KLE Technological University
Creating Value
Leveraging Knowledge

Data Mining and Analysis Course Project (18ECSC301)

Predicting response time of the Paris Fire Brigade Vehicle
Project Id 5DMACP16 Team A4

Course Instructor : Prof. Neha T

Name	USN
Aishwarya.G.M	01FE17BCS016
Yashas J	01FE17BCS244
Akash Bilgi	01FE17BCS020
Akhila.A.G	01FE17BCS022

KLE Technological University
Creating Value
Leveraging Knowledge

About the Challenge

- The response time is one of the most important factors for emergency services because of their ability to save lives depends on it
- If a non optimal choice of emergency vehicle for a rescue request is made which may lengthen the arrival time of the rescuers and can even impact the victim's condition
- Hence this choice directly depends on their ability to predict precisely the arrival time of different emergency vehicles available
- The main goal of this challenge is to predict delay between selection of rescue vehicle and the time it arrives at the scene of rescue request
- Challenge Duration :Challenge opening date:8th July 2019 Challenge closing date:31st December 2020

This challenge is chosen from:https://challengedata.ens.fr/challenges/21

KLE Technological University
Creating Value
Leveraging Knowledge

Problem Statement

Design a model which predicts the delay between selection of rescue vehicle(The time when a rescue team is alerted) and the time it arrives at the scene of rescue request.

KLE Technological University
Creating Value
Leveraging Knowledge

Dataset Description

Training Data:

Size of the data = 200 MB

Number of attributes =26

Number of Tuples=219338

x_train.csv,y_train.csv

Test Data:

Size of the data = 98.2 MB

Number of attributes =26

Number of Tuples=108034

x_test.csv

```
emergency_vehicle_selection                    int64
intervention                                   int64
alert_reason_category                          int64
alert_reason                                   int64
intervention_on_public_roads                   int64
floor                                          int64
location_of_the_event                        float64
longitude_intervention                       float64
latitude_intervention                        float64
emergency_vehicle                              int64
emergency_vehicle_type                        object
rescue_center                                  int64
selection_time                                object
date_key_selection                             int64
time_key_selection                             int64
status_preceding_selection                    object
delta_status_preceding_selection_selection     int64
departed_from_its_rescue_center                int64
longitude_before_departure                   float64
latitude_before_departure                    float64
delta_position_gps_previous_departure_departure float64
GPS_tracks_departure_presentation             object
GPS_tracks_datetime_departure_presentation    object
OSRM_response                                 object
OSRM_estimated_distance                      float64
OSRM_estimated_duration                      float64
dtype: object
```

List of attributes

- Emergency vehicle selection : Identifying of selection instance of an emergency vehicle for an intervention
- Alert reason category (category):It has various categories from 1 to 9
- Intervention on public roads(boolean) : whether the intervention is on public road or not
- Floor (int):floor in which intervention happened
- Location of the event(category) :it defines where the event has taken place ex: entrance hall , boiler room , motorway etc
- Longitude intervention(float) : approximate longitude of the intervention address
- Latitude intervention (float) :approximate latitude of the intervention address
- Emergency vehicle type (category) :type of the emergency vehicle

- Rescue centre (category) :identification of rescue centre to which it belongs
- Date key selection (int) :selection date in YYYYMMDD format
- Time key selection (int) :selection time in HHMMSS format
- Status preceding selection: status of the emergency vehicle prior selection
- Departed from its rescue centre (boolean) :whether the vehicle has departed from its rescue centre or not
- OSRM estimated distance (float) :distance calculated by the OSRM route service
- OSRM estimated duration (float) :transit delay calculated by the OSRM route service

