Submitted by:

shrey jain

662046332

jains6@rpi.edu

Guided by:

Prof. thilanka munasinghe

Analyzing Weather Impact on Ambulance Response Times

Final Project Report

Assignment 7

2022

Table of Contents

* Abstract and Introduction 2
* Data Description and Exploratory Data Analysis 3
* Analysis 5
* Model Development and Application 8
* Conclusion and Discussion 9
* References 11

1. Abstract and Introduction

In this world of rapid development, having a perfect health directly means leading a productive life. While prevention is better than cure, there are some cases where some serious, unexpected things happen which require immediate action. In those scenarios it is crucial to provide immediate responses to such emergencies as they could turn out be a lifesaver. Another such lifesaving action is to provide the pre-hospital care to any patient present in an ambulance or waiting for one. Therefore, having an accurate arrival time of ambulance could be very helpful.

New York City is the centre of all sorts of emergencies, because it is one of very crowded city in the United States of America. The first dataset is the weather dataset which is of the Manhattan borough in NYC. Emergency medical services data for the same borough is used to analyze trends and ultimately predict response time.

While the EMS data is sufficient to roughly predict the response times for situations that fall in similar categories, however, the aim of the project is also to highlight how co-related weather data is with EMS response data and how much performance gains do we get when we use additional weather attributes for model development.

This project covers process of building a machine learning model that predicts a response time of any ambulance in relation with the weather on that day. It also covers details about the datasets in focus, what are the steps that are to be followed to get the data ready for any model building or evaluation. This report also includes some of the plots which helps identify the relation between variables and finally contains the results and the conclusion drawn out from them.

All the references and resources used for making this project are included in the end of this project report.

1. Data Description and Exploratory Data Analysis

The project works on two datasets:

* EMS Dataset: This dataset is provided by Fire Department of New York City (FDNY). This dataset contains data for many regions hence it has 11863759 rows and 32 columns.
* Weather Dataset: This dataset is provided by Global Historical Climatology Network (GHCN). Since this dataset only contains weather data of Manhattan Borough, hence, it has comparatively very less rows as compared to EMS dataset. It only contains 3288 rows and 34 columns.

Weather on a particular date is used to link with the response times for emergencies that occurred on that date using the EMS datasets. Datasets are joined using INCIDENT\_DT column of EMS dataset and DATE column of Weather dataset. The target column to be predicted using various machine learning regression models is INCIDENT\_RESPONSE\_SECONDS\_QY which is present in the EMS dataset.

Before jumping to applying regression models it is very crucial to get data prepared, which means, applying data pre-processing techniques to make data ready for model development. The following data preparation steps (Figure 2.1) are applied:

* Quality Assessment: High level checks to determine if the data meets the required quality standards.
* Data Cleaning: Fixing incorrect, corrupted entries in the dataset.
* Data Munging: Modified or changing dataset beyond its original state.
* Exploratory Data Analysis: Performed initial investigation on the data to discover patterns, trends and to spot anomalies.
* Model Preparation: Applied 4 regression models on the pre-processed dataset.
* Model Evaluation: Evaluated the accuracy and performance of all these models and determined which works best for given set of data.

