KLE Society's

KLE Technological University

Data Mining and Analysis Course Project Report

On

Predicting response time of the Paris Fire Brigade Vehicles

Under the guidance of

Prof. Neha T

Submitted By

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

SCHOOL OF COMPUTER SCIENCE & ENGINEERING,
HUBLI – 580031 (India).

Academic year 2019-20

KLE Society's
KLE Technological
University 2018 - 2019

SCHOOL OF COMPUTER SCIENCE &
ENGINEERING

CERTIFICATE

This is to certify that Data Mining and Analysis project entitled “Predicting response time of the Paris Fire Brigade Vehicles” is a bona fide work carried out by the student team Mr. Yashas J – 01FE17BCS244, Mr. Akash B A– 01FE17BCS020, Ms. Akhila. A.G – 01FE17BCS022, Ms. Aishwarya.G.M – 01FE17BCS016, in partial fulfilment of completion of Fifth semester B. E. in Computer science and Engineering during the year 2019– 2020. The project report has been approved as it satisfies the academic requirement with respect to the project work prescribed for the above said course.

Guide	**Head of SoCSE**
Prof. Neha T.	**Dr. Meena S. M**

External Viva:

Name of the Examiners	**Signature with date**
1.	
2.	

ABSTRACT

This project focuses on to predict the delay between the selection of a rescue vehicle (the time when a rescue team is warned) and the time when it arrives at the scene of the rescue request. In this course we mainly focus on the prediction of response time related to Paris Fire Brigade. The predictions are to be made based on the previous month's observations, i.e. prediction of year 2018 from Jan to Dec. The data analysis offers us an opportunity to develop new methods in the field of prediction of such kind of events. The project uses different methods such as Ridge Lasso and Linear regression the details of which are described in subsequent segments

Acknowledgement

We have been bestowed the privilege of expressing our gratitude to everyone who helped us in completing the dissertation work.

We sincerely thank our guide Prof. Neha T for her guidance, inspiration and wholehearted cooperation during the course of completion.

We sincerely thank Prof. Shankar. S, Department of Computer Science and Engineering, Prof. P. G. Sunitha Hiremath, Department of Computer Science and Engineering for their support, inspiration and wholehearted co-operation during the course of completion.

We also take this opportunity to thank Dr. Meena S M, Professor and Head of Department, Department of Computer Science and Engineering for having provided us with an academic environment which nurtured our practical skills contributing to the completion of our project.

Finally, we thank one and all who have directly and indirectly assisted us in the project work.

Aishwarya.G.M	01FE17BCS016
Yashas J	01FE17BCS244
Akash B	01FE17BCS020
Akhila.A.G	01FE17BCS022

1.INTRODUCTION

In Emergency services the fire brigade plays a very important role because their ability to save lives and rescue people depends on it , and hence The Fire services in France are organized and the Paris Fire Brigade(Army) holds the data related to the underlying causes of these events.

In an effort to provide early response time, the Paris Fire Brigade programs machine learning techniques and OSRM response using publicly available data sources to predict a good response and an optimal choice of emergency vehicle with a high degree of accuracy.

The Paris Fire Brigade seeks innovative solutions and good response with optimal choice of vehicle for the rescue request

The specific event classes of interest are:

The response times of the Paris Fire Brigade vehicles which is the delay between:

- the selection of a rescue vehicle (the time when a rescue team is warned)
- and the rescue team arrival time at the scene of the request (information sent manually via portable radio)

This measurement is composed by the 2 following periods of time:

- the activation period of the rescue team
- the transit time of the rescue team

1.1Literature Survey

The **Paris Fire Brigade** (French: Brigade des sapeurs-pompiers de Paris, **BSPP**) 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 brigade's main area of responsibility is the City of Paris and the surrounding départements of Seine-Saint-Denis, Val-de-Marne, and Hauts-de-Seine (the petite couronne). It also serves the Centre Spatial Guyanais in Kourou, the DGA Military Rocket Test Centre in Biscarosse, and the Lacq gas field. As with the other fire services of France, the brigade provides technical rescue, search and rescue and fire prevention services, and is one of the providers of emergency medical services.

