Optimizing Pit stop duration in Formula One Race

Karthikeyan Devarajan (Karde799)
Mowniesh Asokan (mowas455)

September 2021

Abstract

F1 races are one of the expensive sport and a big marketing area for many motor brands. In the current development of data analysis, many sports are involved in data analysis in order to improve their strategy in the game and F1 is not a outlier of it. So, In this project, we are optimizing pit stop duration. We have used different model and came to an conclusion to Random Forest Classification to predict pit stop duration and it's significant features. The duration spent in pit stop is one of the most crucial parameter for winning a race. The parameters includes number of laps, driver, constructor, number of wins, at what time the pit stop is used, and number of times pit stop is used. We have to pit stop duration by using the significant features selected from the model.

1 Introduction

1.1 Formula One

Formula One or F1 is an international single seated racing conducted across the world. It consists of different circuits and many closed road races. Formula refers to the rules and regulation for the constructors and drivers should abide by. F1 is currently increasing popularity and has always been a technology driven sport with increase in research and development for increasing efficiency and durability of cars. Software such as ANSYS, Solid Work is used for theoretical analysis of car oriented feature. Delta Topco is the private company which organizes and controls F1. There are formulated specification for all the components of the car and how to design the car given by the F1 committee. F1 has a history from 1920's in Europe but it was stop during the Second World War. After the Second World War, it restarted after 1951 and the number of races has increased from 1951 to 2020.

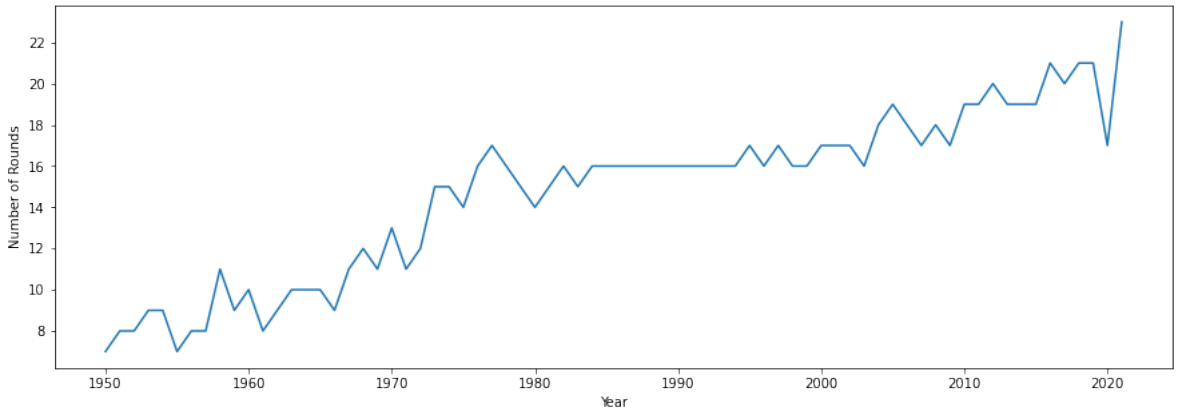

Figure 1: Number of races over the years

This can be viewed in the graph plotted from the data. Formula 1 is very expensive game which includes design of good high performance car, maintenance of car, pit stop team, and recruiting driver. Many construction teams were unable to manage their funds for racing which lead to reduction in number of construction teams. F1 is much technology driven which led to new race called Formula E which focus on the single seated racing for electric cars. ABB is currently organizing the Formula E after the approval of Federation International of motor sport. In Formula E, reinforcement learning is used for stargazing for faster completion of race and better adjustment towards ambient temperature. This is a good example how machine learning is used in sports. Mostly research is done in designing the cars but research is not done what is happening in the race tracks. Machine learning and data analysis will help in stargazing the speed of cars, breaks taken for pit stop. Many people are researching in designing the engine and use simulation on how to drive the car with less damage to car engine. F1 is huge commercial marketing for brands which could follow normal marketing.

