Predicting Response Times of Firefighters in NYC

Mark Espina
November 2018

Data

Source: Data for this project was gathered from New York City Open data portal.

Main dataset used contains Incident Dispatch logs for New York City Fire Department from 2013 - 2017.

Size: >2.75M records

Data was supplemented with a listing of Firehouses in NYC

Preparing the Data

Cleaning:

Removed records where no Responding units (<500,000)

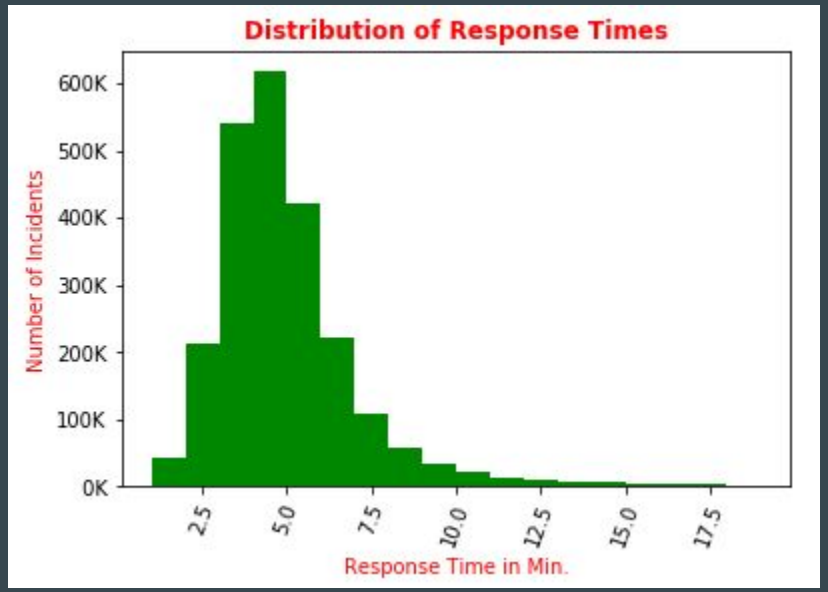
Extracted time stamp components from Incident datetime variable via Pandas DateTime

Factorized incident location data to compute avg Response time per location

Converted target variable from seconds to minutes

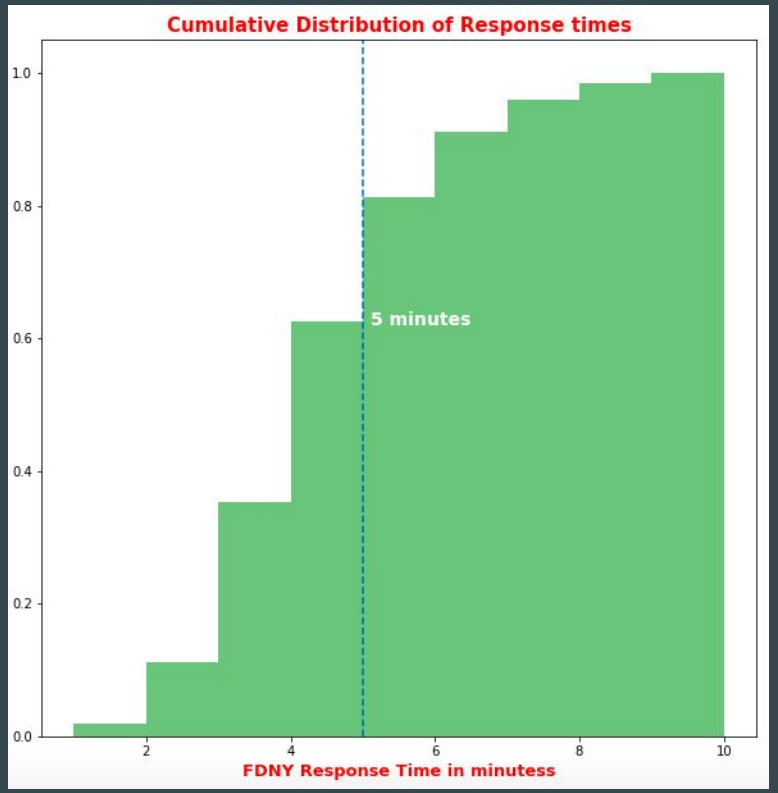
Removed Outliers in target variable

Distribution of target variable



Response Times of Target variable are normally distributed with long tail due to outliers