The brigade is one of two fire services in France that is part of the armed forces, with the other being the Marseille Naval Fire Battalion (BMPM). It is a unit of the French Army's Engineering Arm (l'arme du génie) and the firefighters are therefore sappers (sapeurs, thus sapeurs-pompiers). With 8,550 firefighters, it is the largest fire service in Europe and the third largest urban fire service in the world, after the Tokyo Fire Department and New York City Fire Department. Its motto is "Save or Perish"

The Paris fire Brigade was Founded in 1793 as the Corps des gardes-pompes de la ville de Paris and following the 23-hour Austrian Embassy Fire in 1810 became a military organisation by imperial decree of Emperor Napoleon. On 18 September 1811, it became the Bataillon de sapeurs-pompiers de Paris and was expanded to the Régiment de sapeurs-pompiers de Paris in 1867. On 1 March 1967 became the Brigade des sapeurs-pompiers de Paris.

The operational personnel (hommes du rang i.e. privates) are usually engaged for five years. They must have French nationality, be between 18 and 25 years old, have a clean criminal record and have at least a vocational training CAP certificate. The selection is three days long, with sports tests, psychomotor tests, medical examination, etc.

Training takes place in the Instruction Grouping (Groupement d'instruction, GI), at the fort of Villeneuve-Saint-Georges. The first period lasts two months, with the first aid and first responder training, and basic military instruction (including shooting). They then undertake practical training of four months in an operational fire company (compagnie d'incendie); this includes taking part in personal assistance and utility safety operations. The last stage of training is a further two months at the Instruction Grouping. Upon completing training, the firefighter joins a fire company.

1.1 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.

1.2 Objectives

1.To develop models that predicts the delay between selection of rescue vehicle and the

time it arrives the scene of rescue request

2.To develop a data mining technique to approach the problem.

3.Selection of relevant attributes using relevant techniques.

4. Compare the performance of different classification techniques

1.3 Data Description

The dataset contains information of 2018's data

- Training Data:
  - Size of the data = 200 MB
  - Number of attributes =26
  - Number of Tuples=219338
- Test Data:
  - Size of the data = 98.2 MB
  - Number of attributes =26

- Number of Tuples=108034

List of Attributes and its description

- 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

2 RELATED WORKS

1) Using Data Science to predict Response Times of Firefighters

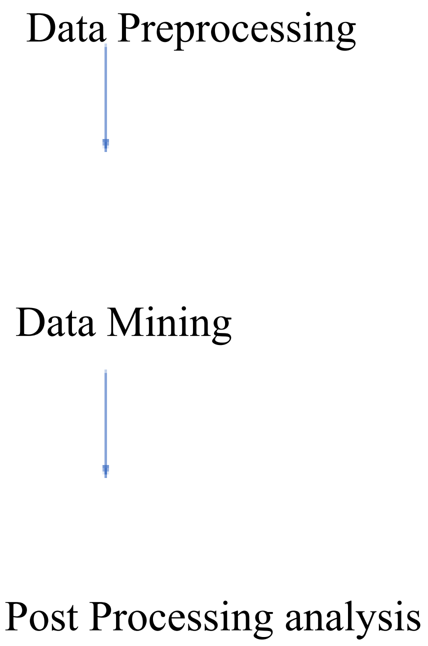
- They present a data science framework to predict turnout time and travel time and demonstrate its efficiency using data from the fire department of Montreal.

