consisted of duration (denoted as 0 here), restitution, and in-state (NY) vs. out of state.



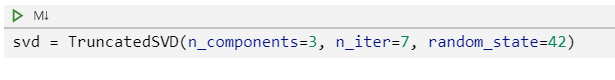
However, we know that these clusters could be better. With a completeness score of 23% and a homogeneity score of 27%, the quality of data points within the same cluster is low, as is the quality of data points between clusters.

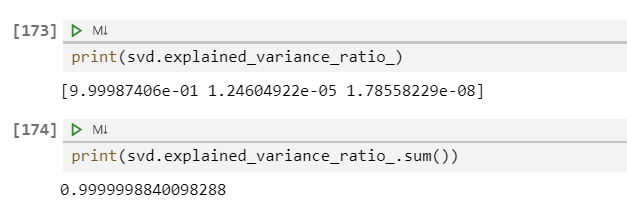


The v-measure was calculated to be 28.5%, which considers both completeness and homogeneity.

# Component Analysis:

There are a large number of features to be considered in the dataset (141 columns). We used Singular Value Decomposition to factorize our matrix. We used slightly more iterations than the default (default = 5, we used 7). Looking at the top three components, we can tell that a single factor accounts for over 99% of the variance in the data:





Rerunning the decision tree leveraging the single component, accuracy dropped down to a little over 59%

# Theories

We explored potential theories as to why Debt Collection Agency Involved and in-state business vs out-of-state business may factor so highly in satisfaction. One theory regarding debt collection is that both parties would prefer a mediated settlement versus debt collection. Debt collection usually means that the person owed the debt gets pennies on the dollar, and the person who has the debt collection claim does damage to their credit score, hampering prospects for future financial dealings.

Second, if you are an in-state business, it could mean that you are a sole proprietor or a local business who is dependent on local reputation. An out-of-state business such as Walgreens or BestBuy has less at stake from a damaged reputation than a local business such as a home improvement contractor or parking garage. Again, these are only theories and not based on any survey data or other research.

# Future Exploration

No analysis is ever complete, and this is no exception. Given more time, we would like to pursue the following:

1. We could continue to refine the K nearest neighbors using the optimal parameters from Grid Search, and even tweaking them further. This includes trying to use collaborative filtering—we hoped to use clustering to help with the KNN but the clustering model was not efficient.
2. Cluster analysis to improve the quality of the intercluster and intracluster dynamics.
3. Regression analysis: we would study the appropriate techniques in order to perform a non-linear regression analysis in order to model how restitution can be predicted. We wanted to drop record where the restitution is 0 but this would drop the volume data and we may lose true records where 0 restitution should be included in a linear model.
4. Finally, there is another dataset on the New York City data portal that records if a business is charged by NYC. This would have made for a useful binary column to add to our analysis to see if charged companies adds any improvement to the quality. Unfortunately, neither dataset denotes a business by a unique identifier. Merging the data based on the business name would have been inefficient. We, therefore, considered it as out-of-scope for this research.