Distribution of target variable (cont'd)

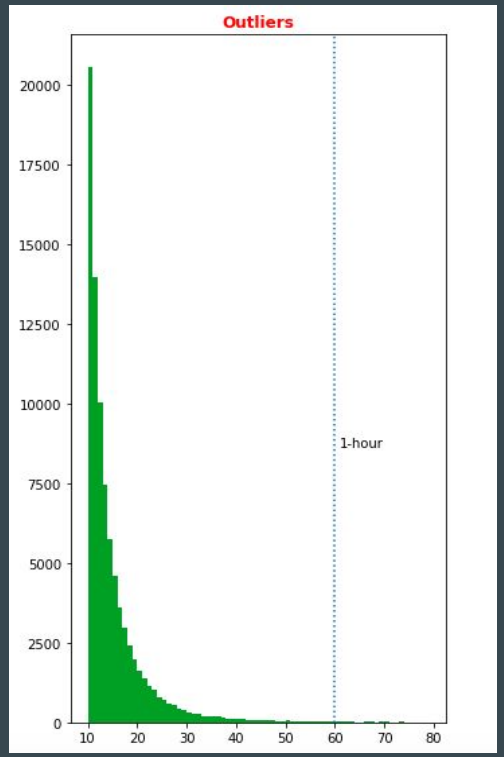


97% of response times

Under 10 minutes,

Outliers varied wildly

62% of Response times under 5 minutes



INITIAL DATA PREPARATION

Created Binary Variables

to model seasonality found in data

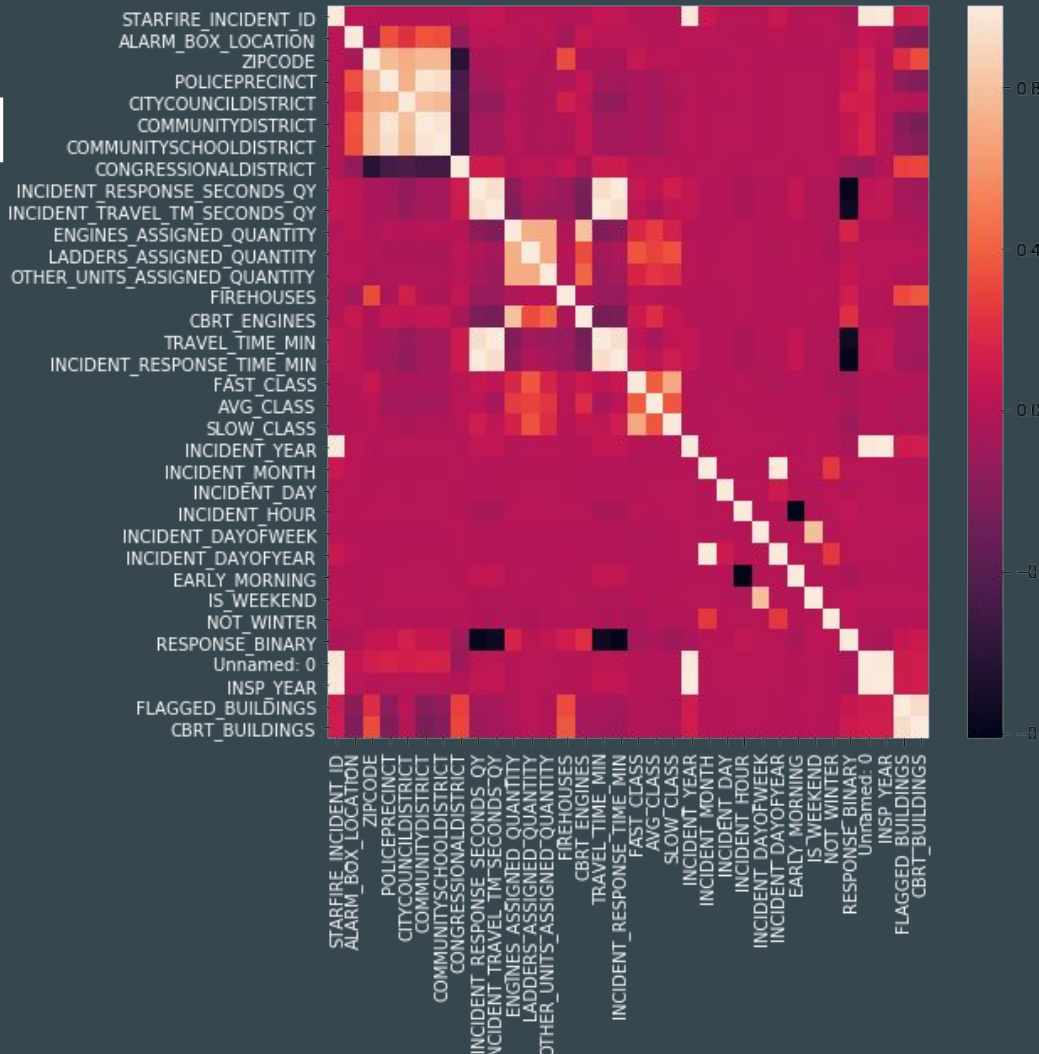
Cubed Root transformation of right skewed variable

