Big Data Science, Final Report

Cole Smith
css@nyu.edu
Undergraduate

May 2019

1 Introduction

This paper outlines the extent of my final project on restaurant closures in New York City. The motivation for this project stemmed from an interest in the economic situations that restaurateurs face on a monthly basis – In this project, I considered factors like health violations, area demographics, economic growth, and economic volatility to estimate this scenario.

Data is primarily gathered from NYC Open Data. The most important of these data sets for the purpose of this project is the DOHMH New York City Restaurant Inspection Results. This data set provides location and violation information for a subset of New York City Restaurants and Sidewalk Cafes. Other data collected form NYC Open Data includes housing developments, area demographics, and Legally Operating Businesses in New York City.

2 Problem Statement

My original formulation of this problem was in the following question: *Will a given restaurant close within a period of time based upon factors like restaurant type, number of violations, cuisine, area, etc.?*

This original classification problem proved difficult due to the lack of a universal identifier for restaurants across data sets. I therefore reformulated the problem into both a **classification** and **regression** problem. That is, given a restaurant's violation history can we predict if they will be closed by the Health

Inspector? Or, given a month in time for a certain zip code, can we estimate the number of closures we will encounter in that zip code?

2.1 Defining a Closure

A restaurant closure can be defined in two different ways. Sometimes a restaurant must close due to a serious health code violation, for which they may re-open at a later time pending a successful inspection, and rectification of previous critical violations. This closure is presumably involuntary, and due its tentative nature, it is considered a **soft closure** for the purpose of this paper. The source of soft closures is the DOHMH New York City Restaurant Inspection Results Data Set.

While a restaurant must be inspected, it must also hold a DCA license so that it may legally operate within New York City. This licenses have expiration dates. If a restaurant does not choose to renew its license this is considered a closure for the purpose of this project. It is also assumed that a restaurant owner would not willingly sit on rented property while not being legally allowed to operate, or without the intention of re-opening, thus, if a license is not renewed, this is considered a **hard closure.** The source of hard closures is the Legally Operating Businesses Data Set.

2.2 Relationship Between Hard and Soft Closures

The Master Data Set proposed later in this paper will eventually aggregate soft and hard closures into one count of total closures for a given month and zip code. Thus, it is critical to note the relationship between these two closure types.

Since restaurants must both receive ongoing inspections, and hold a license, then it follows that both data sets have collected these restaurants from the same distribution. In the Master Data Set, the features found related to Soft Closures are used to model the total count of all closures (Soft plus Hard Closures) for an area. The accuracy of this method is discussed in the Conclusion and Future Work Section.

3 Data Sources

Data is sourced primarily from NYC Open Data. However, additional demographic, income, and tax data is sourced from the IRS, with economic trends (Dow Jones Industrial Average, VIX) from Yahoo Finance.

All data sets will have their string information and column names cleaned, stripped, spaces turned to underscores, and lowercased.

There also exist data sources not finally used in the Master Data Set. They are included in the code (see Conclusion and Future Work) but are not included here. Namely, these are the Housing Developments and Business Inspections Data Sets.

3.1 DOHMH New York City Restaurant Inspection Results

This data set is used to gather the following features:

- Zip Code
- Year
- Month
- Health Score
- Health Letter Grade
- Violation Count
- Total Inspection Count
- Critical Inspection Count
- Cuisine Type
- Soft Closure (Boolean)

3.1.1 Preparing Data Set For Prediction Soft Closures

The DOHMH Data Set is arranged such that each row is a violation. Therefore, a restaurant will appear multiple times for each inspection it will receive.

Moreover, some rows contain incomplete information, and some columns contain categorical variables. There also exists seemingly complete data but is ultimately nonsensical. For example, the provided data dictionary notes that a restaurant will receive an inspection date of January 1st, 1990 if it has not yet been inspected. Thus, I initially include a check for this variable.

The steps to prepare this data set, at a high level, are as follows:

1. Prepare the Grade column, only grades A, B, and C are considered. An A is transformed to a 3, B a 2, C a 1, and all others a 0. In this way, the grade's ordinal nature is preserved in an increasing way.
2. Transform violation action into a closed/not closed decision. This column contains text data which is self-explanatory. If the action included "establishment reclosed" or "establishment closed" then this is labelled as closed and the violation count is incremented. If the action was to re-open the restaurant, the closure is reversed, but the violation count remains static.
3. The other violations are added to the violation count.
4. The date is transformed into Year, Month, and Day columns.
5. A fix to the Critical Flag is fixed to make it numeric.
6. Unnecessary or redundant columns are dropped.
7. The violation-rows are sorted by date in reverse order
8. The rows are grouped by all restaurant identifiers. An option is given to the programmer to additionally group by violation types.
9. The group by aggregation is performed in the following way: critical flags, violation counts, and total inspections are summed. The score and grade is averaged.
10. The date and closed/not closed label will receive the latest entry in the data set.
11. A violation ratio is made by dividing the violation count by the total number of inspections.
12. A newer version of this same data set is then prepared in the same way, and joined on the year, month, and restaurant ID (CAMIS).
13. If a restaurant exists in the older dataset, but not the newer data set, it is labelled as closed. This is still treated as a soft closure for consistency in features and measurement bias.
14. The Cuisine Type column is then pivoted as one-hot columns.

