**Formula 1 Race Prediction**

**Group 56**

Simran Bhatia

Radhika Jayachandran

(929) 592-8604

(857) 318-6720

[Bhatia.s@northeastern.edu](mailto:Bhatia.s@northeastern.edu)

[Jayachandran.r@northeastern.edu](mailto:Jayachandran.r@northeastern.edu)

Percentage of effort contributed by Simran: 50

Percentage of effort contributed by Radhika: 50

Signature of Student 1: *Simran Bhatia*

Signature of Student 2: *Radhika Jayachandran*

Submission Date: April 25, 2022

**Introduction**

At the forefront of emerging and innovative technology with passionate skillful drivers stands Formula 1 Racing. Formula 1 racing started in 1950 with various teams developing highly advanced technology. Formula 1 car racing is arguably the most popular motorsport championship. Behind every car racing on the track, there’s a team of data scientists hard at work, crunching data from millions of sensors, measuring lap times, tire and brake temperatures, airflow and engine performance to advise drivers and constructors on their next move. The volume of data being collected provides a perfect playground for predictive analytics.

A season comprises various races ranging 19 – 20 across countries around the world. Each race is divided into three day events. The first day has two free practice sessions provided to test the circuit and tracks and ensure everything is in place with the car and for the drivers. The second day has one additional free practice and a qualifying session. Qualifying session performance decides the position the driver starts on Race day and is judged on the basis of fastest laps set in record. Third day or Race day is the main event where the position of driver decides their position in the driver and constructor championship at the end of the season. A red and white race car

Description automatically generated with medium confidence

**Problem Setting**

Research and implement data mining models to predict the course of a season given the inherent unpredictability of the sport. Through the project, considerations and processes used to implement models to predict the course of the 2021 Formula 1 season were explored.

**Problem Definition**

The intention of this analysis is to identify factors that successfully determine race outcomes and to apply data mining techniques to design a model which predicts race statistics for upcoming championships.

**Data Sources**

The data was webscrapped from Ergast F1 data repository and the official Formula 1 website.

**Data Description**

The dataset contains information about all championships and races from 2012 to 2021, including their circuit location, drivers result, grid and finishing position of each driver, their teams, constructor standings, qualifying position on that particular race day.

**Variable Description**

Following are the main features used in the dataset along with their respective data types:

• Race ID - Unique key to associate Races (Numeric)

• Circuit ID - Unique key to associate Circuits (Numeric)

• Year - Year when the race was hosted (Numeric)

• Circuit - Circuit where the race will be hosted (Character)

• Country - Location of Grand Prix (Character)

• Driver Name - Name of the Driver along with their racing number, team they represent and Country they belong to (Character)

• Constructor Name - Name of the Constructor along with the country they represent (Character)

• Driver Standing - Position of Driver at the end of each race (Numeric)

• Constructor Standing - Position of Constructors at the end of each race (Numeric)

• Results - Final Position of Drivers and Constructors in the World Championship (Numeric)

**Predictor Data**

| Races | Results | Driver Standings | Constructor Standings |
| --- | --- | --- | --- |
| Season | Season | Season | Season |
| Round | Round | Round | Round |
| Circuit ID | Circuit ID | Driver | Constructor |
| Country | Driver | Driver Points | Constructor Points |
| Date | DOB | Driver Wins | Constructor Wins |
|  | Nationality | Driver Position | Constructor Position |
|  | Constructor |  |  |
|  | Points |  |  |
|  | Podium |  |  |

**Data Cleaning and Preprocessing**

Each year and each round were queried iteratively from the [Ergast API](https://ergast.com/mrd/) to get information about all the drivers and constructor result and standings. Points are awarded during each race based on where drivers and constructors finish the race. The first 10 drivers finishing are awarded points, with the highest being 25 points. The number of points and positions were extracted from Ergast API as well.

**Cleaning the Data**

Several variables were discarded from the dataset during pre-processing due to their irrelevance to the scope of analysis being conducted. These variables are mentioned below:

* URL - Many URL columns across all files listed Wikipedia links to various subjects, seasons, drivers, constructors, circuits etc. Since URLs will not aid in predicting race values, URL columns were dropped.
* Latitude, Longitude, Altitude - Latitude, longitude and altitude of circuit was listed for each circuit. These were removed due to their irrelevance to the analysis.