Exploratory Data Analysis

- emergency_vehicle_type VS OSRM_estimated_duration

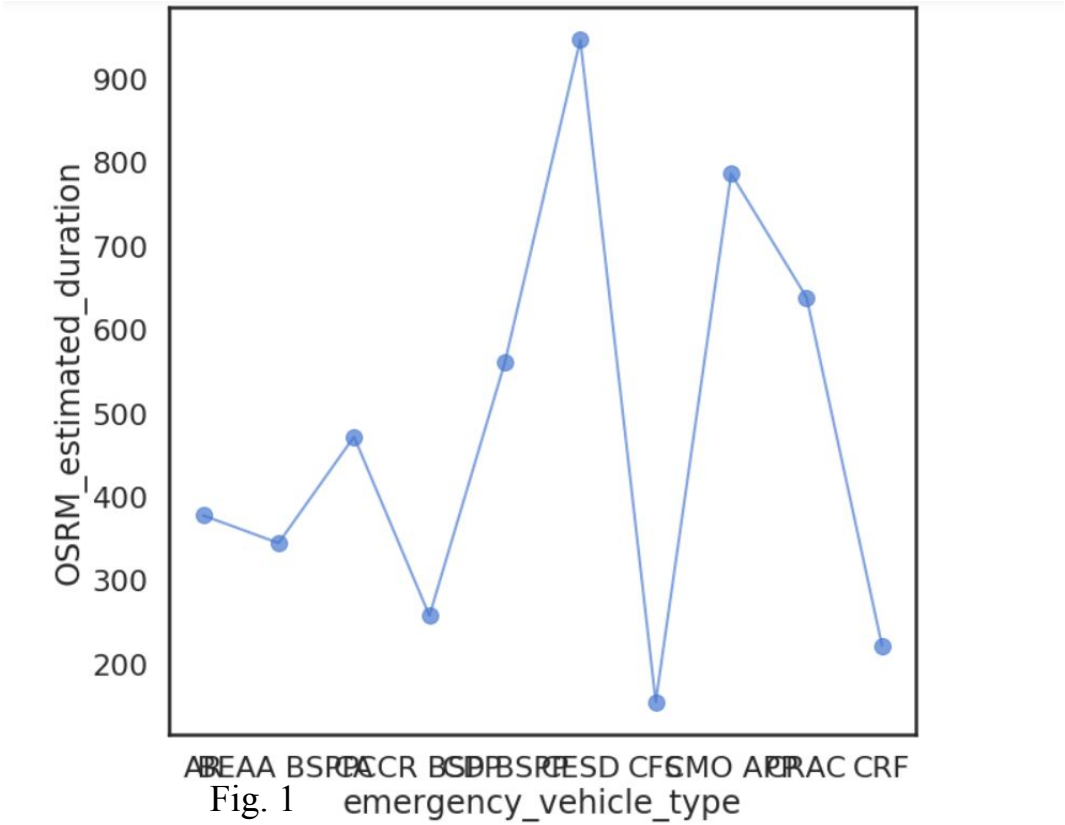


Fig. 1

- Inference –For CESD emergency_vehicle_type OSRM_estimated_duration in high

Frequency of the accidents vs Month