3.1.2 Source

3.2 Legally Operating Businesses

This data set is used to define hard closures. The only information needed for this is the following:

- Zip Code
- Year
- Month
- License Expired

This data set is simply transformed as follows:

1. Data is filtered to only include Restaurants and Sidewalk Cafes.
2. Date column is transformed into Year and Month columns.
3. Rows are grouped by unique business identifier
4. If the latest entry has an expiry later than the current date, then the restaurant is assumed to be closed.

3.2.1 Source

3.3 Demographic Statistics By Zip Code

Demographic stats are joined as-is, but filtered only by NYC zip codes for faster processing.

In the master set, they will be joined by zip code.

3.3.1 Source

3.4 SOI Tax Stats - Individual Income Tax Statistics - 2016 ZIP Code Data

The Income Tax Statistics are cleaned by making all columns numeric. Only those related to the Adjusted Gross Income and total return count are used. This columns are then pivoted into columns, since their values are bucketed categorical variables.

3.4.1 Source

3.5 DJIA and VIX Information

Dated DJIA and VIX close prices are pulled from Yahoo Fiance for the total term of the entire data set so far. An EMA with a 5 month lag is also included as additional columns. These then aggregated into monthly averages.

3.5.1 Sources

DJIA VIX

4 Classification Analysis of Restaurant Inspection Data Set

The prepared DOHMH New York City Restaurant Inspection Results receives its own prediction output for soft closures. Multiple models are compared for the classification task.

Since each data is a unique restaurant, the rows are randomly shuffled and an 80% training data, 20% test data split is created. The values are then Min/Max Scaled to range [0,1].

The results of three different estimators are compared below, for the task of predicting the closure column for each row.

4.1 Rebalancing

This data set resulted in an extremely imbalanced classification problem. Before merging the newer version of the data set, the positivity rate was **0.05%**. Once merging the data set gathered 2 months later, the positivity rate was increased to **3.2%**.

For the final model input, the data was selectively down-sampled to a final **25%** positivity rate. Sparse, negative rows were removed first, then a bootstrap resampling was performed to achieve the desired 25% positivity rate.

4.2 Visualization

The following figure demonstrates the initial hypothesis that predicting the closure of restaurants is a dynamic, non-linear relationship of features. The shape of the 2-component PCA decomposition produces another hypothesis, and possible explanation of non-linearity, that the data points lie on an n-dimensional manifold.

In order to properly transform the data set to become linearly separable, a manifold learning task such as Isomap or Spectral embedding will need to be performed.

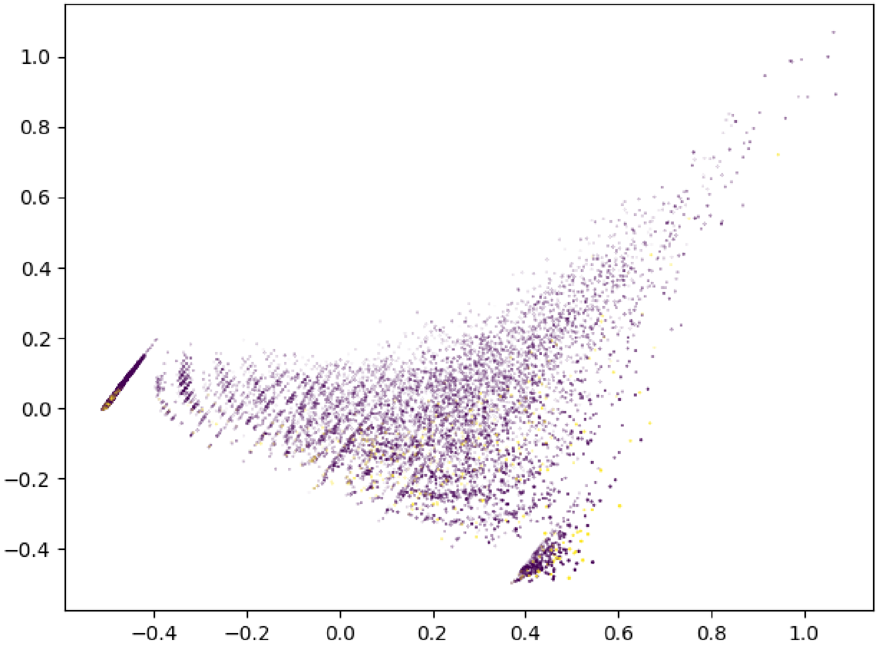

Figure 1: All Restaurants by First Two PCA Components (Closed Restaurants Labelled in Yellow)