2) ETA Phone Home: How Uber Engineers an Efficient Route

- They used Ultra-efficient route planning and highly accurate ETAs which are critical. and they even used modern routing algorithms to build a carefully optimized system capable of handling hundreds of thousands of ETA requests per second, with single-digit milliseconds response time. All of their new services, such as uberPOOL and UberEATS start with this system. The importance of accuracy and efficiency to contextually determine the best way to get somewhere, and when, will continue to rise as we expand and improve products like uberPOOL and beyond. They're going to make phoning home even more reliable and intelligent.

3 METHODOLOGY & RESULTS

The approach towards the solution for predicting response time was as follows:



Data Preprocessing

Data Mining

Post Processing analysis

The first step towards the solution is to pre-process the data. The pre-processing of data involves:

1. Removing Id's in dataset.
2. Filling in the NULL values in every attribute.
3. Checking for Outliers in the dataset.
4. Removal of less relevant attributes.
5. Checking the relationship between different attributes of the dataset by plotting.
6. Conversion of categorical data to numerical data format by get dummies function.

The Second step is the Data mining which involves:

1. Analysis of data.
2. Building of a prediction model.
3. Prediction.

The third step involves analysis of the outputs and checking the correctness of the predictions.

3.1 DATA PREPROCESSING

Step -1 : **Data Cleaning**

- Dropping date_key_selection after extracting month and weekday.
- Dropping time_key_selection after extracting hour .
- Column includes Id's like emergency_vechicle_selection , intervention , emergency_vehicle which need to be dropped .
- Certain attributes contained NULL values; these values were replaced with relevant data or dropped them due to too many null values.
- For example,: 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.

Step – 2 : **Data Transformation**

- Convertion of categorical to numeric values
  It is done through

Label Encoding

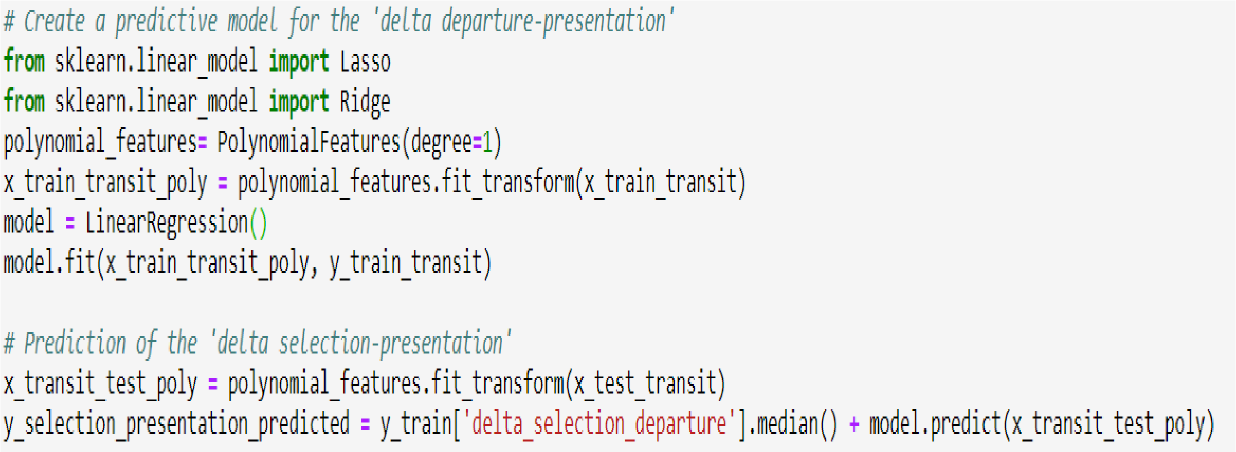
- This technique refers to transforming the word labels into numerical form so that the algorithms can understand how to operate on them.

- Ex in this dataset : 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
- 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).
- Extracting month from date_key_selection.
- Extracting weekday from date_key_selection.
- Dropping selection_time which contain date_key_selection. now from these steps – New attributes are constructed from the given set of attributes to help the mining process.