The world championship is given for driver and constructors using a point scheme which is given in table 1. There are many races in addition to world championship such as South African Formula one Championship, European non-championship, and British Formula One championship. F1 world championship has a series of grand prix races and ranked based on the point system. After the series of races, the team of drivers and constructors with highest points are decided as world champions. Race tracks are known as circuits. The highest number of circuits is present in United States of America. The number of circuits present in each country and most used circuits are shown in table 1 and table 2.

Circuit	Number of Races
Autodromo Nazionale di Monza	70
Circuit de Monaco	66
Silverstone Circuit	54
Circuit de Spa-Francorchamps	53
Nürburgring	41

Table 1: Number of Races in each circuit.

Country	Number of Circuits
USA	11
France	7
Spain	6
United Kingdom	4
Portugal	4
Italy	4

Table 2: Number of circuits in each country.

Typically a race should have 306 kilometers and each circuit's length will decide the number of laps. The maximum duration of a race should 2 hours and when the duration exceeds 2 hours before the total length is completed, race will be completed when the lead racer completes. The driver can enter to the pit stop whenever they require with a controlled speed according to the law. The tyres are given by the race association for all the teams. In the early times, drivers can refuel in the pit stop but a law was made that they cannot refuel. The weight of car was less before this law came into existence, so the cars were faster.

There will be about 20 teams in a race and conducted across different circuits in the world. Some cars may be disqualified because of engine failure and crash in between the races. A team consists of 3 drivers

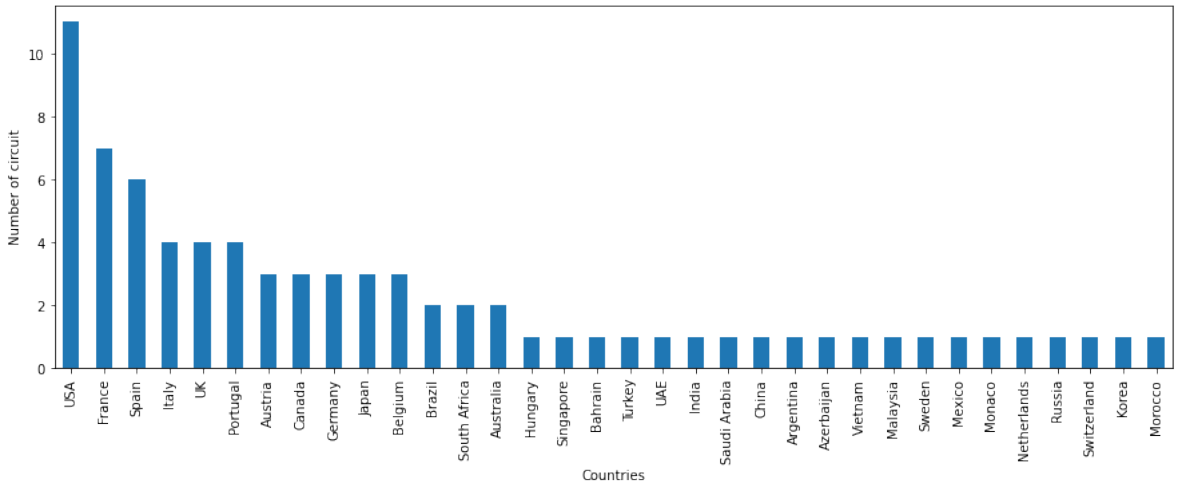

Figure 2: Number of races over the years

and 1 construction team. The construction team usually design and produces the car according to the F1 regulation. Most famous constructor teams are McLaren, Red bull, Ferrari, Mercedes and Honda. Each construction team can have two teams, which was regularized in 1960's. The construction team and driver's team will be allocated points based on the winning position. Some times, 2 construction teams can collaborate for manufacturing chassis and engine. The team which produces chassis is given importance than the engine constructor team.