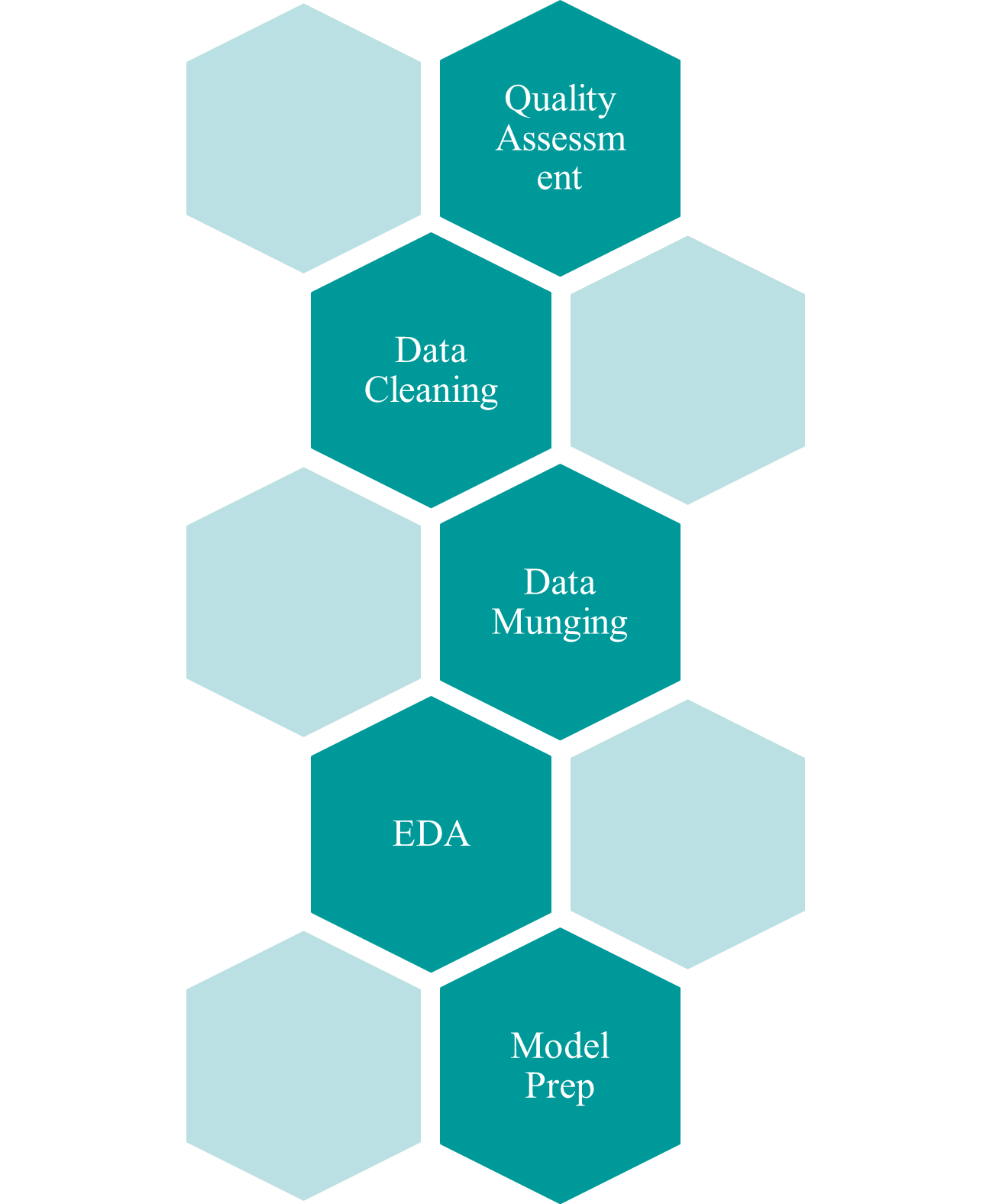


Figure 2.: Data pre-processing steps

On further analyzing the data, there are some key useful functions, which can help us give high help summary of the data variables. Example: *summary*, *fivenum, dim, etc*

* The summary function provides us very crucial information on the dependent variable that I want to predict. It seems there are about 0.3M N/A values in the dataset, just in the response time column itself.



Figure 2.: Summary of response time column

A picture containing text

Description automatically generated

Figure 2.3: Dimensions of EMS data with and without N/A values

1. Analysis

The following section contains various graphs and plots and how each of these helped in identifying trends and patterns in the data. Figure 3.1 shows a bar chart highlighting the number of EMS incidents that happened every year ranging from 2008-2016.

Chart, bar chart

Description automatically generated

Figure .1: Bar chart highlighting number of incidents by year

It can be inferred from Figure 3.1 that; number of incidents have a positive co-relation with the year. With each year, the incidents have gradually increase in number.

