Prediction of Emergency Medical Services (EMS) Response for Fire Incidents

Project Report
CS-GY 6053 Foundations of Data Science.
Prof. Rumi Chunara.
Tandon School of Engineering
New York University, New York.

SUBMITTED BY –
RAKAVEE ANANDAN (RA2635) AND SHISHIR SINGAPURA LAKSHMINARAYAN (SSL495)

Background and Motivation

According to NFPA (National Fire Protection Association), a fire-related incident is reported every 24 seconds in the United States [1]. Thus, there is an important need to analyze fire incidents data to better understand the cause of fires, mitigate the incidents, and most importantly, improve the response of the association to fire calls.

Moreover, fire incidents are known to cause significant harm to people and property. Whenever a fire incident occurs, people are vulnerable to injuries, intoxication and other life-threatening hazards. Therefore, providing Emergency Medical Services in time during such scenarios helps in significantly controlling damage and casualties.

The need to understand the reasons and responsiveness of EMS during fire incidents motivated this project of analyzing and finding meaningful inferences from the available data. The National Fire Incident Reporting System (NFIRS) collects information once a year about the various fire incidents across the country including details pertaining to the EMS team such as the number of personnel that attended to an incident. For this project, the NFIRS dataset of General Incident Information from the year 2014 was used [2][3].

Problem Statement

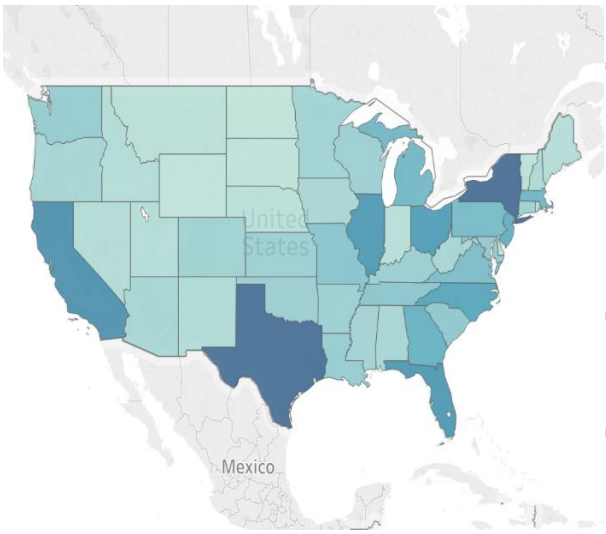
The goal of this project was to predict whether a fire incident required emergency medical services based on existing incident data such as the type of the incident, injuries and casualties incurred, and action responses.

Target and Predictor Variables

The target variable was emergency medical services personnel (ems_per). The dataset contained the target variable. The target variable comprised of discrete numeric values. However, since the prediction task was to predict whether a incident required EMS response or not, the target variable was converted to binary.

The dataset contained 57 features. However, careful exploratory data analysis and feature selection were used to select predictor variables for the prediction problem. The following steps were taken.

- Descriptive data analysis was performed on the features to gain better insight. The results helped understand information such as the states that had the highest number of incidents, distribution of incidents based on the cause of the fire, commonly used actions for response among others and type of property damage due to the incidents.



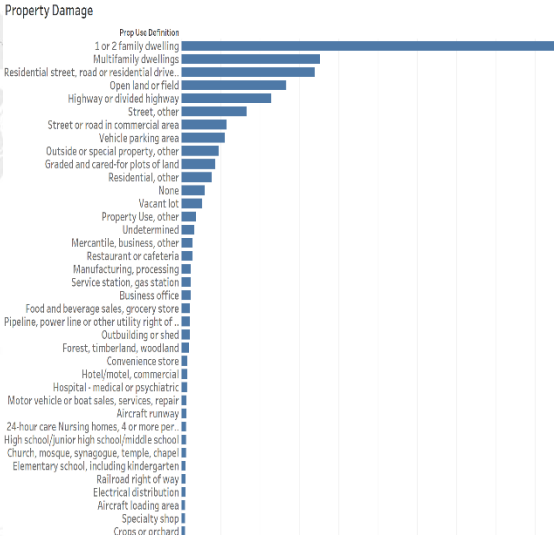


Fig 1: State-wise Incident occurrences. Fig 2: Types of property damaged.
(Please refer to the original Tableau workbook or image files for complete images)

- Exploratory data analysis - Histogram plots were used to find out the distribution of the target variable for each of the predictor variables.