The point's scheme is displayed in table 3.

Position	Points
1st	25
2nd	18
3rd	15
4th	12
5th	10
6th	8
7th	6
8th	4
9th	2
10th	1

Table 3: Points Table.

The points will given to both constructor and driver. The points table for driver and constructor is displayed below.

1.2 Data Analysis in F1

Many companies have used Machine learning for reducing cost for testing of car and improving marketing strategies. Most companies use data analysis in Computational Fluid Dynamics for improving aerodynamics and reduce down force. Down force is measure of vertical aerodynamic load on the surface of car. More down force will cause more stress on tire, which will in turn increase the number of time visiting pit stop. Monte Carlo is used for simulation for race between two drivers and strategy can be

Constructor	Points
williams	66.0
Mercedes	50.0
Red Bull	43.0
Ferrari	43.0
McLaren	43.0

Table 4: Points Table for Constructor

Driver	Points
Hamiliton	50.0
Massa	36.0
Bottas	30.0
Max Verstappen	26.0
Vettel	25.0

Table 5: Points Table for Driver

made for the next race. These types of simulations are done by sampling by changing tire type, pit stop lap, and positions. Due to high unpredictability, sometimes human decision is highly required. Most common machine learning technique is used Reinforcement learning. The reinforcement learning along with neural networks and function approximation, we could reduce the computation complexity.

[Hei+20] developed a Virtual Strategy Engineer (VSE) which uses 2 Neural Network i.e) One for deciding whether to take a pit stop break or not. If it takes a decision to use pit stop break, the other neural network choose the type of tyre to use.

[DM09] found that multi team cars had more victories and control the pole position. The multi team cars also increase revenue motivations also.

[Sto17] explained that artificial neural networks can predict the race winner and ANN outperform other model in predicting the winner.

We have plotted graph, that plots the driver who won with initial pole position as 1.

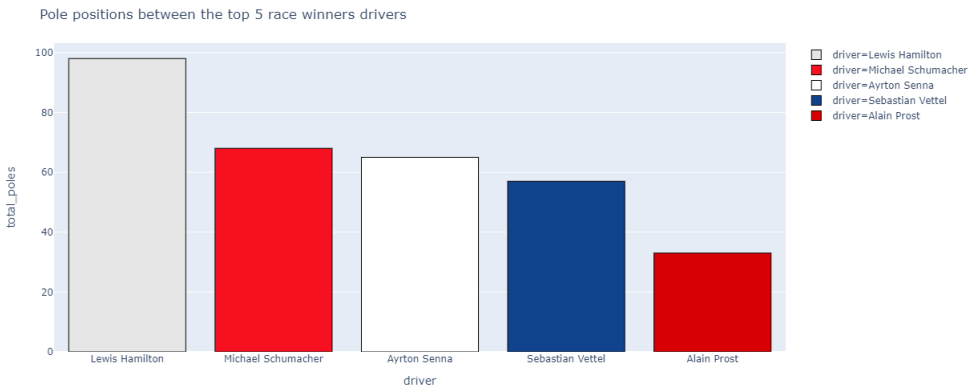

Figure 3: Influence of Initial poles on winning race

2 Objective

In this project we are going to discuss about the time taken in pit stop for changing the vehicle parts. The data set contains many parameters collected in F1 races from 1951 to 2018. This full paper revolves around the following statements:
‘ i) The main objective of this project is to model for pit stop duration and find the model accuracy
ii) Significant feature selection that influences pit stop duration

3 Data

We worked in the data set obtained from the Kaggle competition. There are about 13 different files which contains raw data. Raw Data contains details of each constructor and driver personal details, circuit details, each lap times, qualifying details, race results, pit stop details and status of the cars. The data sets contained each race details and the points scored by the constructor an driver, number of laps, duration at the pit stop, fastest lap time, at which lap driver entered pit stop, number of time he visited pit stop, driver Id, race Id, constructor Id , and many such parameters. Merging all the data set was difficult because they had a lot of NaN values. The merging of data set was done on basis of Race Id, Driver Id, and Constructor Id. The final data set contains the following feature:
i) Grid Position
ii) Fastest Lap Time
iii)Lap at which pit stop is used
iv) Duration at pit stop
v) Total laps
vi) Points of constructor
vii) Points of driver
viii) Time at which the pit stop is used
ix) Results of each race