**Data Cleaning**

After removal of variables mentioned above, the data frame was combed for missing values, inconsistencies and outliers. None were found, hence no further action was taken.

**Reducing Data Dimension**

The variables were examined and it was determined that data dimension reduction would not be necessary for this dataset.

**Data Exploration**

The aim of visualizing acquired data is to enable exploration of data and detect potential trends to deep-dive into at a later stage. This is done to help discover and make sense of the data. The features were examined using myriad exploratory data analysis methods. Following questions were explored on the dataset:

**Which features are strongly correlated?**

Feature Correlation Matrix:

Chart

Description automatically generated

As can be seen from above correlation matrix, the features that are highly correlated are the constructor position and driver position, drive points and their podium finish and the least correlated features are the season with all other features.

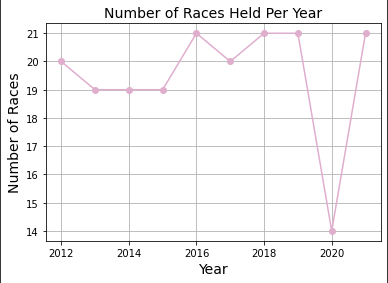
**Where around the world are circuits located?**

Map

Description automatically generated

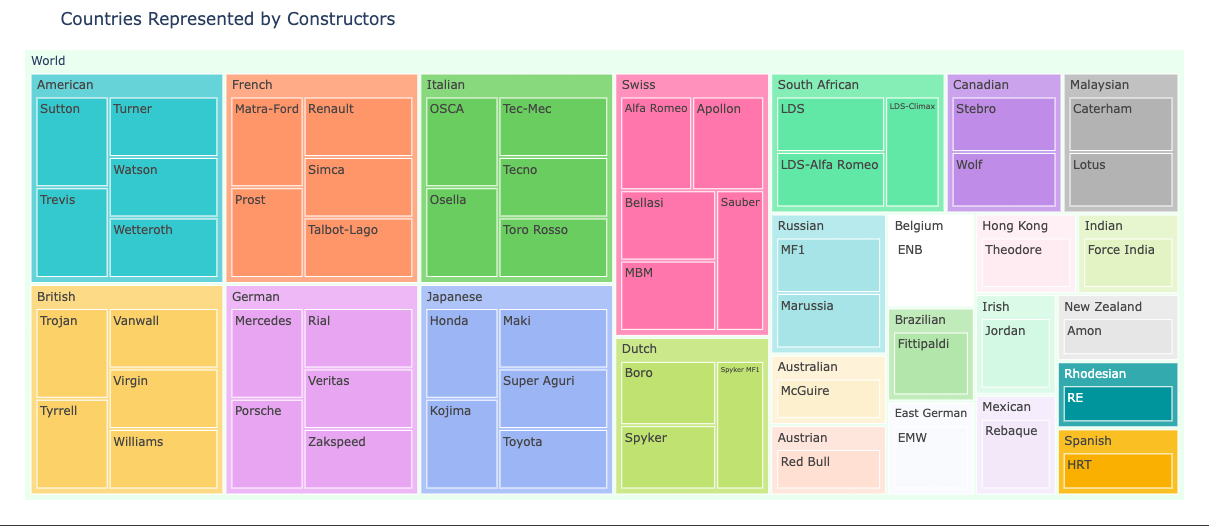
As we can see from the above visualization European countries host the most races in a season.

Line graph of number of races held each year:



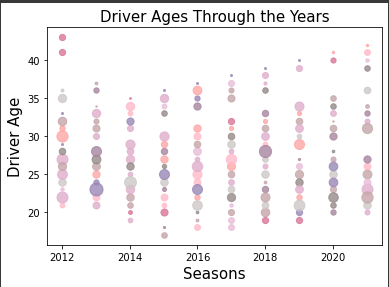
As can be seen from above visualization, number of races is more or less consistent with the exception of the large dip in 2020, presumably due to the COVID-19 pandemic.

Tree map of Countries Represented by Constructors:



Top five constructors representing each participating country can be observed from the above tree map. Some countries have fewer than five participating constructors.

