Restaurant Closures: Project Update

Cole Smith

Recap

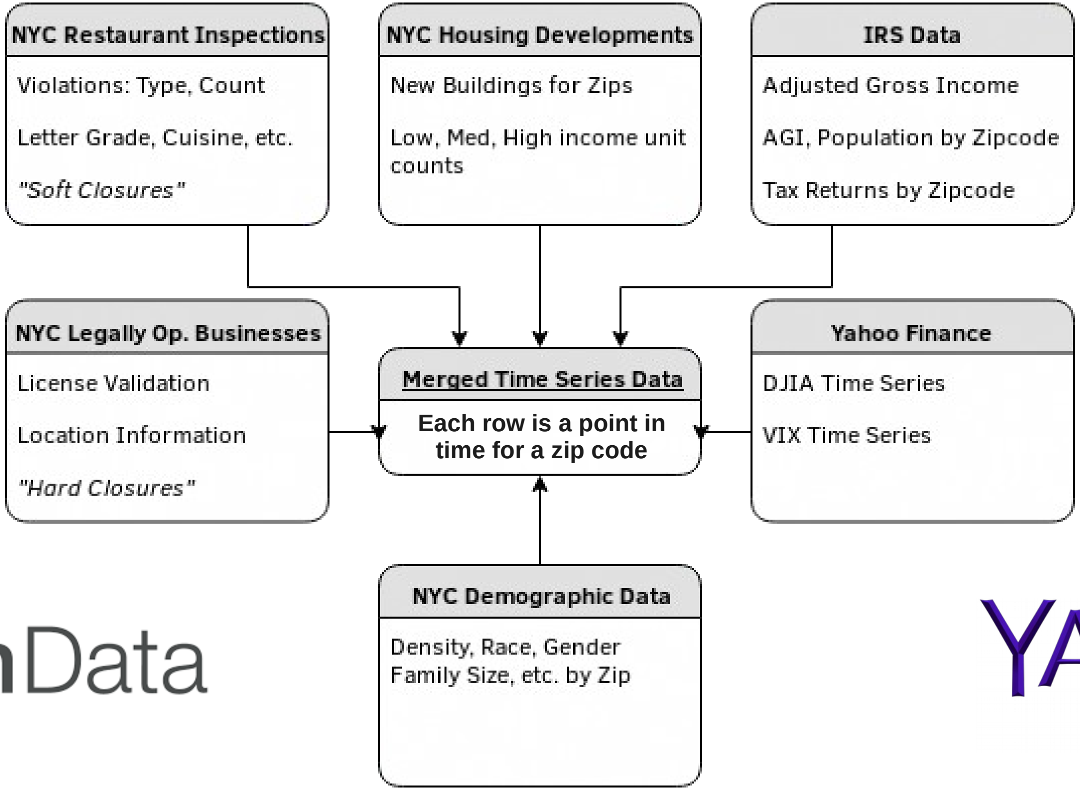
Objective is to predict

whether or not a restaurant will close within a given period of time

Based upon factors like

Health Violations (type, frequency), Location (demographics, income), Cuisine Type, Restaurant Type (restaurants vs. sidewalk cafes), and Economic Trends (DJIA, VIX)

Data Sources



NYC OpenData

YAHOO! FINANCE

Data Merging

- No universal identifier for restaurants between data sets
- **Fuzzy Joining** found infeasible due to restaurants using either legal or DBA name in data set.
- Common keys are therefore only **zip code, year**, and **month**
- Closed restaurants not represented in Inspections data set

thousands of restaurants start business and go out of business every year; only restaurants in an active status are included in the dataset.

Rethinking The Problem

- Two Types of Closures
  - *Hard Closures*: License expired and not renewed, cannot legally operate – presumably *voluntary*
  - *Soft Closures*: Health Inspection shutdown, may re-open at some point – presumably *involuntary*
- *Hard closures* appear related to *soft closures* on monthly aggregations
- **Regression** can be performed on *All Closures* while **Classification** can be performed on *Soft Closures*

Data Preparation

Data is set up to be analyzed in two ways:

Restaurant-row **format**

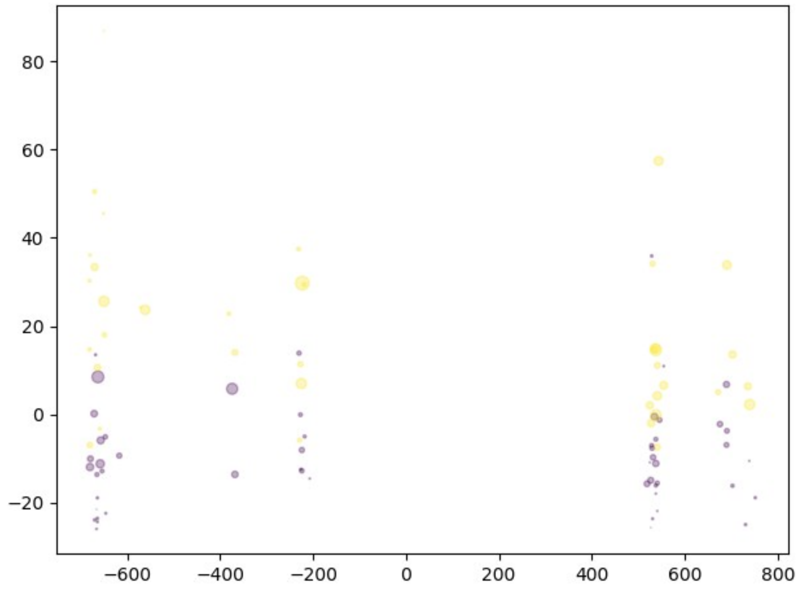
(Restaurant Inspection Data Set)

Monthly Zip-row **format**

(Master DataFrame)

Data Exploration: Restaurant Inspection Data Set

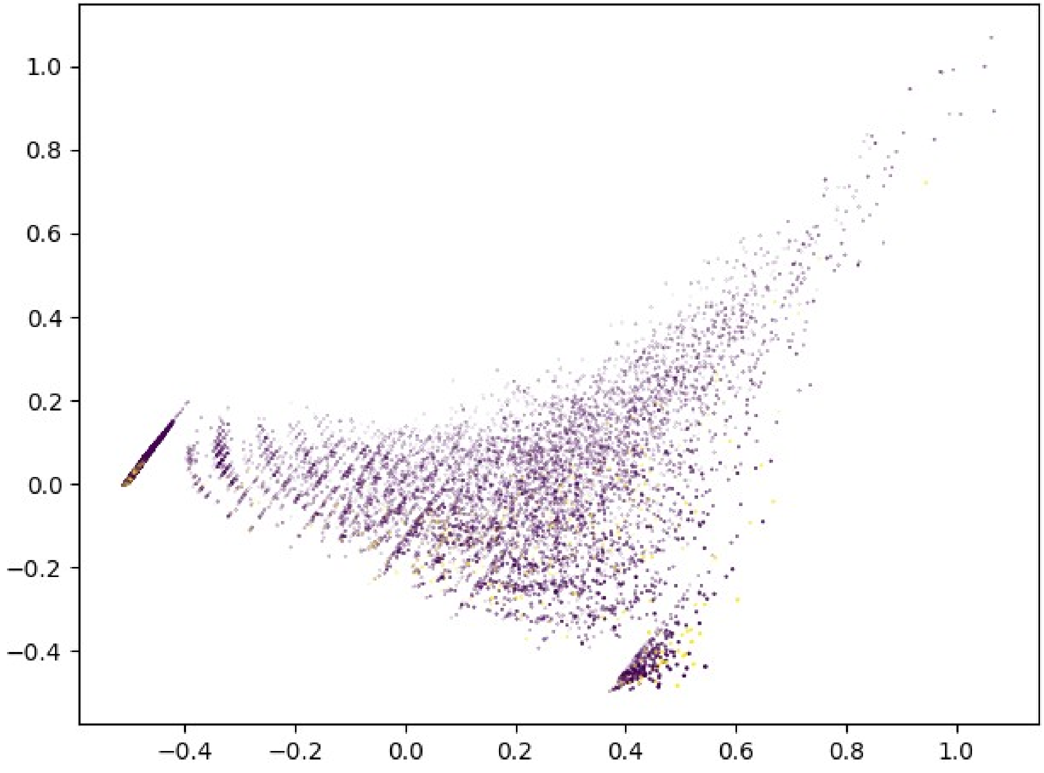
- Data extremely imbalanced (0.5% closure rate)
- Estimators perform poorly
- Bootstrap Sampling, Downsampling: Estimators perform well but over-fit (lack of specificity)



Fixing Imbalance

- Waiting until NYC releases new data set…
- Netted extra **750** closures over 2 months
- New total closure rate: **3.2%** of sampled restaurants

Restaurant Inspections: PCA Visualization



Classification Performance (*n=3,476*)

- Gradient Boosting Machine
  - F1: 0.51
  - Cohen Kappa: **0.38**
- K-Nearest Neighbors
  - F1: 0.12
  - Cohen Kappa: **0.07**
- Neural Network
  - F1: 0.46
  - Cohen Kappa: **0.38**
- **(Positivity Rate Bootstrapped to 25%)**