Factorized Incident location data

Added categorical variables to feature matrix via get_dummies

Feature selection with Adaptive boost feature importances

Custom feature ranking via value counts



INITIAL MODELING

First Attempt: multi-class classification

Initial models included outliers as a third class:

Although initial modeling benefit from derived time variables,

The boosting classifiers benefited most from unaltered time component data,

This could be indicative of a complex influence of seasonality not seen during exploratory analysis

DATA EXPLORATION

Time stamps:

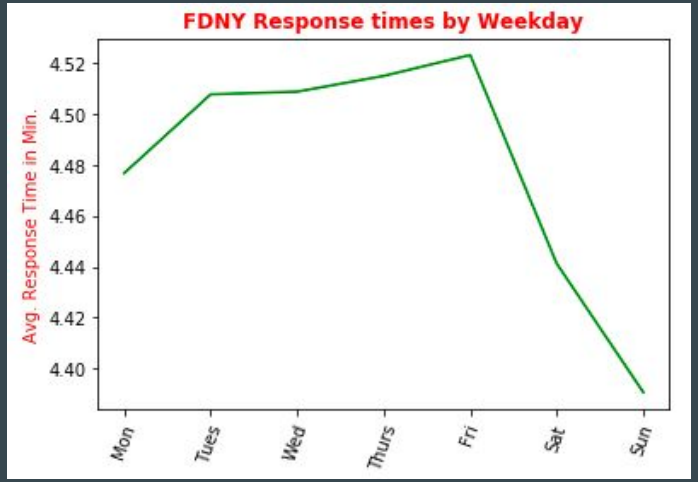
In terms of seasonality, whether or not the

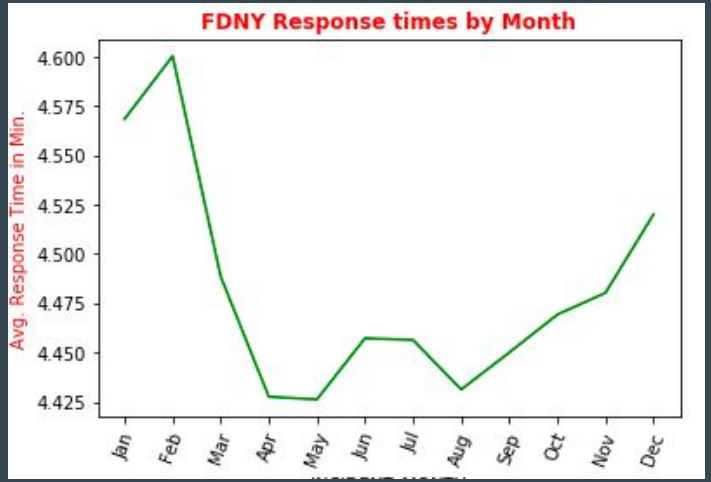
Incidents occured on the weekend or in the winter

