KLE Society's

KLE Technological University



**Data Mining and Analysis Course Project Report On**

**Predicting response times of the Paris Fire Brigade vehicles**

***Under the guidance of***

**Dr P.G. Sunitha Hiremath**

**Submitted By**

|  |  |
| --- | --- |
| **Name** | **USN** |
| Shantanu Kumar | 01FE17BCS181 |
| Shivam Wadhera | 01FE17BCS189 |
| Satwik Belaldavar | 01FE17BCS178 |
| Samay D Naik | 01FE17BCS168 |

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

HUBLI – 580 031 (India)

Academic year 2019-20

|  |  |  |
| --- | --- | --- |
| **S.No** | **Table of Contents** | **Page No.** |
| 1 | Introduction | 3 |
| 2 | About the Challenge | 4 |
|  | 2.1-Problem Statement | 4 |
|  | 2.2-Data Description | 4 |
| 3 | Methodology | 5 |
|  | 3.1-Exploratory Data Analysis | 5 |
|  | 3.2-Data Preprocessing | 6 |
| 4 | Experimentation | 8 |
| 5 | Result | 10 |
| 6 | Conclusions | 10 |
| 7 | References | 11 |

**1.INTRODUCTION**

Response time is one of the most important factors for firefighters because their ability to save lives and rescue people depends on it. Every fire department in the world seeks strategies to decrease their response time, and several analyses have been conducted in the past years to determine what could impact response time. A non-optimal choice of an emergency vehicle for a rescue request may lengthen the arrival time of the rescuers and impact the future of the victim. This choice is therefore highly critical for emergency services and directly relies on their ability to predict precisely the arrival time of the different units available.

**Definition of Response Time**

Fire departments usually divide the response times of their firefighters in two main parts: turnout time, which corresponds to the seconds elapsed while firefighters prepare themselves at the station, and travel time, which refers to the time taken by the vehicle to arrive at the location of the incident.

In this project we address the above problem by investigate potential factors that may affect response time prediction. After​ the KDD process we will be able to identify the most important factors that contribute towards response time of Fire Brigade. With these attributes we will be able to suggest the Fire Brigade which kind of vehicle are better for intervention for locations present at long and short distances, also which vehicles are suited for a particular alert reason.

In this project we help the Fire Brigade community to work on the factors that affect the response time which could lead to better and faster response time thereby ensuring proper rescue operations.

**2.About the Challenge**

The Paris Fire Brigade is a French Army unit which serves as the primary fire and rescue service for Paris, the city’s inner suburbs and certain sites of national strategic importance. The dataset provided by The Paris Fire Brigade covers the entire year 2018 for which inoperable data have been squeezed out.

Start Date – June 10,2019

End Date – Dec 31,2020

**2.1 Problem Statement**

The task is to predict the delay between the selection of a rescue vehicle and the time when it arrives at the scene of the rescue request.

**2.2 Data Description**

Training Data (219337 lines)

* x\_train.csv
* y\_train.csv

Test data (108033 lines)

* x\_test.csv
* y\_test.csv

**3.Methodology**

**3.1 Exploratory Data Analysis**

Before preprocessing of Data, it is required to find out important attributes, missing values and plotting of attributes. The Dataset consisted of 534210 missing values which are distributed among four attributes (in percentage) as mentioned below.

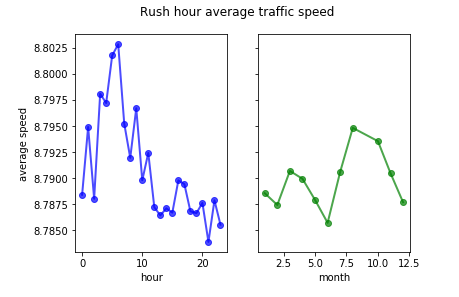
location of the event 5.794736

delta position gps previous departure-departure 97.764627

GPS tracks departure-presentation 69.998678

GPS tracks datetime departure-presentation 69.998678

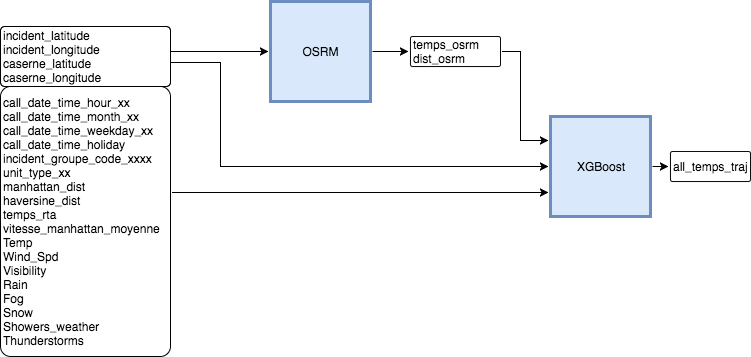
The Dataset didn’t contain any redundant tuples.



Inference:

* During the Day what time traffic is less and when its more – Fire Brigade can leave early if they can predict when the traffic is more.
* Traffic during Each Month.

**Overview of Response Time Prediction**



**3.2 Data Preprocessing**

The data has to be preprocessed before it can be fed to any learning model.