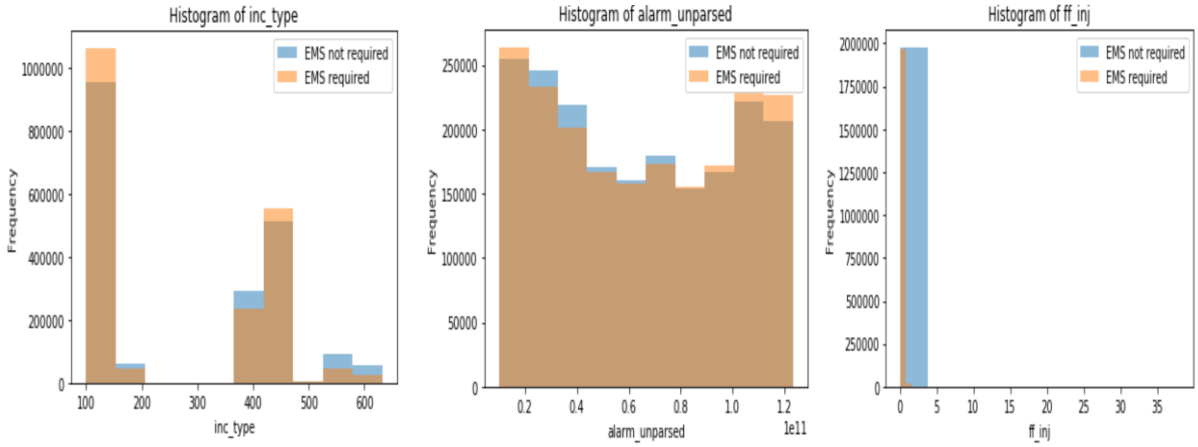


Fig 3: Histogram plots of target variable over three of the predictor variables.

- Data Cleaning - The dataset contained over 2 million rows. Therefore, the features where more than half the entries were missing were dropped. Features like fire injures and death had relatively fewer missing data and contained significantly low missing values. Missing data for such features was imputed with the mean value of that column as done in previous similar work and pandas official documentation [4].

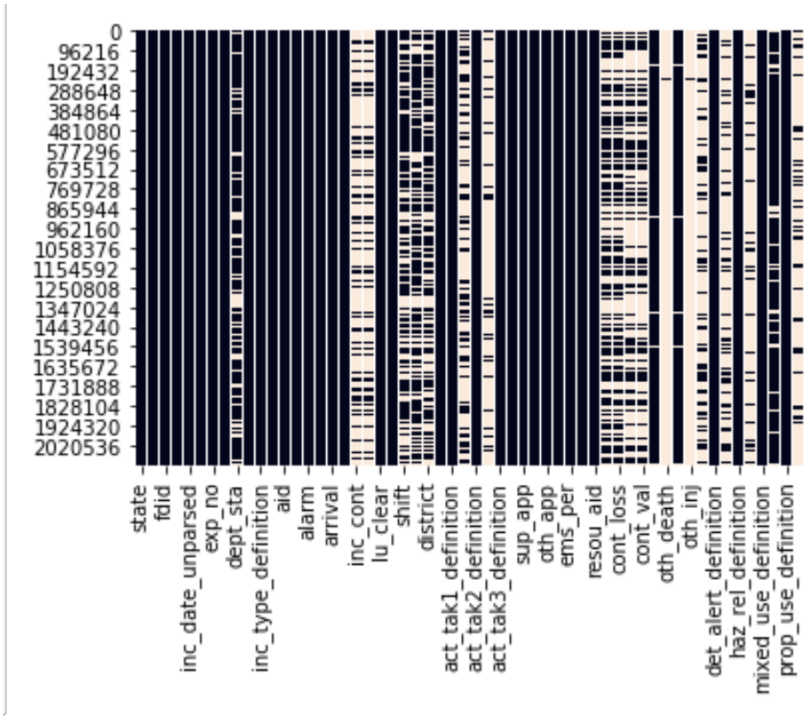


Fig 4: Heatmap of missing data.