(December, January, February

T-Statistic for Response times on Weekend vs Weekday: -38.83

T-Statistic for Response times onWinter vs Other Season: 43.35



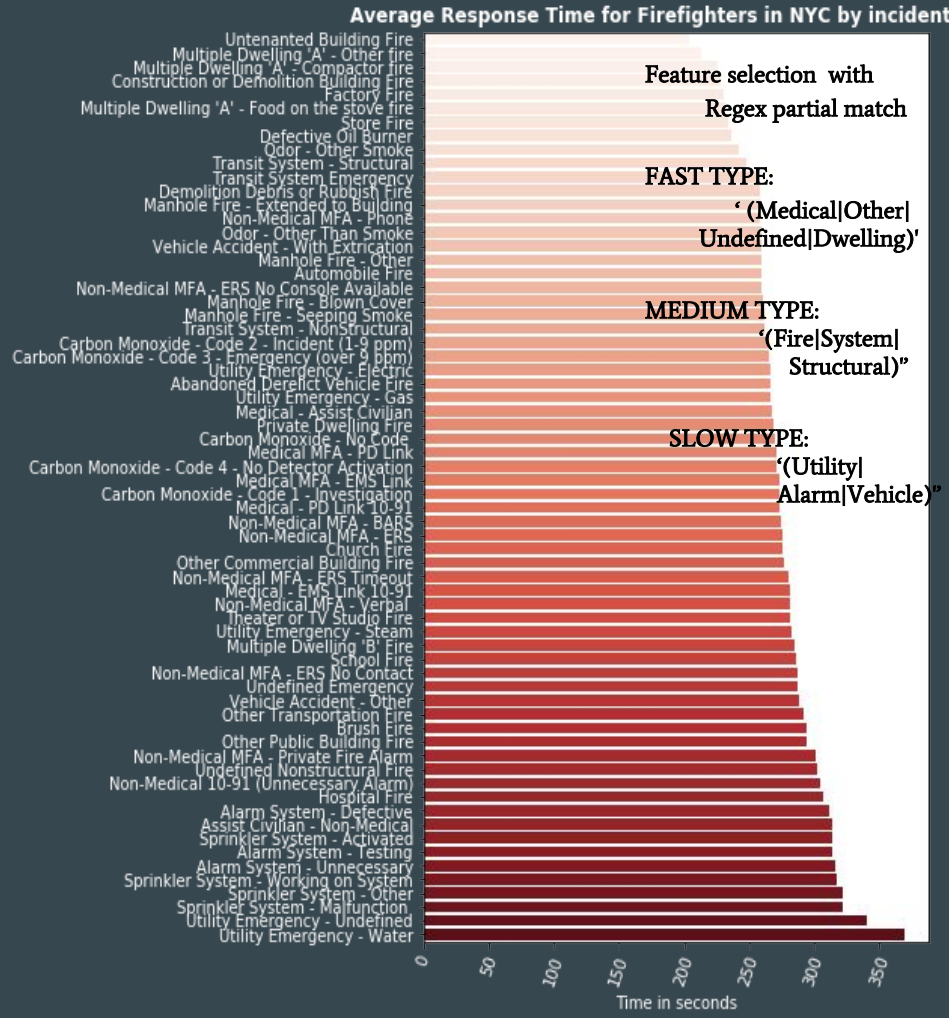


Dealing with categorical Variables

Incident classification was the most important categorical variable for increasing predictive accuracy.

Initial Attempts:

Manual rank of variable on 1-5 based on volume counts in data



The best performing early models benefited from addition of incidents types as dummy variables in feature matrix.

Preliminary attempts to reduce dimensionality with PCA, SELECTKBEST resulted in decreased accuracy across models

Limitations of Dataset

Best Performing models, Adaptive Boosting Classifier and Gradient Boosting Classifier achieved a max of 75% accuracy with maximum number of categorical features. Best features classification type is only revealed when processed as dummy variable and used in Adaptive Classifier model and or Gradient boosting

Other limitations:

avg response time derived from training data, was based on aggregated locations, not the actual incident location.

Most features had low/no correlation with target variable.

