Capstone Project 1

Boston Fire Alarm Type Prediction

Summary

This report analyzes Boston Fire Alarm Incidents and establishes a predictive ML algorithm to predict these incidents, which helps Boston Fire Department improve operational efficiency. Five parts are included in this report: Project Overview, Data Wrangling, EDA Analysis, ML Algorithm Development and Conclusions/Deliverables. Source code is shared here.

Project Overview

Background

Boston fire department receives various alarms every day. Some alarms are urgent and might lead to disastrous incidents, which require more firefighters. However, more reports are minor, even false or unwanted, and fewer firefighters are needed. So types of fire incidents has a great impact on Boston fire department's resource distribution and operational efficiency.

Problem

Prediction on the daily types of fire incident would become beneficial. It will pre-filter those incidents report on phone or fire alarm first. Then the departments would benefit from higher operational efficiency and more optimized labor distribution.

Client

Boston Fire Departments wants to improve their operating system efficiency to deal with fire incident reporting. As a data scientist in a consulting company, I will analyze data, gain insights from data, build an algorithm to predict daily fire alarm type for Boston fire department to improve their operations.

Approach

I will extract data mainly from monthly Boston fire incident reports from 2012-2018. Meanwhile, data from daily historical weather, and community population density would be useful. (So far, The entire incidents would be about 350,000). Based on zip code and address, these data would be merged in one. The main features in merged data would contain incidents reporting time, location(zip code, district, address etc.), alarm type, property owners information, property assessment results, daily weather information, local population density and nearby events. Through EDA, those data would be visualized and more stories would be discovered. Finally, a predictive model will be built to predict the type.

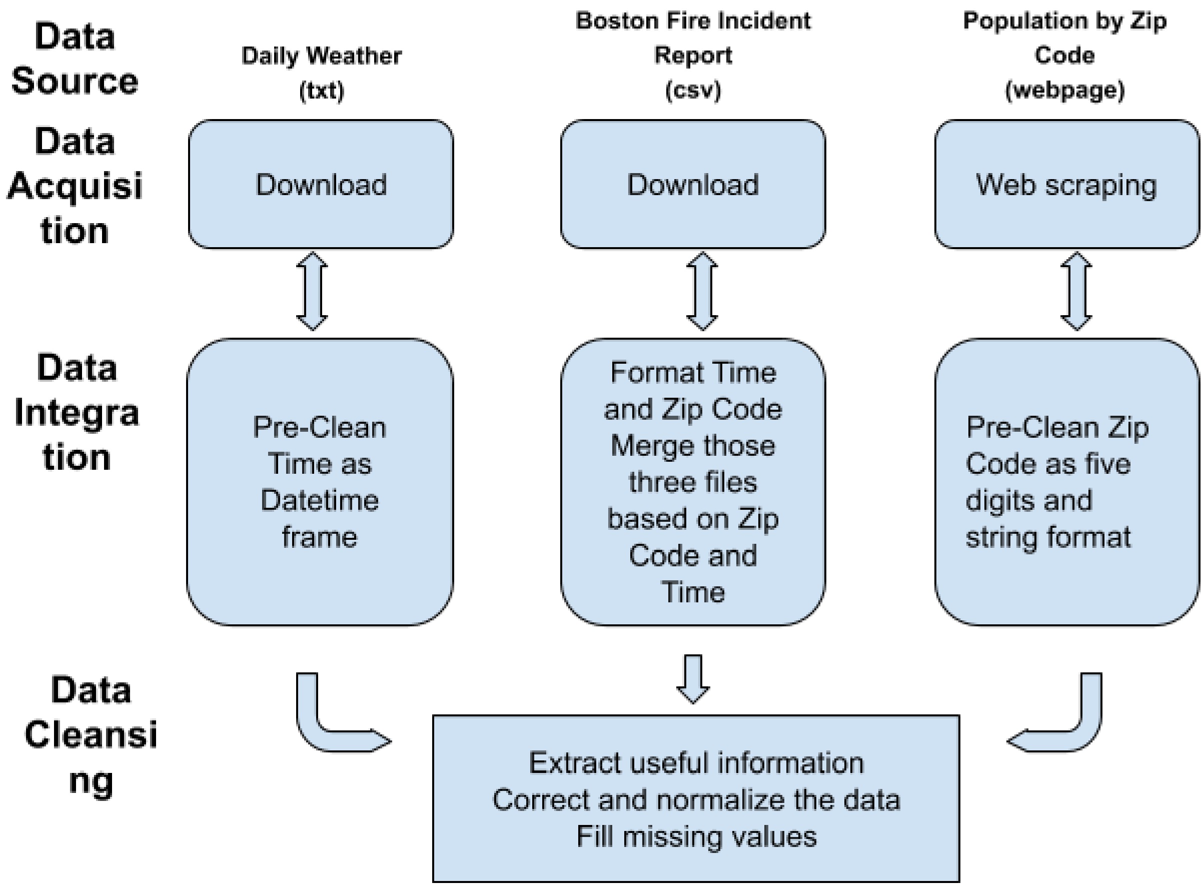
Deliverables

Source code will be displayed in Github A report and powerpoint slides would be shared. And I will post the results in Medium (Hopefully, I can try to deploy the model in AWS or Google Cloud or build an API for the dataset etc.)