4.3 Results

Training Set Size: 2,780

Test Set Size: 696

4.3.1 Gradient Boosting Machine

Number of Estimators: 300

Learning Rate: 0.3

F1 Score: 0.52

Cohen Kappa Score: 0.38

4.3.2 K Nearest Neighbors

Neighbors Considered: 30

Distance Metric: Euclidean

F1 score: 0.13

Cohen Kappa Score: 0.07

4.3.3 Multilayer Perceptron

Hidden Layers: 20, 7, 2

Alpha: 0.00005

Learning Rate Initialization: 0.001

Max Iterations: 500

Activation: ReLU

Solver: Adam

F1 Score: 0.47

Cohen Kappa: 0.32

5 Regression Analysis of Master Data Set

5.1 Preparing the Master Data Set

The Master Data Set is prepared by merging all previously discussed tables on the Year, Month, and Zip Code columns. This results in a data set will exactly 100 feature columns and one prediction column: total closures. Each row is therefore a zip code at a month in time.

The Master Data Set is performed in the same way as the Restaurant Inspections Data Set. Each row is treated independently, and thus, independent of time aside from the EMA measures for economic trends. Data is Min/Max Scaled, and split into 80% 20% Train Test sets.

Because of the large number of features, the prepared data set also undergoes a feature selection operation. Recursive Feature Elimination is run with Cross Validation and a Gradient Boosting Machine to reduce the number of features to about 40. Since this process is automatic, the number of features may change with newer data. Some columns are extremely spare and have low variance, these are columns that constitute the majority of information removed by the RFE-CV operation.

5.2 Visualization

The following chart plots the total closures against the DJIA and VIX closing prices over time. Since the date range is too short to indicate any seasonality, the decision was made to treat each zip code-month invariant of the past. This also makes sense when including features from the Restaurant Inspection Data Set, which is arranged in a restaurant-row format.

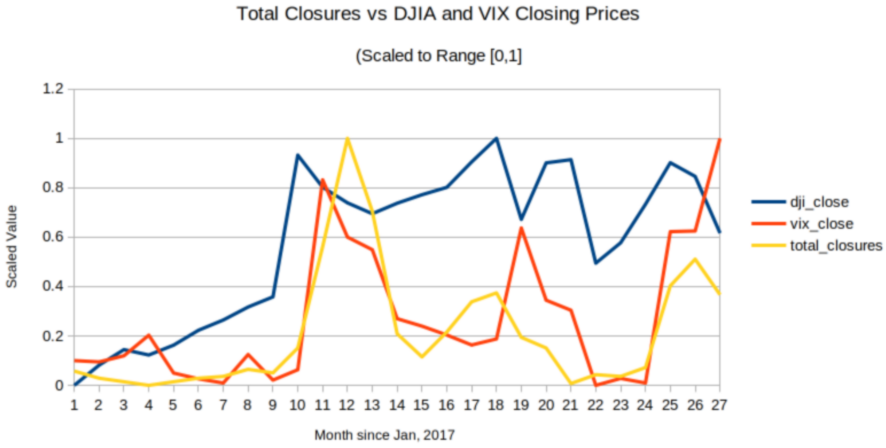

Figure 2: Total Closures Against Economic Growth and Volatility

5.3 Results

Predicting: Number of total closures per month per zip code

Training Set Size: 1,240

Test Set Size: 310

5.3.1 Gradient Boosting Machine

Number of Estimators: 300

Learning Rate: 0.13

Explained Variance Score: 0.79

MAE: 0.26

R-Squared: 0.78

5.3.2 Linear Regression

Explained Variance Score: 0.46

MAE: 0.35

R-Squared: 0.46

5.3.3 Multilayer Perceptron

Hidden Layers: 15, 5

Alpha: 0.001

Activation: ReLU

Solver: Adam

(Did not converge to result)

5.4 Important Features

Important Features are determined by the Gradient Boosting Machine, which achieved the best results of the three methods.

The following are the top 5 most important features as found by latest decision tree in a Gradient Boosting Machine:

1. Cuisine: American
2. DJIA Close Price
3. Cuisine: Cafes (Coffee and Tea)
4. Violation 08c Count
5. Cuisine: Chinese

6 Conclusion and Future Work

From the regression results of the Master Data Set, I found that about 40% of the explained variance is non linear, and could not be captured by linear regression. This confirms my initial hypothesis that both the Restaurant Inspections Data Set and Master Data Set likely lie on a manifold that is not linearly separable. An extension of this project would be to run this data through an embedding task in an attempt to decompose this non-linearity. The most important features also confirm assumptions that particular violations and particular cuisines are the most important factors in closures.

However, the cuisines are likely less important than what is noted here. Since the feature space is quite large, and many features are quite sparse, it is likely that these cuisines were simply the most populated features, since each cuisine is a one-hot column. Another possible extension to this project would be to determine the causality of the important features. This however, would require data to be collected over a long period of time to assess any seasonality, which likely arises in restaurant property leases at a yearly timescale.

6.1 Code and Distribution

The code and data for this project is publicly available under the Apache 2.0 license.

GitHub Repository