

Forecasting Fatalities in NYC

DSCT Capstone 1 - Data Story
Jonathan D. Williams



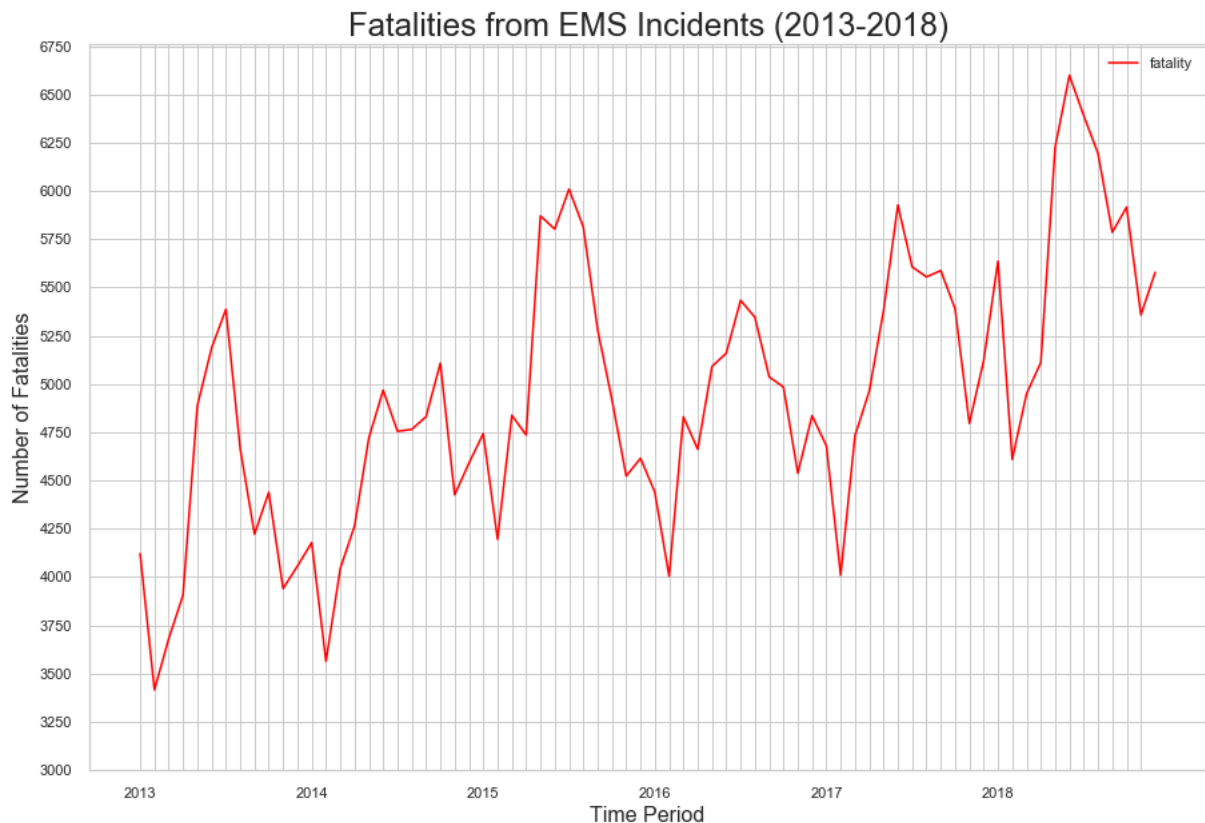
The Fire Department of the City of New York (FDNY) is the second largest fire department in the world, and the largest within the United States. More than 8.5 million residents and tourists within the five boroughs of New York City are protected on a daily basis by just under 15,000 uniformed personnel. This elite team of individuals is comprised of: firefighters, EMTs, paramedics, and administrative staff. The agency provides invaluable services to all within the city. Yet, in spite of their vigilance and bravery, the members of the FDNY cannot save everyone.

The goal of this project is to develop machine learning models that predict whether or not the outcome of an EMS incident will result in a fatality. This is a supervised, binary classification problem. Analyses will be performed on a collection of nearly 8.5 million records of documented incidents, which span the six year period from January 2013 through December 2018, and appropriate predictive models will be developed to achieve the primary objective. This dataset is robust and contains several feature variables that describe both various attributes of each incident as well as the responsive action taken by the FDNY. All of the aforementioned factors affect an individual's survivability once a response is initiated.

Exploratory data analysis was conducted prior to the development of any ML algorithms. The purpose of this process was to reveal trends that can help identify key feature variables and also provide contextual insights about the dataset. To this end, a natural starting point was to answer the following question: How many total recorded incidents resulted in a fatality, and how does this outcome vary over time? It turns out that the FDNY is highly effective at saving lives. Of the 8,4313,649 observations used in this analysis, only 355,939 resulted in a fatality.

EMS Incident Outcome	Value Count	Percentage
Survival	8075710	0.957785
Fatality	355939	0.042215

The time series below illustrates an upward trend in the annual number of fatalities that result from EMS incidents across the six-year observation period. In addition, the frequency of fatalities tend to spike during the middle of each year. This observation warrants a closer inspection to determine whether or not the incident month is a deterministic factor of a fatality.



In addition, there are many interesting feature variables within the dataset that lend themselves to the primary target: fatalities. These factors are examined in greater depth, using exploratory data analysis¹ and inferential statistics, in order to refine the ML models used to address the original objective. Some questions that are investigated include, but are not limited to, the following:

- How do the call type (`final_call_type`) and severity level (`final_severity_level_code`) assigned to an incident impact a patient's survivability?
- To what extent does the incident assessment determined from caller-provided information (`initial_severity_level_code`) correlate with the assessment made by response personnel on-the-scene (`final_severity_level_code`)?
- Which responsive actions, if any, by the FDNY have the greatest impact on a positive outcome of an incident?
- What is the frequency of fatalities within each borough of the City of New York?

Ultimately, the insights gathered from this analysis will help identify the feature variables that have the most significant impact on the outcome of an EMS incident. The final results can be used by the FDNY to minimize the number of lives lost within New York City.

¹ Jupyter Notebook: https://github.com/jdwill917/SB-DSCT-Repo/blob/master/Capstones/Capstone%201/CP1-02_EDA.ipynb