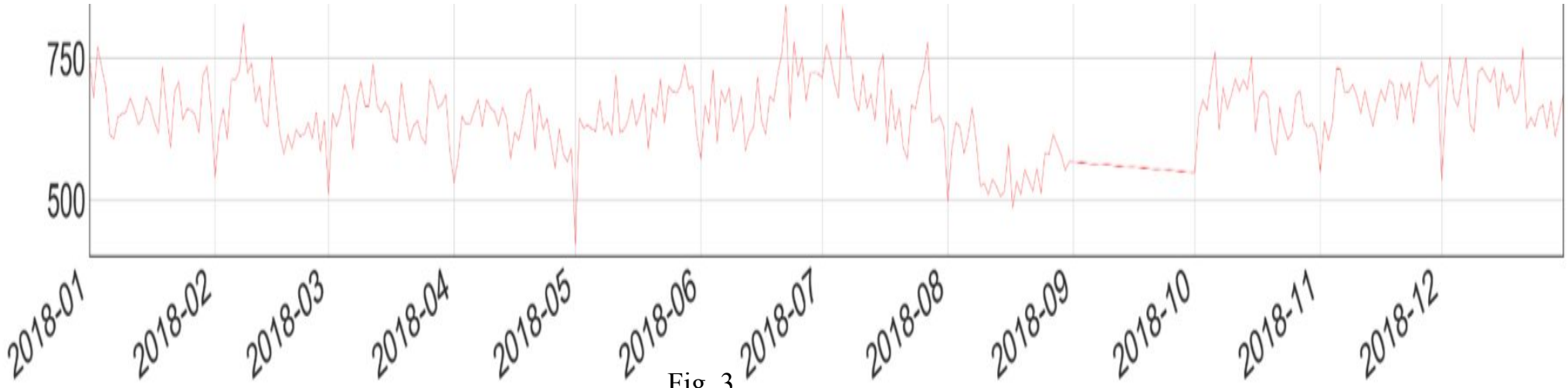


Fig. 3

- Inference –Most of the accidents where recorded during July

- Frequency of accident vs day of week

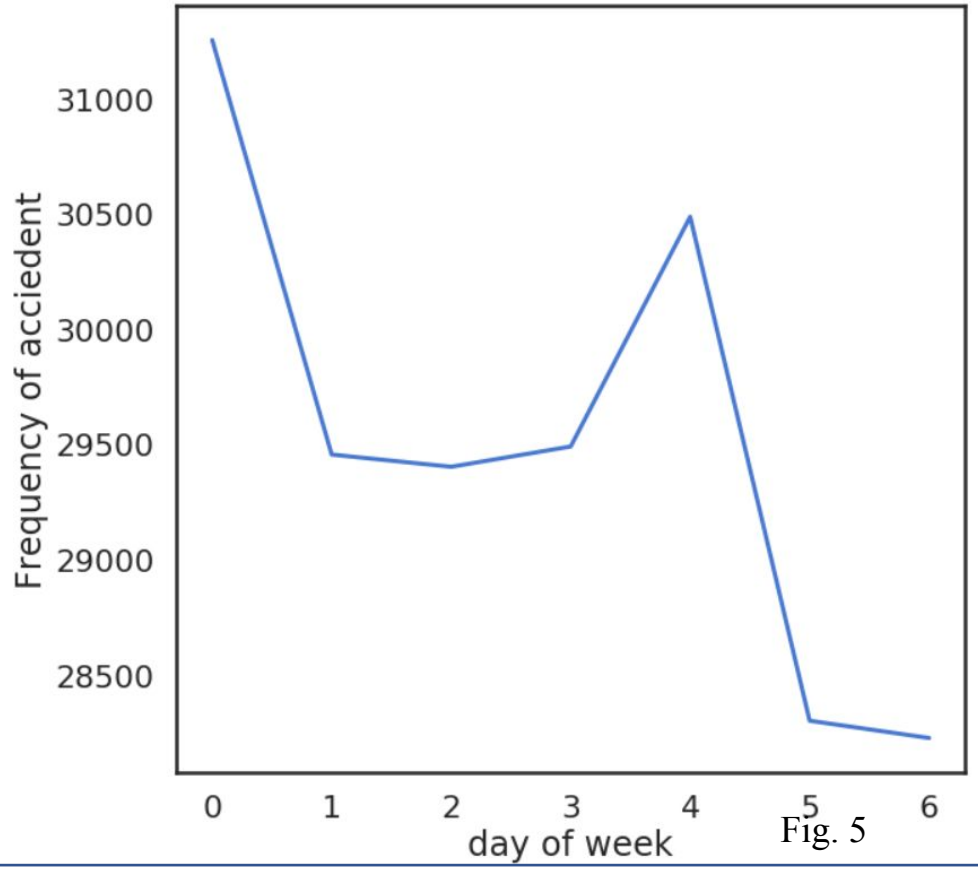


Fig. 5

- Inference –Most accident took place on Sunday.