Driver Ages:



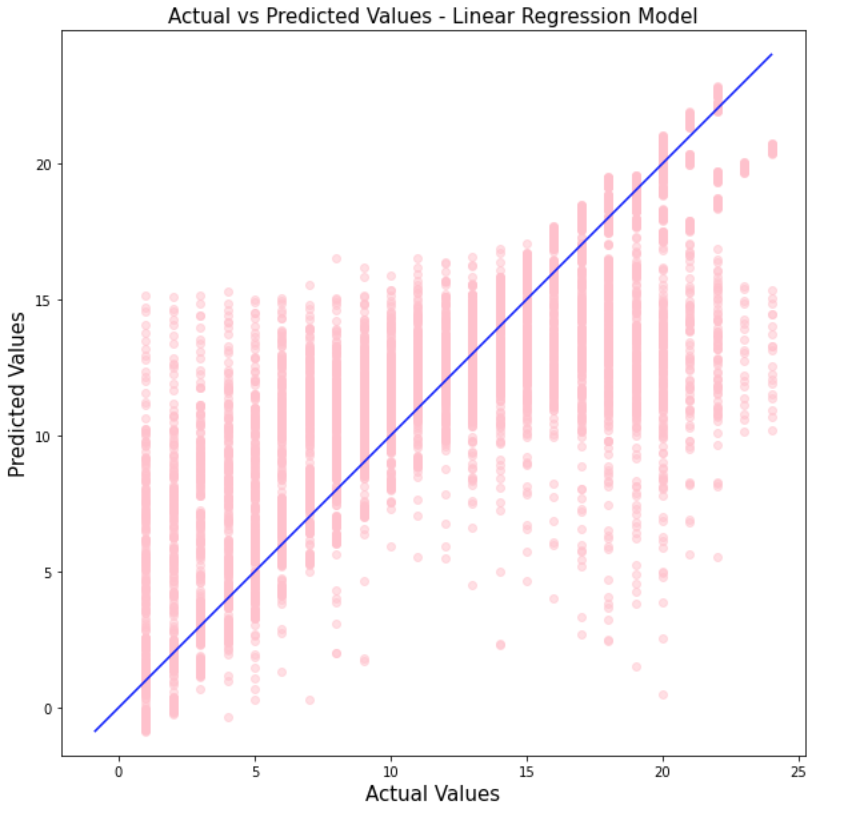
Driver ages for each year can be observed from the scatter plot above. As can be seen, ages have diversified with increasingly younger and older participants every year since 2012. 2021’s F1 season has seen the oldest F1 drivers since 2012, even though the general distribution has shifted towards younger side.

**Data Mining Models/Methods**

The models were implemented and assessed to find the best performer across both the Driver and Constructor Championships simultaneously using the 'Predicted Splits' Target. Models implemented were Linear regression, Random forest, Support vector machine and Neural Networks models. After conducting in-depth research followed by implementation, the Support vector machine model turned out to be the best among all tested, closely followed by the Random Forest model as higher accuracy and good scores were achieved for both.

**Linear Regression:**

Also known as simple linear regression, it uses a straight line to establish the relationship between two variables. Linear regression seeks to find the slope and intercept that define the line and minimize regression errors in order to draw a line that is closest to the data. One can use regression analysis to determine the strength of relationships between variables. Regression analysis, which employs statistical measures such as R-squared / adjusted R-squared, can speak towards how much of the total variability in a dataset is explained by the chosen model.