Some feature is label encoded and the duration was in milliseconds, so the value was very high. This will lead to very low accuracy. So duration at pit stop is scaled. The final size of the data set is 19 parameter and 3959 observation.

3.1 Data Understanding

We plotted some graph and displayed some table to better understand the data. The fastest lap time by driver and minimum pit stop duration in different circuits is displayed in table 6 and table 7 respectively.

Circuit	Driver	Fastest Lap time
Bahrain International Circuit	russell	0:55.404
Red Bull Ring	sainz	1:05.619
Indianapolis Motor Speedway	barrichello	1:10.399
Autódromo José Carlos Pace	bottas	1:10.540
Circuit Gilles Villeneuve	bottas	1:13.078

Table 6: Fastest Lap time

Circuit Pit stop duration	Driver	Constructor
Yas Marina Circuit 128.97	maldonado	Williams
Hungaroring 131.73	hamilton	McLaren
Circuit de Barcelona-Catalunya 132.59	massa	Ferrari
Shanghai International Circuit 139	perez	Sauber
Circuit de Spa-Francorchamps 139.14	glock	Virgin

Table 7: Minimum Duration of Pit Stop

4 Model Selection

We tried various models for target response – Duration in Pit stop. We started with Regression models but classification models gave better accuracy. We decided to use Support Vector Mechanism and Random Forest Classification. SVM supports only for two class response variable but whereas Random Forest Classification can be used for multi class response variable.

4.1 Theory

Random forest is a type of supervised learning model and a combination of decision tree so that the model is more stable and accurate. Random Forest can be used as a classifier and regressor. Random Forest classifier uses gini index and entropy as a impurity criterion whereas Random Forest regressor uses mean absolute error and mean square error. Since we are using Random forest classification model, we will using either gini index or entropy for impurity criterion.
Gini Index:

$$\sum_{i=1}^{C} f_i(1 - f_i)$$

Entropy:

$$\sum_{i=1}^{C} -f_i log(f_i)$$

f_i is the frequency of the ith label at the node and C number of unique labels.

The general definition of entropy is measure of randomness. So while splitting the data, model has to decrease the entropy. This is known as Information Gain.

$$Gain(T, X) = Entropy(T)^{-}Entropy(T, X)$$

Where T – Target variable

X – Feature to be split on
Entropy(T,X) – Entropy of the split data on X feature

4.2 Hyper parameter

Hyper parameters of random forest classification are
i) n estimators = number of trees the algorithm builds before taking average of predictions. Increase in number of trees improves better performance but slows computations.
ii) Max depth = maximum number of features random forest considers to split node.
iii)n jobs = number of processors allowed to computation
iv) min samples split = minimum number of observation that are split
v) min sample leaf = minimum number of leaf required to split an internal node
They are more focused on the model than the data set.

4.3 Feature selection

The decrease in node impurity weighted on the probability of the reaching the node is used for feature selection.

$$NodeProbability = \frac{Numberofsamplethatreachnode}{Totalnumberofsamples}$$

Random forest is collection of decision tree, each node importance is calculated as

$$ni_j = w_j C_j - W_{left(j)} Cleft(j) - w_{right(j)} Cright(j)$$

ni_j – Importance of node j
w_j - weighted number of samples reaching node j
C_j - Impurity value of node j
left(j) – child node from left split on node j
right(j) – child node from right split on node j