Data Preprocessing

Data Cleaning

Finding null values in each column

```
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
OSRM_estimate_from_last_observed_GPS_position           69.998678
OSRM_estimated_distance_from_last_observed_GPS_position 70.018738
OSRM_estimated_duration_from_last_observed_GPS_position 70.018738
time_elapsed_between_selection_and_last_observed_GPS_position 69.998678
updated_OSRM_estimated_duration                         70.018738
dtype: float64
```

Data Preprocessing

Data Cleaning

- Filling nan values of location_of_the_event column with 139 (most frequent location).
- Dropped delta_position_gps_previous_departure_departure because it had 97% NULL tuples.
- Dropped GPS_tracks_departure_presentation because it had 70% NULL tuples.
- Dropped GPS_tracks_datetime_departure_presentation because it had 70% NULL tuples.
- Null values of updated_OSRM_estimated_duration is filled with previous OSRM_estimated_duration .
- Null values of OSRM_estimated_distance_from_last_observed_GPS_position is filled with previous OSRM_estimated_distance.
- Null values of OSRM_estimated_duration_from_last_observed_GPS_positionis filled with previous OSRM_estimated_duration.
- Inference – There are no null values in the dataset.

KLE Technological University
Creating Value
Leveraging Knowledge

Data Preprocessing

Data Cleaning

- Dropping date_key_selection after extracting month and weekday.
- Dropping time_key_selection after extracting hour .

- Column having ID's

1. OSRM_response.
2. emergency_vehicle_selection.
3. Intervention.
4. emergency_vehicle.

- Inference – Dropping above column as they are unique values and does not make any contribution in prediction.

Data Transformation

Convertion of categorical to numeric values

Label Encoding

- This technique refers to transforming the word labels into numerical form so that the algorithms can understand how to operate on them.

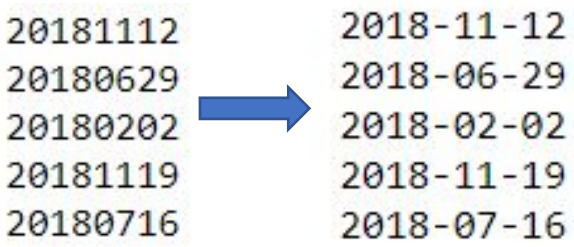
Inference –

- Alert_reason_category is converted into numeric form
- Month is converted into numeric form
- weekday is converted into numeric form
- location_of_the_event is converted into numeric form

Data Transformation

- Converting time_key_selection (int) to date time format.
- Extracting Hour from time_key_selection.

- Converting date_key_selection(int) to date time format.

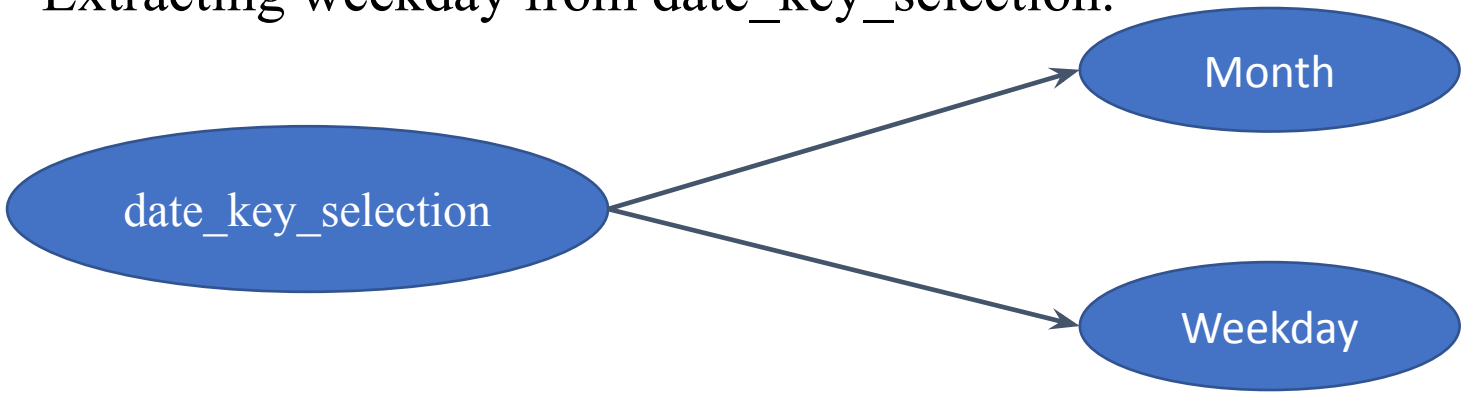


- Creating speed feature (speed=distance/time).