- Feature Selection - Features (e.g. ems_apparatus) that were intuitively related to our target variable were removed. These features also had a near perfect linear correlation with the target variable which made sense. For example, EMS apparatus will only be needed when EMS personnel are needed. Also, the dataset contained binary categorical data which were converted to numeric.

  After data cleaning and feature selection, the variables shown in the following graph were used for prediction:

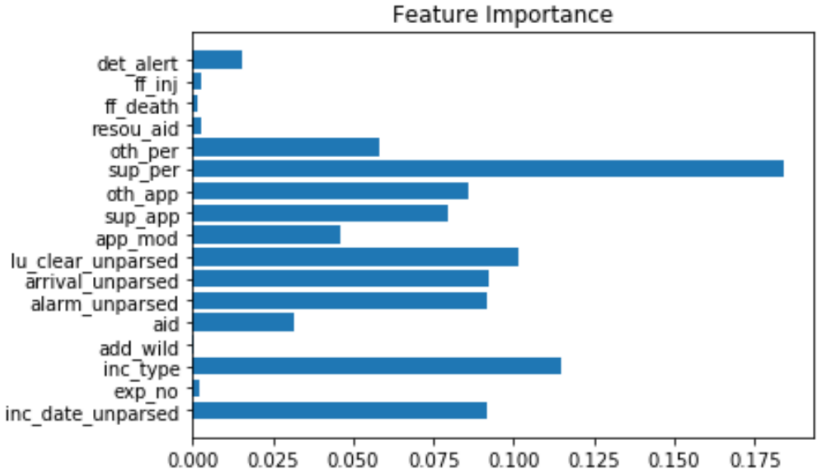


Fig 5: Feature importance of predictor variables.

Models

The goal was to predict whether a particular fire incident required response from Emergency Medical Services (ems_per) based on features such as area of occurrence information, type of fire, etc. As the dataset contained the labels used for prediction, supervised learning approach was used. Since the outcome variable (ems_per) was binary, binary classification models were chosen for the prediction task.

Logistic Regression: - The target variable (ems_per) was binary and the features were numeric values, therefore logistic regression was considered a suitable model for prediction.

Multinomial Naive Bayes: - As the features arose from a distribution that contained discrete values, multinomial Naive Bayes seemed a suitable choice.

Decision Tree: - The dataset used contained a large amount of data (2 million rows). Decision trees due to its fast learning tasks are known to work well with such large datasets. Therefore, decision trees were used for the prediction model.

Thus, the above models were used for the classification task.

Evaluation

Whenever a violent fire incident occurs, response from emergency medical services personnel is extremely critical. Therefore, the tradeoff was that it was better even if EMS personnel responded when they were not needed rather than EMS personnel not responding to incidents where they were absolutely needed. Thus, sensitivity or recall was a very important evaluation metric for this prediction model. Decision tree showed the highest recall score of 0.9798 among the three models.

The prediction model desired should be one where the predictions for the number of cases where EMS responded whenever the incident required a response significantly outrank the number of cases where the EMS did not. Hence, another metric suitable for evaluation in this context would be the AUC curves. The performance of the three models were compared using the AUC curves and the following results were obtained.