Importance for each feature on a decision tree is

$$fi_i = \frac{\sum_{j:nodejsplitsonfeaturei} ni_j}{\sum_{k \in allnodes} ni_k}$$

where
fi_i = Importance of feature i
ni_j = importance of node j

After importance is measured, it is normalized between 0 and 1.

$$normfi_j = \frac{fi_i}{\sum_{j \in allfeatures} fi_j}$$

Final significant features are given below

$$RTfi_j = \frac{\sum_{j \in alltrees} normfi_{ij}}{T}$$

$RTfi_j$ - Importance of feature i calculated from all trees in the model
$normfi_j$ - Normalized feature importance for I in tree j
T - Total number of tree

5 Analysis

5.1 Hyper parameters

Since we decided to model using random forest classification, we have to tune the hyper parameter. We have plotted the graph of hyper parameter and their influence on the accuracy score.
i) Graph related how number of estimators is influencing the accuracy score

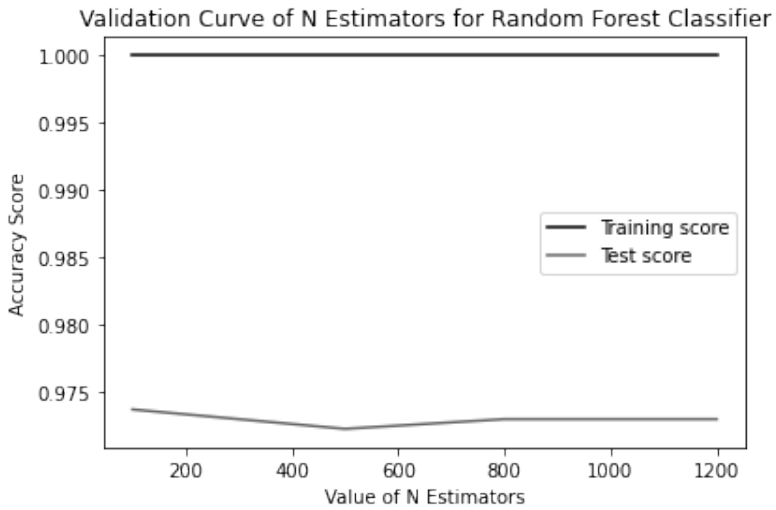

Figure 4: Influence of Number of estimator on Accuracy

The test accuracy decreases with increase in number of estimators.
ii) Graph related to how minimum number of sample split influences the accuracy score

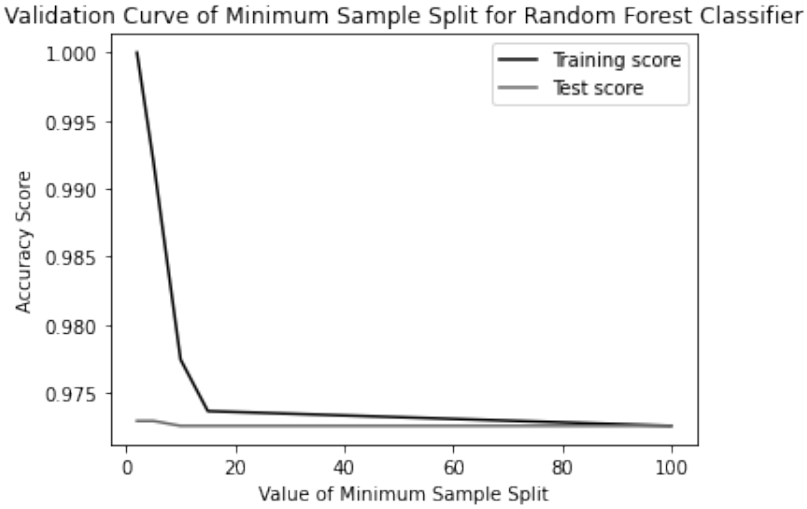

Figure 5: Minimum Number of sample split vs Accuracy

The accuracy value in testing data is more stable than the number of estimators.

iii) Graph related to how minimum number of sample leaf influences the accuracy score.