3.2 GETTING THE MODEL READY

Iteration -1 : Linear regression

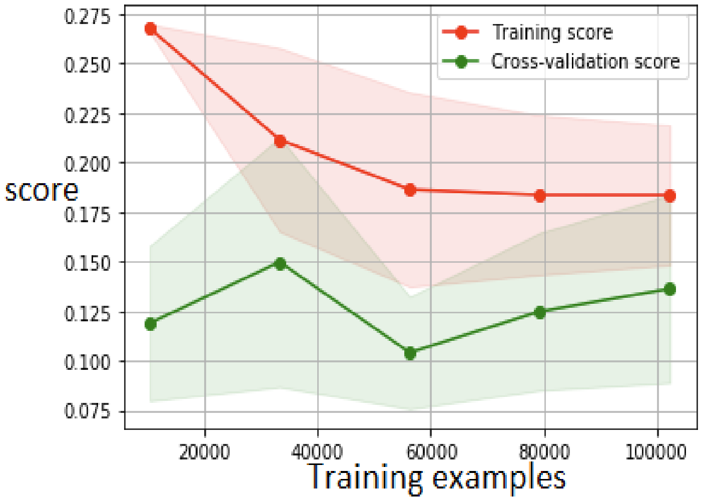
In this iteration we tried to use Linear regression . To model the relationship between two variables by fitting a linear equation to observed data.



```
# Create a predictive model for the 'delta departure-presentation'
from sklearn.linear_model import Lasso
from sklearn.linear_model import Ridge
polynomial_features= PolynomialFeatures(degree=1)
x_train_transit_poly = polynomial_features.fit_transform(x_train_transit)
model = LinearRegression()
model.fit(x_train_transit_poly, y_train_transit)

# Prediction of the 'delta selection-presentation'
x_transit_test_poly = polynomial_features.fit_transform(x_test_transit)
y_selection_presentation_predicted = y_train['delta_selection_departure'].median() + model.predict(x_transit_test_poly)
```

Figure 1: Linear regression model



Results

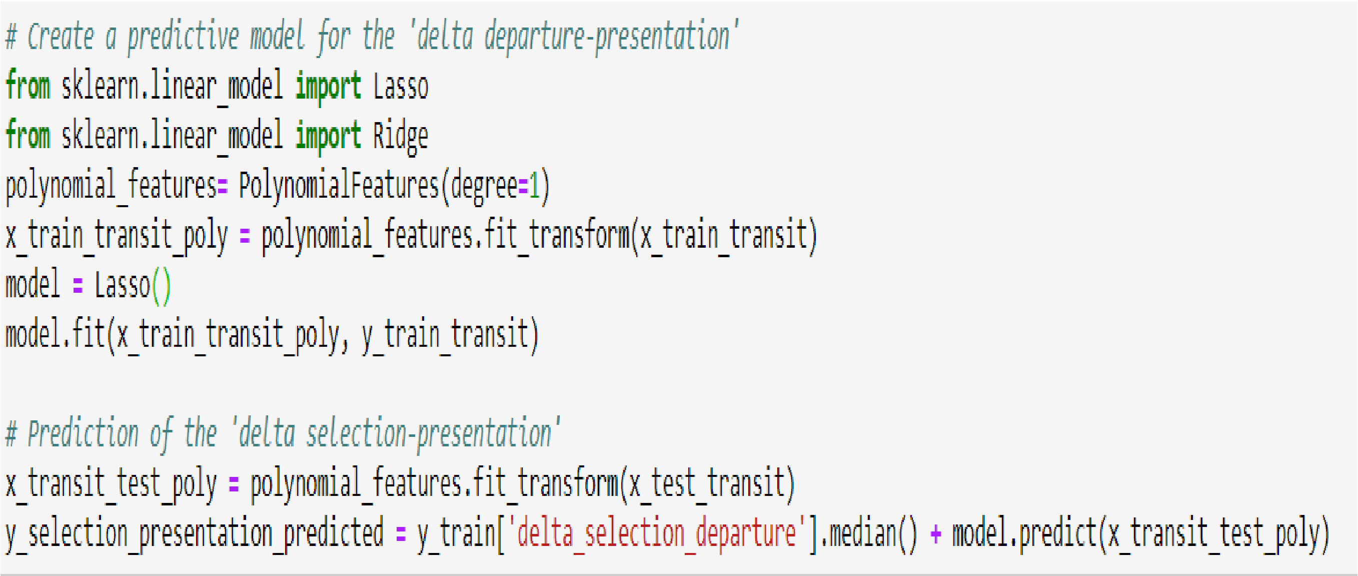
R2 score of delta_selection_departure for train data is 0.166
R2 score of delta_selection_departure for test data is 0.125