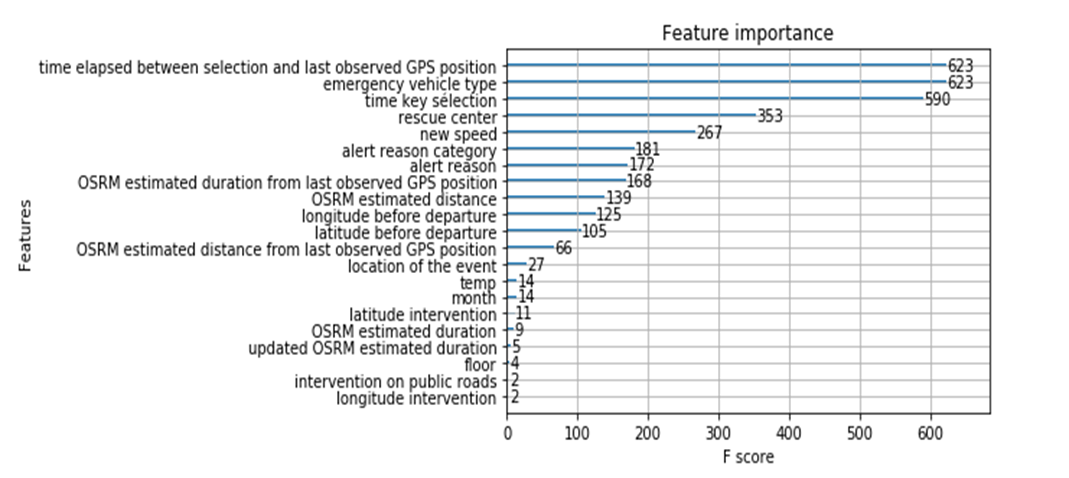
Handling of Missing Values

* Besides ‘location of the event’ attribute, the other three were dropped.
* ‘Location of the event’ attribute’s missing values were filled with the mode i.e. the most probable location id during intervention.

Reduction and Transformation

‘Status preceding selection’ attribute datatype was categorical which was converted to numerical using one-hot encoding.

Feature Importance:



Inference:

• From this we can take the required features for building our model and thus enhance the accuracy.

**Feature Engineering:**

* Month (Month of intervention)
* Hour (Hour of intervention)
* Average speed of different vehicle type
* Temperature (Temperature of each day of month)
* OSRM estimated distance from last observed GPS position
* OSRM estimated duration from last observed GPS position
* Time elapsed between selection and last observed GPS position
* Updated OSRM estimated duration

Now we have obtained the data that we could use for model building in the next step.

The X\_train data after preprocessing was split into a Training and Validation sets with a split ratio of 20%:

* xTrain
* xTest
* yTrain
* yTest

The xTrain and yTrain data would be used for the model building and the xTest and yTest would be used for the final evaluation of the model after ensembling.

We develop 3 models which would later be ensembled to obtain the final model.

We use the X\_train and y\_train original files and further split them with a validation set ratio of 20%.

After building the final model, each model is trained on the entire dataset for further testing on the X\_test and y\_test original files. This is similar to cross-validation.

**4.Experimentation**

**Model 1: Linear Regression**

Linear regression is used for finding linear relationship between target and one or more predictors.

We took two attributes of y\_train and trained each attribute with selected attributes of x\_train and then added both the prediction.

**Model 2: XGBoost**

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. Gradient Boosted decision trees consist of building a simple model and then cycle consists of Calculate Errors, Build Model Predicting Errors, Add Last Model to Ensemble.

XGBoost has a few parameters that can dramatically affect model’s accuracy and training speed.

These are n\_estimators, early\_stopping\_rounds and learning\_rate.

Too low value of n\_estimators can cause underfitting, Too high value causes overfitting. Typical values range from 100 – 1000.

Early\_stopping\_rounds cause the model to stop iterating when the validation score stops improving. For this n\_estimators can be set high.

Learning rate allows use of higher value of n\_estimators without overfitting. In general, a small learning rate (and large number of estimators) will yield more accurate XGBoost models but will take longer to train the model.

XGBoost Regressor yielded a higher accuracy and hence is used as the final model.

**Model 3: Random Forest**

Random forests, also known are a popular ensemble method that can be used to build predictive models. The model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer.

The random forest uses many trees, and it makes a prediction by averaging the predictions of each component tree.

**Delta selection departure**

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean Squared Error** | **R2 Score** |
| Linear Regression | 40.844 | 0.041975 |
| XG Boost | 37.530 | 0.07505 |
| Random Forest | 36.913 | -0.00901 |

**Delta departure-presentation**

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean Squared Error** | **R2 Score** |
| Linear Regression | 101.2738 | 0.40229 |
| XG Boost | 89.715 | 0.5171 |
| Random Forest | 86.464 | 0.44367 |

**5.Result**

From the tables we can infer that XG Boost algorithm is better model for our dataset.

Public Score – 0.2286

**6.Conclusion**

* Calculate feature importance of attributes.
* Gradient Boosting algorithm was the best for feature selection.
* Observed comparatively less Error and higher accuracy in XG Boost compared to other Models.

**7.References**

# Using Data Science to Predict Response Times of Firefighters - https://medium.com/crim/predicting-the-response-times-of-firefighters-using-data-science-da79f6965f93

# How Uber Engineers an Efficient Route - <https://eng.uber.com/engineering-an-efficient-route/>

* Kaggle Challenge: New York City Taxi Trip Duration -

<https://www.kaggle.com/c/nyc-taxi-trip-duration/notebooks>