- Inference – New attributes are constructed from the given set of attributes.

Data Transformation

- Extracting month from date_key_selection.
- Extracting weekday from date_key_selection.



- Dropping selection_time which contain date_key_selection.

- Inference – New attributes are constructed from the given set of attributes.

Building Learning Model

1. Linear Regression

- Linear regression is useful for finding relationship between two continuous variables
- To model the relationship between variables by fitting a linear equation to observed data

R2 score for test data 0.0848

R2 score for train data 0.1666

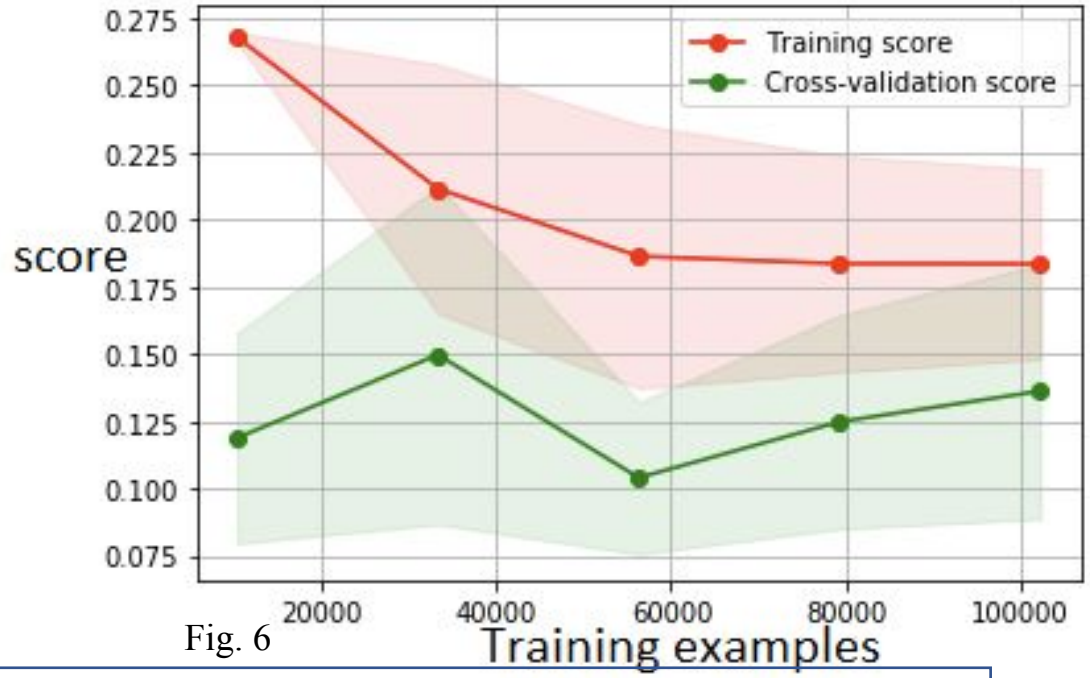


Fig. 6

- Inference –Model is to much overfitting and there is high variance.

2. Lasso Regression

- Lasso regression analysis is a shrinkage and variable selection method for linear regression methods
- It is generally used when we have more number of feature as it automatically does the feature selection