Inference

Model is too much overfitting and there is high variance.

Iteration -2 Lasso regression

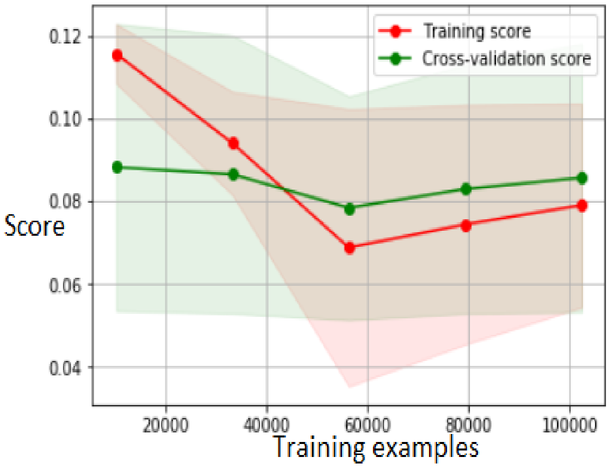
In this iteration we tried to use L1 regularization technique. It is generally used when we have more number of feature. Because it automatically does the feature selection and it is a shrinkage and variable selection method for linear regression methods.



```
# Create a predictive model for the 'delta departure-presentation'
from sklearn.linear_model import Lasso
from sklearn.linear_model import Ridge
polynomial_features= PolynomialFeatures(degree=1)
x_train_transit_poly = polynomial_features.fit_transform(x_train_transit)
model = Lasso()
model.fit(x_train_transit_poly, y_train_transit)

# Prediction of the 'delta selection-presentation'
x_transit_test_poly = polynomial_features.fit_transform(x_test_transit)
y_selection_presentation_predicted = y_train['delta_selection_departure'].median() + model.predict(x_transit_test_poly)
```

Figure 3: Lasso regression model



Results

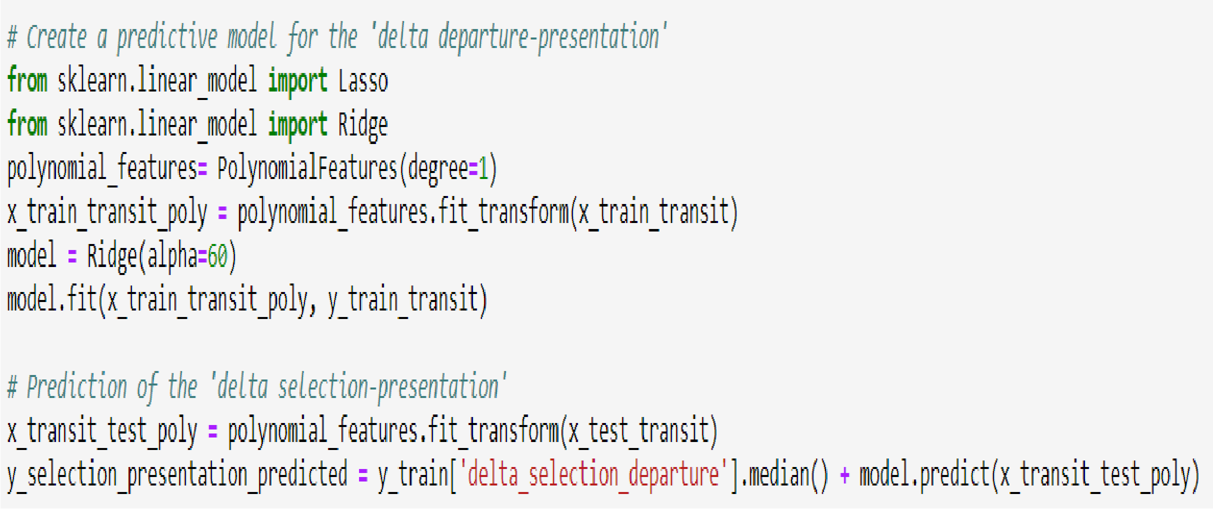
R2 score of delta_selection_departure for train data is 0.15
R2 score of delta_selection_departure for test data is 0.105