R² Value: 0.6421869963778043

Intercept: 11.045322647506524

Mean squared error: 13.524764641452316

Root mean squared error: 3.6776031109205243

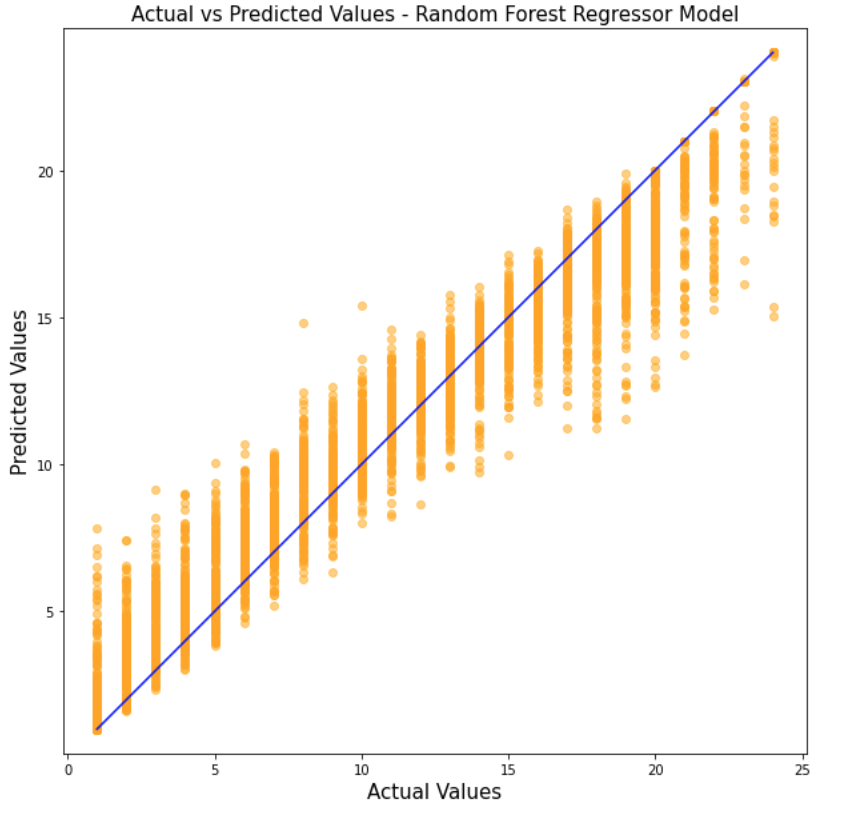
Coefficient: [ 43.49313231, -43.49913698, -1.29656154, -0.36871102, 0.10377775, 3.11130495, -2.90104373, 0.67940449, 0.56989836, 0.13322055]

**Random Forest Regressor:**

Random Forest is a popular classification algorithm that can perform both classification and regression. It is capable of accurately classifying large amounts of data. The name "Random Forest" comes from the fact that the algorithm is made up of decision trees. Each tree in the "forest" is based on the values of a random vector sampled independently with the same distribution as the others. Each one has been grown to the greatest extent possible.

Predictive analytics algorithms strive for the lowest possible error by either "boosting" (a technique that adjusts the weight of an observation based on the previous classification) or "bagging" (which creates subsets of data from training samples, chosen randomly with replacement). Bagging is used in Random Forest.

If there’s a large amount of sample data, instead of training with all of it, one can take a subset and train on that, followed by another subset and train on that (overlap is allowed). All of this can be done concurrently. An average is calculated by taking multiple samples from the data.