Top Features by RFE-CV and GBM

Score

Grade

Total Inspections

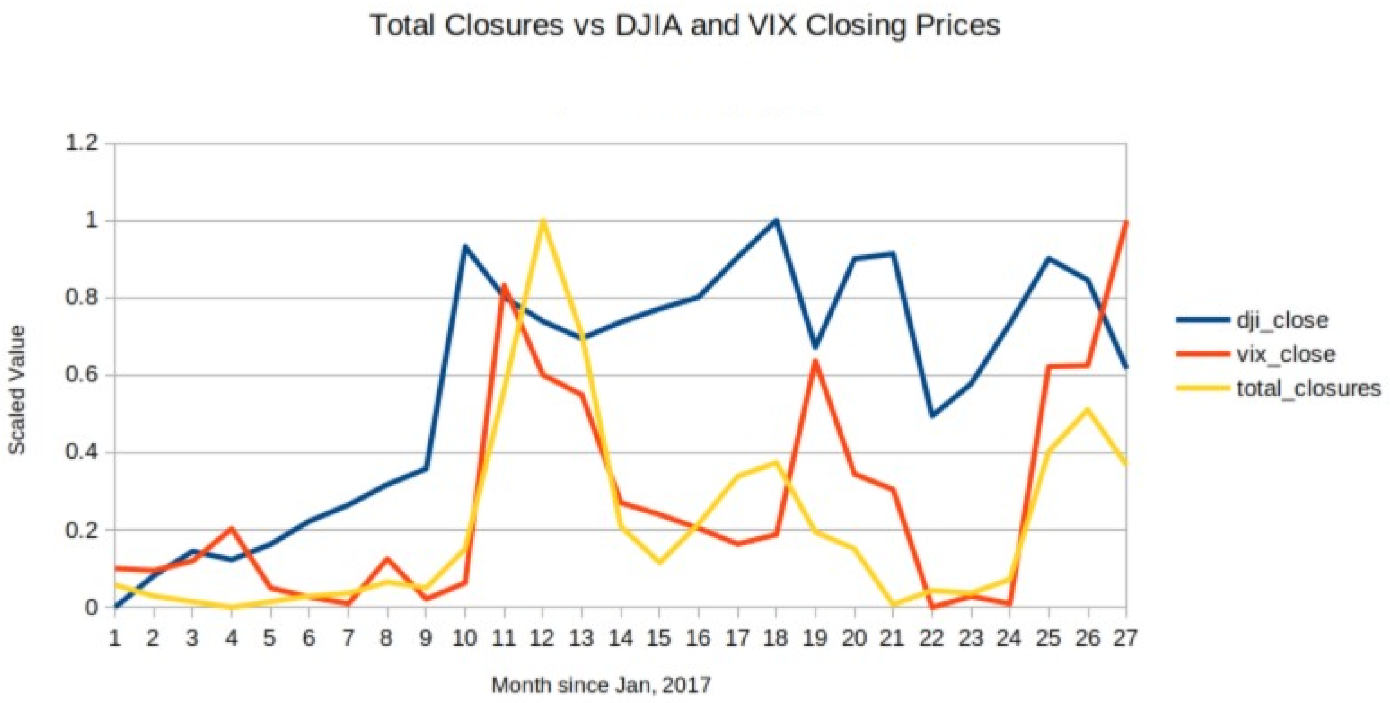
Violation Count

Critical Violations

Data Exploration: Master Data Set

- Includes all previous data plus
  - Economic Trends
  - Demographic
  - Adjusted Gross Income
- Time series: **Zip codes** in **monthly increments**
  - Data was treated **independent of time**, however

Master Data Set Over Time



Master Regression Performance (*n*=1,550)

- Gradient Boosting Machine
  - Exp. Var.: 0.78
  - MAE: 0.28
  - R2: **0.78**
- Neural Network
  - *No convergence (R2: ~ 0.0)*

- Linear Regression
  - Exp. Var.: 0.50
  - MAE: 0.36
  - R2: 0.50

Observations

- **Zip** was not important
  - Nor was demographic data.
- Cuisine type and DJIA/VIX was important (As found by RFE/GBM)
  1) American
  2) Cafes
  3) Chinese

- About **40%** of explained variance is **non-linear**
- Certain violations will always result in a *soft closure*
- *Soft closures* can **reliably model all closures**

Future Work

- Finding unique restaurant identifiers
- Restaurant Inspections Data Set can be collected over time
  - Points may lie on a **manifold** → Isomap/Spectral Embedding
- Analysis of Social Media Presence