Inference

- Lasso regression lead to feature selection.
- After 60000 training examples model is underfitting and there is variance.

Iteration -3 Ridge regression

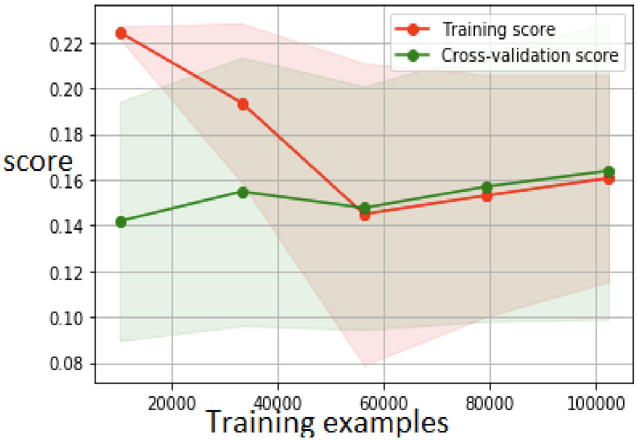
It is a regression technique that shrinks the parameters. Therefore, it is mostly used to prevent multicollinearity. It reduces the model complexity by coefficients shrinkage. It uses L2 regularization technique.



```
# Create a predictive model for the 'delta departure-presentation'
from sklearn.linear_model import Lasso
from sklearn.linear_model import Ridge
polynomial_features= PolynomialFeatures(degree=1)
x_train_transit_poly = polynomial_features.fit_transform(x_train_transit)
model = Ridge(alpha=60)
model.fit(x_train_transit_poly, y_train_transit)

# Prediction of the 'delta selection-presentation'
x_transit_test_poly = polynomial_features.fit_transform(x_test_transit)
y_selection_presentation_predicted = y_train['delta_selection_departure'].median() + model.predict(x_transit_test_poly)
```

Figure 5: Ridge regression model



Results

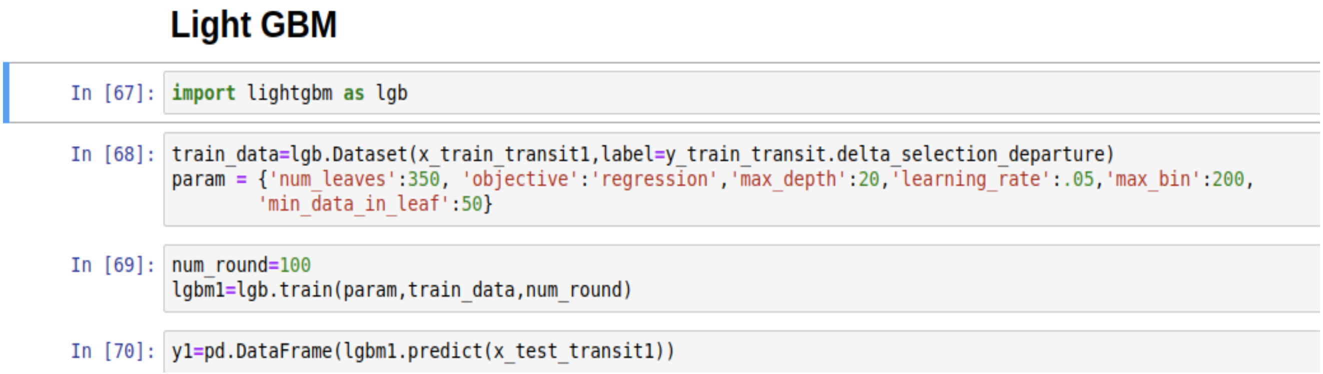
R2 score of delta_selection_departure for train data is 0.1482
R2 score of delta_selection_departure for test data is 0.125