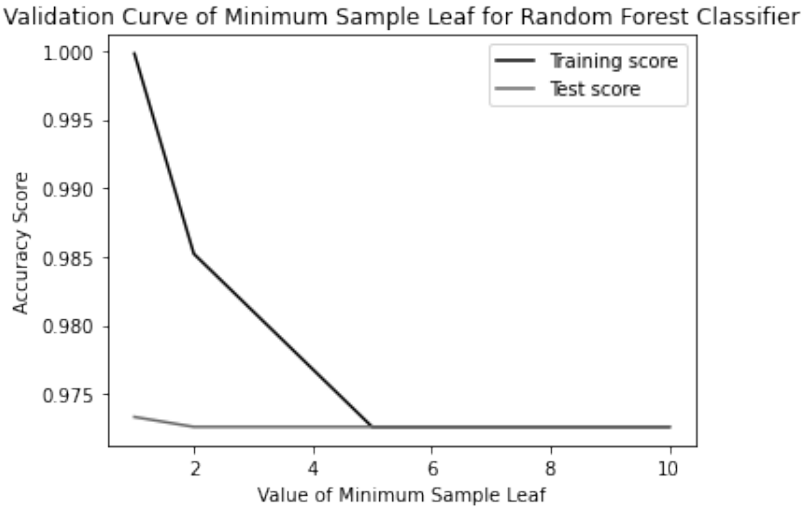

Figure 6: Minimum Number of sample leaf vs Accuracy

The accuracy of testing data revolves around 97 percent and accuracy is constant around 97 percent.

iv) Graph related to how maximum depth influences the accuracy score

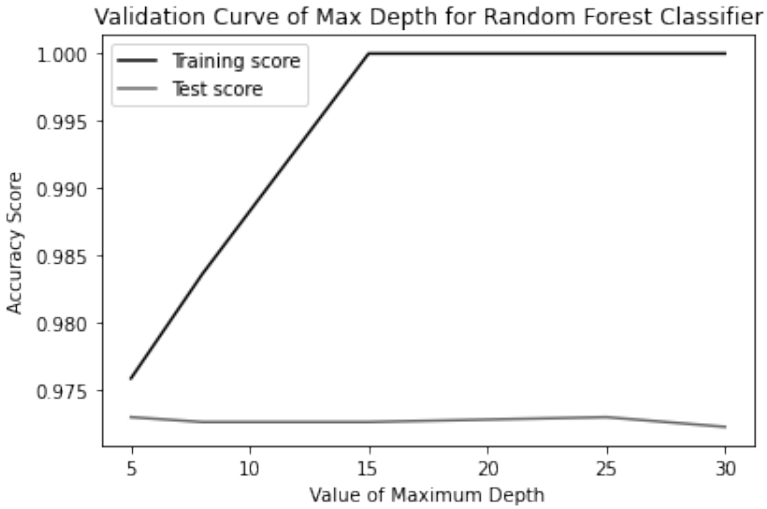

Figure 7: Maximum Depth vs Accuracy

Initial assumption of hyper parameters is
number of estimators = [100,300,500,800,1200]
minimum number of sample leaf = [1,2,5,10]
minimum number of sample split = [2,5,10,15,100]
maximum depth = [5,8,15,25,30]

The optimal hyper parameters are decided using Grid Search CV in python.

From the assumed values of hyper parameters, the optimal set of hyper parameters is
number of estimators = 100
minimum number of sample leaf = 1
minimum number of sample split = 2
maximum depth = 2

5.2 Model and Feature selection

The optimal hyper parameter chosen from the Grid Search CV are used as hyper parameter for the random forest model. We have implemented Random forest classification with hyper parameter and without hyper parameter tuning. The accuracy is increased when random forest classification is implemented with hyper parameter is tuned.