R2 score for test data 0.105

R2 score for train data 0.15

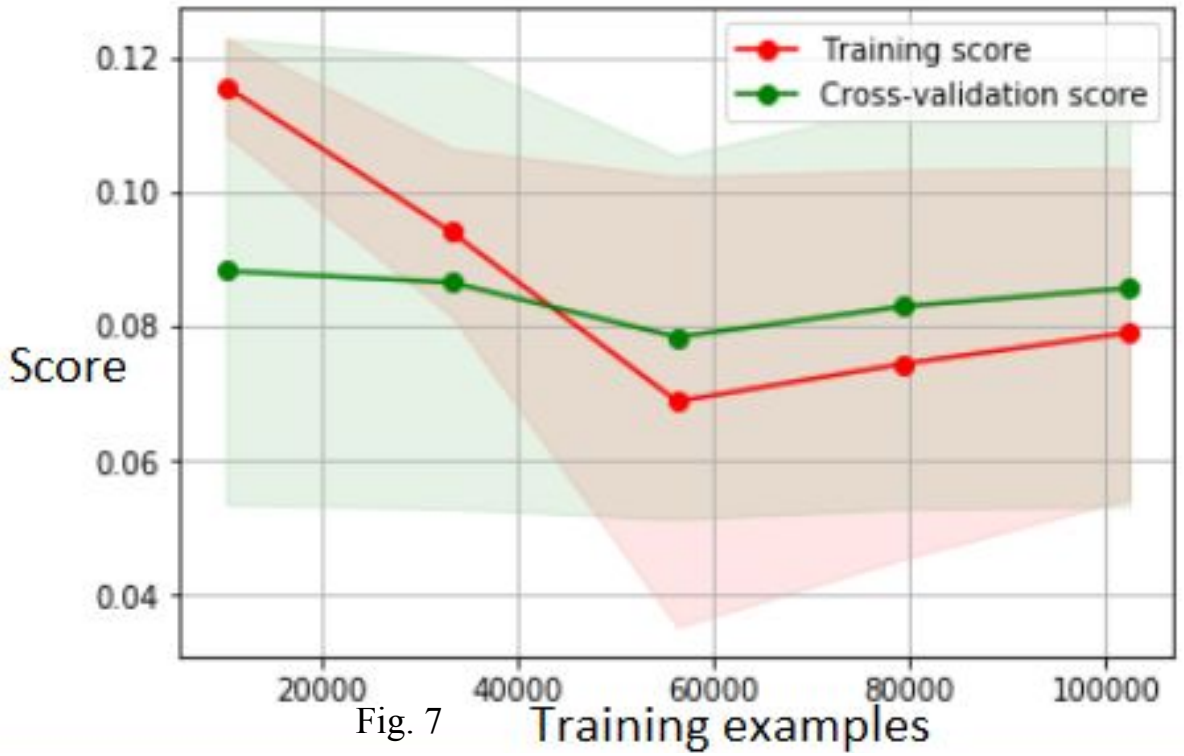


Fig. 7

Inference –After 60000 training examples model is underfitting and there is variance.

3. Ridge Regression

- Ridge regression is a technique for analysing multiple regression data that suffers from multicollinearity ie: the existence of non linear relationships among independent variables
- It reduces the model complexity by coefficients shrinkage
- It uses L2 regularization technique

Fig. 8

R2 score for test data 0.125

R2 score for train data 0.1482

Inference –After 60000 training examples model is underfitting but variance is low.

4.Light GBM

- Light GBM is a gradient boosting framework that uses tree based learning algorithm.
- Light GBM grows tree vertically while other algorithm grows trees horizontally meaning that Light GBM grows tree leaf-wise while other algorithm grows level-wise.