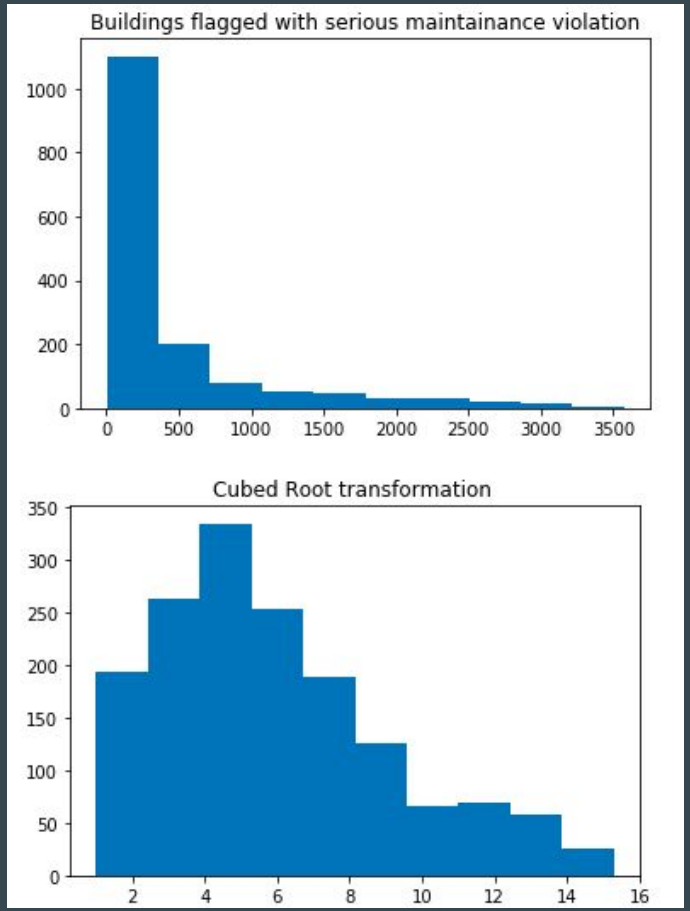
SEARCHING FOR MORE DATA

Other dataset maintained by FDNY either lacked a common field, or had too many missing value in common fields of interest to be used.

Additional Data Used:

NYC Housing Maintenance Code Violation 2012 -2016

Derived feature of interest: Total number of buildings flagged for ‘serious’ code violation by Zipcode from previous year (relative to fire dispatch incident)



Model Performance (with >100) features

RFC Accuracy: .719	Under 5 min	5 - 10 min	Knn (k=16) Accuracy: .727	Under 5 min	5 - 10 min
Precision	.79	.60	Precision	.77	62
Recal	.77	.63	Recall	.79	60
F1	.78	.61	F1	.60	61

More Models!! (with >100) features

Ridge Accuracy .749	Under 5 min	5 - 10 min	Logistic Reg: Accuracy. .745	Under 5 min	5 - 10 min
Precision	.77	.71	Precision	.76	.70
Recall	.87	.55	Recall	.86	.55
F1	.81	.62	F1	.81	.62

Boosting (with >100) features

Gradient Boost (depth=5 .7528	Under 5 min	5 - 10 min	Adaptive Boosting depth=3 .7517	Under 5 min	5 - 10 min
Precision	.77	.71	Precision	.77	.70
Recall	.86	.58	Recall	.86	.57
F1	.81	.63	F1	.81	.63

Model Performance (<15) features

RFC Accuracy: .721	Under 5 min	5 - 10 min	Knn (k=16) Accuracy: .727	Under 5 min Accuracy: .722	5 - 10 min
Precision	.77	.63	Precision	.77	.64
Recall	.80	.59	Recall	.80	.59
F1	.78	.61	f1	.78	.61

Penalized Models (<15) features

Ridge: .745	Under 5 min	5 - 10 min	Logistic Reg: .743	Under 5 min Accuracy:	5 - 10 min
Precision	.76	.71	Precision	.76	.70
Recall	.87	.54	Recall	.86	.54
F1	.81	.61	F1	.81	.61

Boosting (<15) features

Gradient Boost (depth=5 .7535	Under 5 min	5 - 10 min	Adaptive Boosting depth=3 .7515	Under 5 min	5 - 10 min
Precision	.77	.72	Precision	.77	.71
Recall	.86	.56	Recall	.86	.57
F1	.81	.63	F1	.81	.63

What have we learned?

Adaptive boost model performance consistency on reduced feature set suggest relationships between Response times and a number of feature including time of year, day of the week, nature of incident--utility, vehicle-involved, state of buildings in zipcode.

Other variable of Interest to Explore: Malicious False Alarms

The End