Model	Accuracy
Without Hyper parameter tuning	98%
With Hyper parameter tuning	99%

Table 8: Model with their accuracy

The random forest with hyper tuning is plotted below:

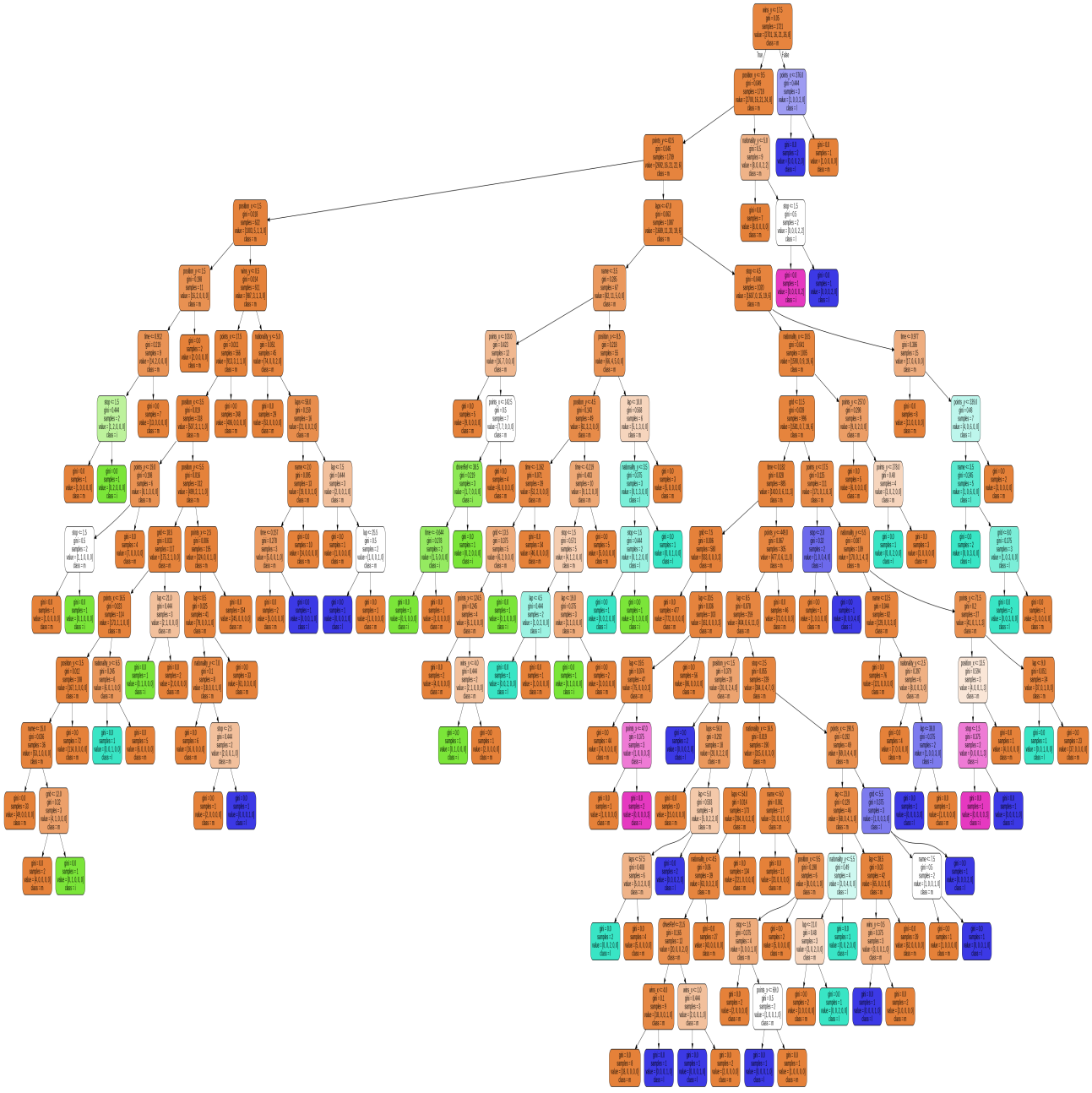

Figure 8: Plot for Random Forest

We have used the random forest model with hyper parameter tuning to select the significant feature for the duration at pit stop.