Inference

- R2 score increases as alpha is increased.
- Ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity
- After 60000 training examples model is underfitting but variance is low.

Iteration -4 Light GBM

- Light GBM is a gradient boosting framework that uses tree based learning algorithm. And 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.



Light GBM

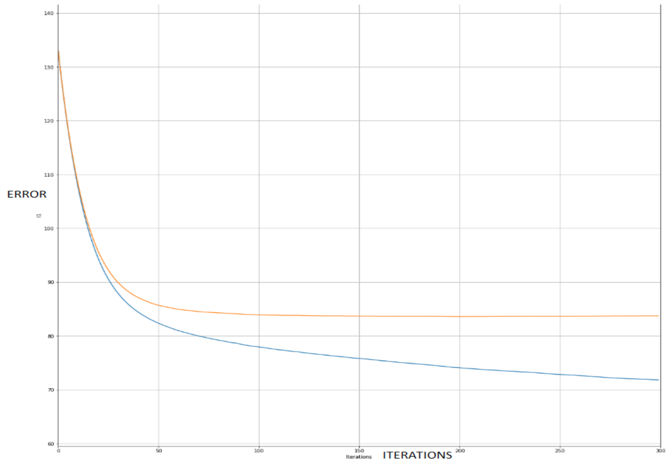
```
In [67]: import lightgbm as lgb

In [68]: train_data=lgb.Dataset(x_train_transit1,label=y_train_transit.delta_selection_departure)
         param = {'num_leaves':350, 'objective':'regression','max_depth':20,'learning_rate':.05,'max_bin':200,
                 'min_data_in_leaf':50}

In [69]: num_round=100
         lgbm1=lgb.train(param,train_data,num_round)

In [70]: y1=pd.DataFrame(lgbm1.predict(x_test_transit1))
```

Figure 5: Light GBM model



Results

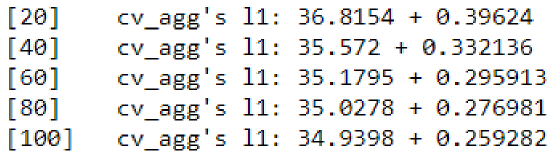
R2 score of delta_selection_departure for train data is 0.292
R2 score of delta_selection_departure for test data is 0.1806

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

Cross Validation technique used: k-Fold Cross-Validation

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
    - Take the group as a hold out or test data set
    - Take the remaining groups as a training data set
    - Fit a model on the training set and evaluate it on the test set
    - Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores



```
[20]    cv_agg's l1: 36.8154 + 0.39624
[40]    cv_agg's l1: 35.572 + 0.332136
[60]    cv_agg's l1: 35.1795 + 0.295913
[80]    cv_agg's l1: 35.0278 + 0.276981
[100]   cv_agg's l1: 34.9398 + 0.259282
```

Inference- Average evaluation scores of all folds is same.

RESULT

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

4 CONCLUSIONS

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 i.e. the accuracy of 0.2703 and 3rd rank .

5 References

- **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**
- **https://eng.uber.com/engineering-an-efficient-route/**
- **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/**