Data Wrangling

In this part, I will explain how to collect, wrangle and clean data for Capstone Project 1, Boston Fire Alarm Type Prediction.The goal for this step is to get clean data and prepare for data analytics and prediction model. There are five steps: Data Acquisition, Data Integration, Data Extraction, Data Normalization and Data Correction. The data sources come from three files: Fire Alarm Incident Report, Hourly Weather and Community Population.

Data_Preparation.ipynb completes the first two steps while Data_Cleansing.ipynb completes the last three steps. Source code can be found here. I will explain these steps in details. The process will be followed as the figure below.


Data Source
Daily Weather (txt)
Boston Fire Incident Report (csv)
Population by Zip Code (webpage)
Data Acquisition
Download
Download
Web scraping
Data Integration
Pre-Clean Time as Datetime frame
Format Time and Zip Code Merge those three files based on Zip Code and Time
Pre-Clean Zip Code as five digits and string format
Data Cleansing
Extract useful information
Correct and normalize the data
Fill missing values

Data Acquiring

Data are downloaded individaully from Boston Fire Incident Report(https://data.boston.gov/dataset/fire-incident-reporting), Historical Boston weather(http://www.frontierweather.com/historicaldataonly/KBOS.txt) and Population Community (http://zipatlas.com/us/ma/zip-code-comparison/population-density.5.html)

Data Integration

1. Format data: Zipcode must contain five digits only and first two digits must be “01” or “02”. Date must be datetime frame format: Month/Day/Year Hour.

2. Data Integration: Based on time, Fire Incident would be merged with hourly weather. Then the data from population would be merged based on zip code.

Data Extraction

We need useful information from some columns, such as Incident Type, District and Weather. For example, to simply the problem, the first number is extracted to represent the Incident Type. Meanwhile, non-digit characters in Incident Type column need to be converted to empty blank.

Data Normalization

Some columns having various representation need to be converted to same format, such as Main Address and Address 2. We need to combine Street Number, Prefix, Type, Suffix. Some columns have to be dropped

Data Correction

Data Some rows with missing values must be removed, such as Incident Type. Some missing values might be filled as 0 or mean values such as Precipitation, Population and Population Density. The NaN of text columns would be filled as the string “None” or “Unknown”.

EDA Analysis:

- **Graphical Analysis**

  Through graphical analysis, a few conclusions can be drawn. More detailed analysis can be found in the powerpoint slides.

  - The real time incident report does not change while the monthly count of incident reporting is increasing every year.
  - The top three Incident reporting type is False Alarm and Fake Call, Service Call and Good Intent Call,
  - The top area of Incident reporting is Boston, Dorchester and Roxbury
  - Weather is an import factor on Incident Reporting.

- **Statistical Inference:**

  From the previous graphical analysis, we have further questions to answer:

  1. Is real-time Incident Type time dependent?
  2. Is the monthly counts of Incident reporting time dependent?
  3. Does Temperature really correlate to Incident Type?
  4. Does Temperature have linear correlation with Dewpoint?

We will answer questions through Hypothesis Testing. So we start from Hypothesis first.

- Hypothesis:

  In order to answer these questions, we need to perform validate our assumptions by hypothesis. We start from hypothesis respectively.

  Question 1:

  Null Hypothesis H0: Real-time Incident type is time dependent

  Alternative Hypothesis Hɑ : Real-time Incident Type is not time dependent

  Question 2:

  Null Hypothesis H0: Monthly counts of Incident reporting is time dependent

  Alternative Hypothesis Hɑ : Monthly counts of Incident reporting is not time dependent

  Question 3:

  Null Hypothesis H0: Temperature does correlate to Incident Type

  Alternative Hypothesis Hɑ : Temperature does NOT correlate to Incident Type

  Question 4:

  Null Hypothesis H0: Temperature has NO linear correlation with Dewpoint

  Alternative Hypothesis Hɑ : Temperature has linear correlation with Dewpoint

- Hypothesis Testing:

  Since Question 1 and 2 are related to time, Adfuller Testing is used. Question 3 is correlation testing, so we use Chi Square testing. For question 4, Pearson correlation is performed.

- Testing Results:

  Question 1: ɑ = 0.0 is less than 0.05 (significance level). So we reject Null Hypothesis H0 and accept Alternative Hypothesis, which means Real-time Incident Type is not time dependent

  Question 2: ɑ = 0.856808 is greater than 0.05. Null Hypothesis fails to be rejected. Real-time Incident Type is not time dependent

  Question 3: ɑ = [0.92368965 0.45178618 0.52189569 0.46904684 0.88481471 0.98657539 0.99999469 0.55348292] for each incident type is greater than 0.05. Null Hypothesis is accepted. Temperature does correlate with Incident Type.

Question 4: ɑ = 0.0 is less than 0.05 and correlation coefficient is 0.890731345258. That means Temperature has a linear correlation with Dewpoint