R2 score for test data 0.1806

R2 score for train data 0.292

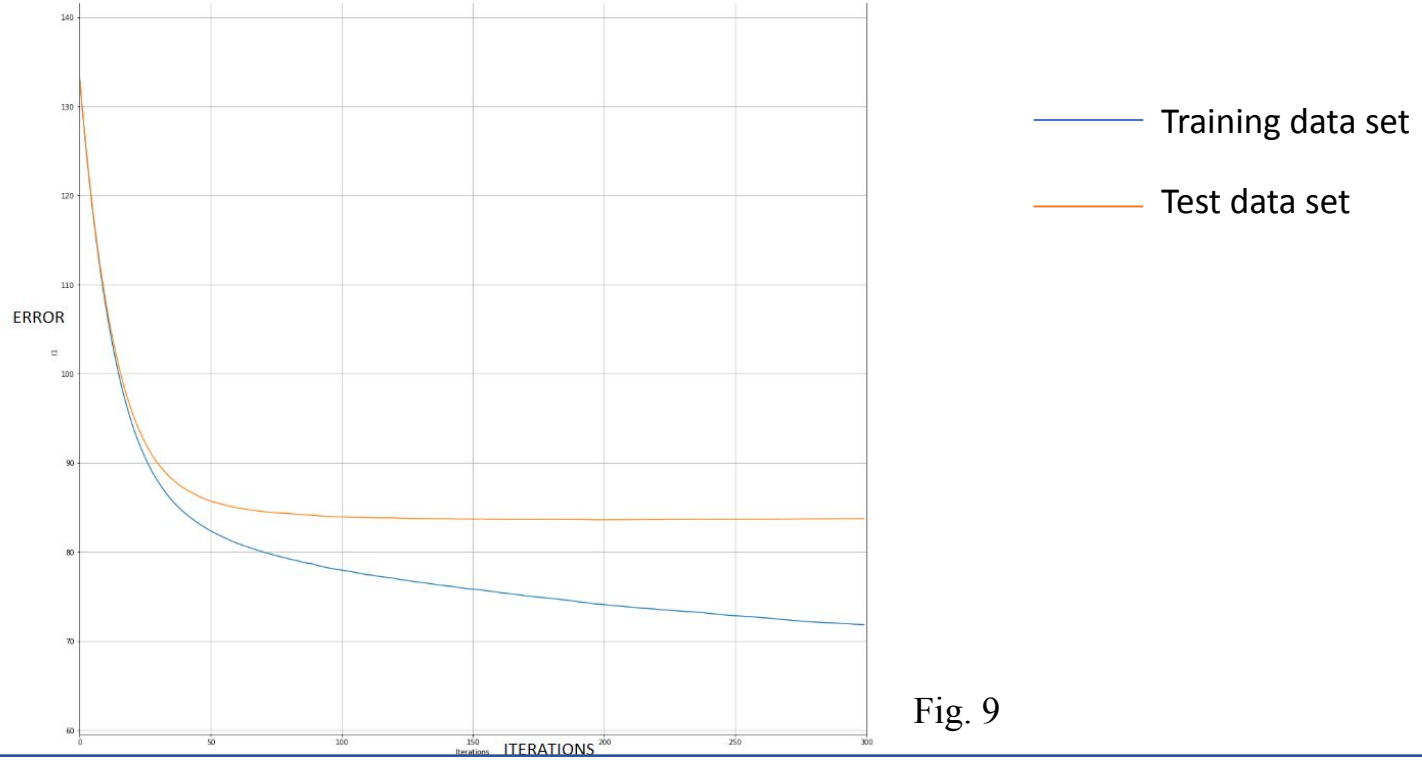


Fig. 9

Inference-After 100 iteration error of test data set remain constant but error of train data set is decreasing which leads to overfitting of model .

Classification Techniques and their accuracy

Sl.No	Model	R2 score
1	Linear Regression	0.0848
2	Lasso Regression	0.105
3	Ridge Regression	0.125
4	Light GBM	0.1806

Inference
Accuracy of Light GBM is highest of all the regression techniques Hence this model has been selected.

Leaderboard Ranking

R2 score	Rank	Total Competitors
0.15	30	35
0.1812	19	48
0.2204	5	70
0.2687	2	120
0.2703	10	160

Conclusion

Our objective was to predict the delay between selection of rescue vehicle and the time it arrives at the scene of rescue request and we achieved the highest accuracy through Light GBM model ie the accuracy of 0.2703 .

Reference

- Dataset description
- https://paris-fire-brigade.github.io/data-challenge/challenge.html
- https://challengedata.ens.fr/participants/challenges/21/
- https://medium.com/crim/predicting-the-response-times-of-firefighters-using-data-science-da79f6965f93
- Implemented example
- _https://eng.uber.com/engineering-an-efficient-route/
- Encoding Categorical Feature
- https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html
- https://www.youtube.com/watch?v=9yl6-HEY7_s
- http://www.datasciencemadesimple.com/convert-column-to-categorical-pandas-python-2/

Reference

- Light GBM
- https://lightgbm.readthedocs.io/en/latest/
- https://medium.com/@pushkarmandot/https-medium-com-pushkarmandt-
- Why Light GBM
- https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/
- learning curve
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.learning_curve.html
- https://www.kaggle.com/tobikaggle/humble-lightgbm-starter-with-learning-curve

THANK YOU