The significant features for the response variable – pit stop duration are

Feature	Importance
Time at which pit stop used	20.99%
Lap at which pit stop is used	14.70%
Stop	10.56%
constructor Points	9%
Total number of laps	7.93&

Table 9: Model with their accuracy

The actual data is more random than the predicted data. The randomness of actual data and predicted data is observed in the upcoming figure.

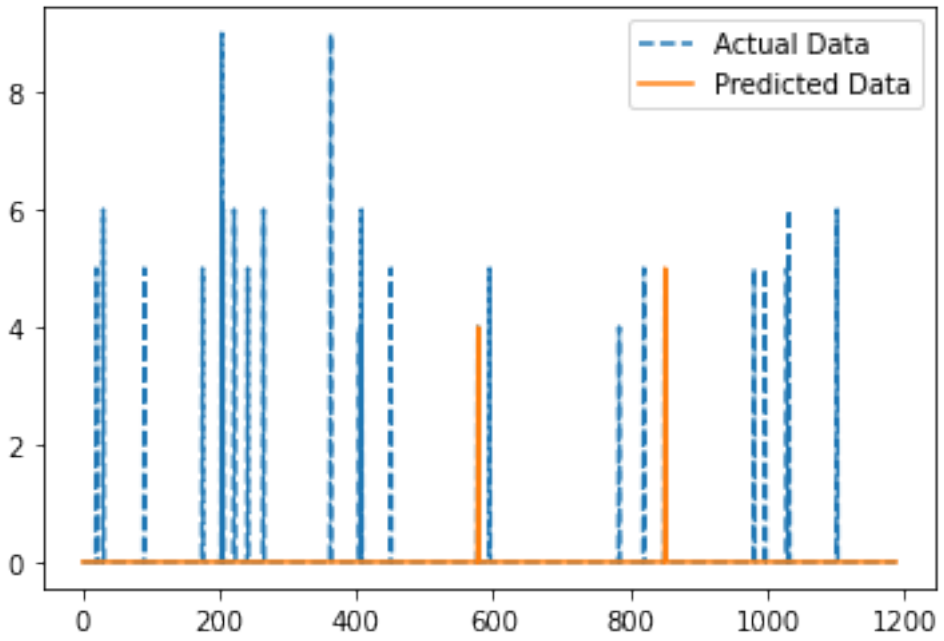

Figure 9: Predicted Value vs Actual Value

6 Conclusion

The outcome of the project is to model for pit stop duration and find the significant feature which affect the pit stop. We had 13 data set in which 2 were meta data set. We plotted the influence of hyper parameter on the accuracy for better understanding. The constructors which have less pit stop duration have more probability of coming first position. Pit stop duration are not controlled by the driver but by the constructors. The pre-processing of data includes merging data sets by removing are repeating parameters and NaN values. The duration at pit stop was in milliseconds, so we had to scale the response variable. The output of this model is that we could alter the lap at which pit stop should be taken, time at which the pit stop should be taken and based on total number of laps to get less pit stop duration. We could analyse for other parameters on ground for proper strategies of game.

References

[DM09] Craig A Depken and Larisa Mackey. "Driver success in the NASCAR Sprint Cup Series: The impact of multi-car teams". In: *Available at SSRN 1442015* (2009).

[Hei+20] Alexander Heilmeier, André Thomaser, Michael Graf, and Johannes Betz. "Virtual Strategy Engineer: Using Artificial Neural Networks for Making Race Strategy Decisions in Circuit Motorsport". In: *Applied Sciences* 10.21 (2020), page 7805.

[Sto17] Eloy Stoppels. "Predicting race results using artificial neural networks". Master's thesis. University of Twente, 2017.