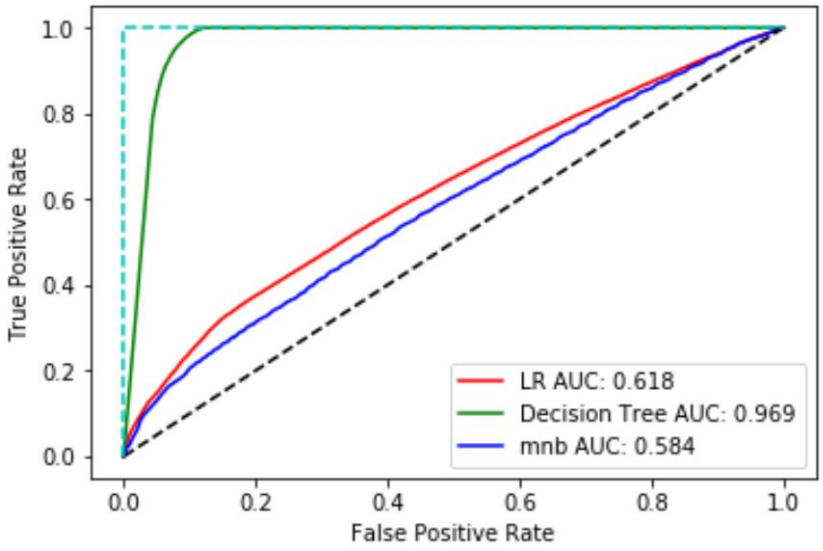


Fig 6: AUC Curves for the tested models.

Cross validation was performed, and a consistent accuracy score was obtained across the 10 folds. For Decision Tree, the following values were obtained - [0.94850994 0.9481888 0.94835809 0.94835809 0.94845165 0.94820889 0.94911418 0.9487374 0.94887901 0.94851993].

Limitations/Assumptions

The major limitation encountered was due to the imbalance in the class labels. The number of instances where the EMS personnel (ems_per) were not required significantly dominated the number of instances where ems_per were required. Therefore, in order to balance the dataset, two approaches were considered – under-sampling and over-sampling. From the initial results obtained, it was observed that under-sampling did not perform quite as well. Therefore, over-sampling was considered to handle the imbalance and for further modelling.

```
Random over-sampling:
1    1977271
0    1977271
Name: ems_per, dtype: int64
```

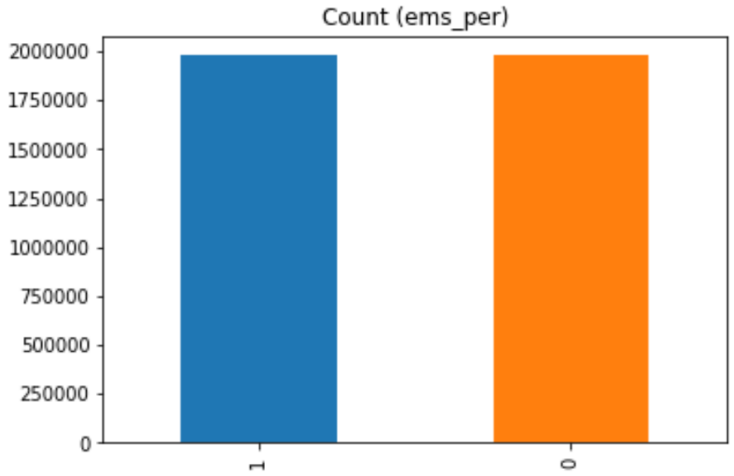


Fig 7: Random Over-sampling of Target Variable

Changes made

Initially, the features were assumed to arise from a continuous distribution and thus Gaussian Naive Bayes was considered to model the data. However, on further examining the data carefully, it was observed that the features comprise of discrete values and rather stem from a multinomial distribution. Hence Multinomial Naive Bayes was considered for further training and evaluation.

References

1. https://www.nfpa.org/News-and-Research/Data-research-and-tools/US-Fire-Problem
2. https://www.usfa.fema.gov/data/nfirs/about/index.html
3. https://public.enigma.com/spotlight/essentials/national-fire-incident-reporting-system-nfirs
4. https://pandas.pydata.org/pandas-docs/stable/missing_data.html