Chart, line chart

Description automatically generated

Figure 3.2: GG-Plot highlighting number of incidents by month

Additionally, from Figure 3.2, it can be inferred that, usually in the month of February and November, the incidents are comparatively lesser. On the other hand, there is a rise in number of incidents during summers/fall, in the month of June, July, and August.

Chart, bar chart

Description automatically generated

Figure 3.3: Bar chart highlighting total incidents with time split in different stages

It can also be seen from Figure 3.3 that, most of the time goes into transporting the patient. As we know that the response time is split across various stages which starts from the time the call was placed to the point where the patient was taken to the doctor. Transporting patients roughly takes about 70% of the total response time in case of an emergency.

Graphical user interface

Description automatically generated

Figure 3.: Chart highlighting total rainfall by year

On a separate note, for the Weather data Figure 3.4, depicts the total rainfall over the year, somewhere in 2011, the rainfall was at an abnormally high peak, which could potentially be an outlier.

* Outlier & Cleaning Stats:
* Outlier detection and removal were done using the IQR method.
* Removed about 0.22M outliers from the Weather and EMS dataset.
* Removed about 3.7M rows with N/A values from EMS dataset.
* Removed 3 columns from Weather dataset as they had majority of N/A values.
* Converted Date Format & String to numbers.

1. Model Development and Application

Before applying models, the datasets were merged using the data column. The final merged dataset obtained had about 0.36M rows. This number is after all data filtering and cleaning were done and dataset was ready for model building. Models performed significantly better on datasets without outliers and N/A values. Four regression models were applied which are Linear Regression, Support Vector Machine, Random Forest, and finally K-Nearest Neighbors. My additional hypothesis is that Hyper tuning will turn out to be an essential tool optimizing the model parameters and maximizing the performances of all these four models.

* **Linear Regression**

Before performing the linear regression, character literals were converted to integers, N/A values were removed, and attributes that have only one level (constants) were removed. After all the above removal, the final merged dataset had about 360K rows. Figure 4.1 and Figure 4.2 depicts histograms of the frequency of the values corresponding to the response times with and without N/A values respectively.

Graphical user interface, application

Description automatically generated

Figure 4.1: Histogram highlighting frequency of response times in the merged dataset with outliers

Chart, histogram

Description automatically generated

Figure 4.2: Histogram highlighting frequency of response times in the merged dataset without outliers

This proves one of our initial hypotheses that outlier removal is very important when building any regression model.

Moving on, while applying linear regression the following columns (as represented in Figure 4.3) were removed as they had constant values and in R if we build a regression model on columns with constant values it gives error.

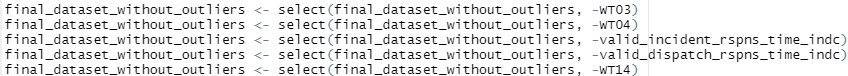


Figure 4.3: Five columns that had constant values and were removed before building regression model

The linear regression model was then built using the lm.fit() function. Figure 4.4 on the next page describes the output of the summary of the linear regression model.

* **Support Vector Regression**

In case of support vector machine, the same dataset without outliers was used which was used for Linear Regression. Before applying models, the datasets were merged using the data column. Models performed significantly better on datasets without outliers and N/A values. My hypothesis is again that Hyper tuning will turn out to be an essential tool optimizing the model parameters and maximizing the performances of Support Vector Regression.

Table

Description automatically generated

Figure 4.4: Summary of the Linear Regression Model

Conclusion and Discussion

References

* R Basics: <https://towardsdatascience.com/r-basics-everything-you-need-to-know-to-get-started-with-r-10c8e566d7b3>
* Scrape-R Package: <https://cran.r-project.org/web/packages/scrapeR/scrapeR.pdf>
* Read Excel File in R: <https://www.datacamp.com/community/tutorials/r-tutorial-read-excel-into-r>
* Dealing with missing data in R: <https://www.statmethods.net/input/missingdata.html>
* Visualization in R: <https://www.analyticsvidhya.com/blog/2015/07/guide-data-visualization-r/>
* KNN Tutorial: <https://quantdev.ssri.psu.edu/sites/qdev/files/kNN_tutorial.html>