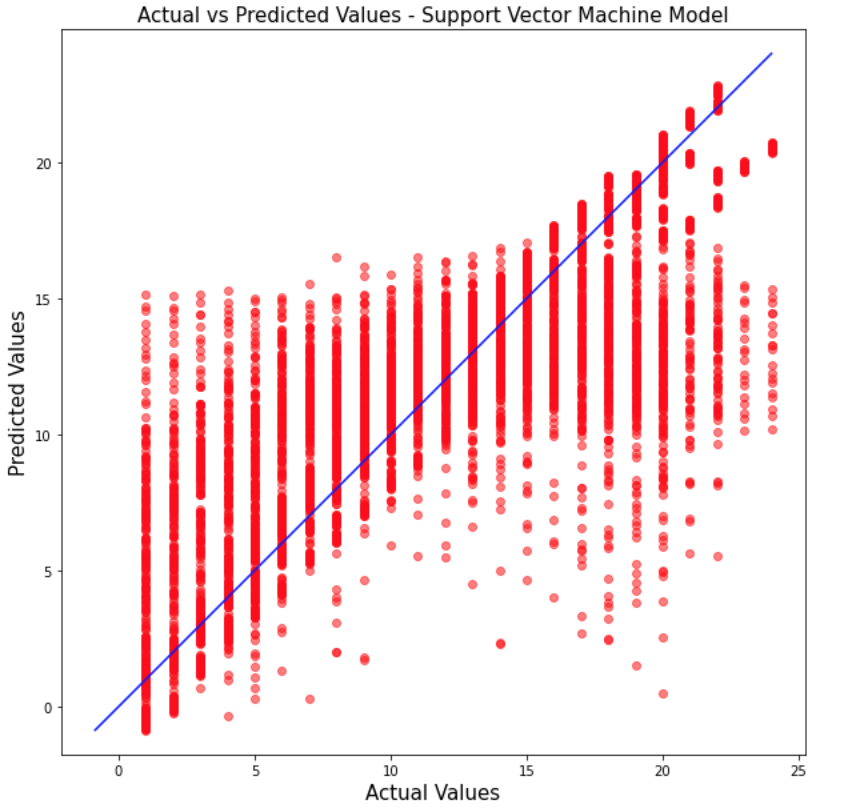
R² Value: 0.9538187177933495

Mean squared error: 1.745579300815255

Root mean squared error: 1.3212037317595098

**Support Vector Machine:**

Another simple algorithm that many people prefer is the support vector machine, which produces significant accuracy while using less computation power. SVM, or Support Vector Machine, can be used for both regression and classification tasks. The support vector machine algorithm's goal is to find a hyperplane in an N-dimensional space (N — the number of features) that clearly classifies the data points.



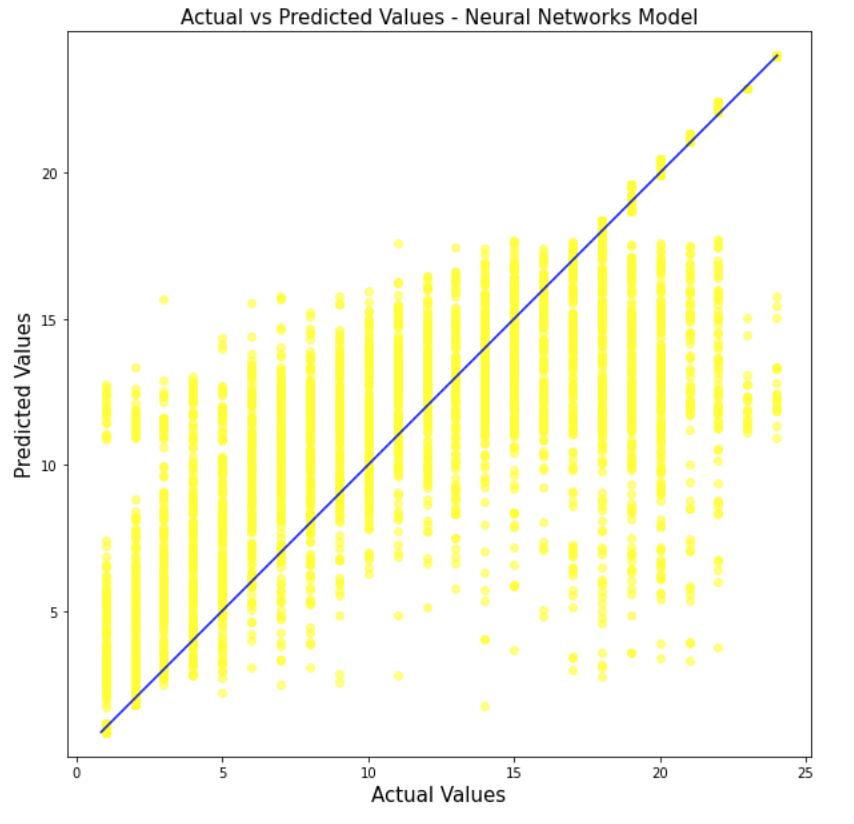
R² Value: 0.6421869963778043

Mean squared error: 13.524764641452316

Root mean squared error: 3.6776031109205243

**Neural Network:**

Neural Network is the model loosely inspired by the human nervous system. It creates networks between classifiers, and is used with models that are not linearly separable. This model creates well connected nodes that are used to cluster and classify the dataset. Although neural networks can be Used for classification and regression, it was in this instance used for regression with highly effective results. Unfortunately, it is complex to implement.



R² Value: 0.7029156169631899

Mean squared error: 11.22931899777016

Root mean squared error: 3.351017606305607

**Model Performance Evaluation**

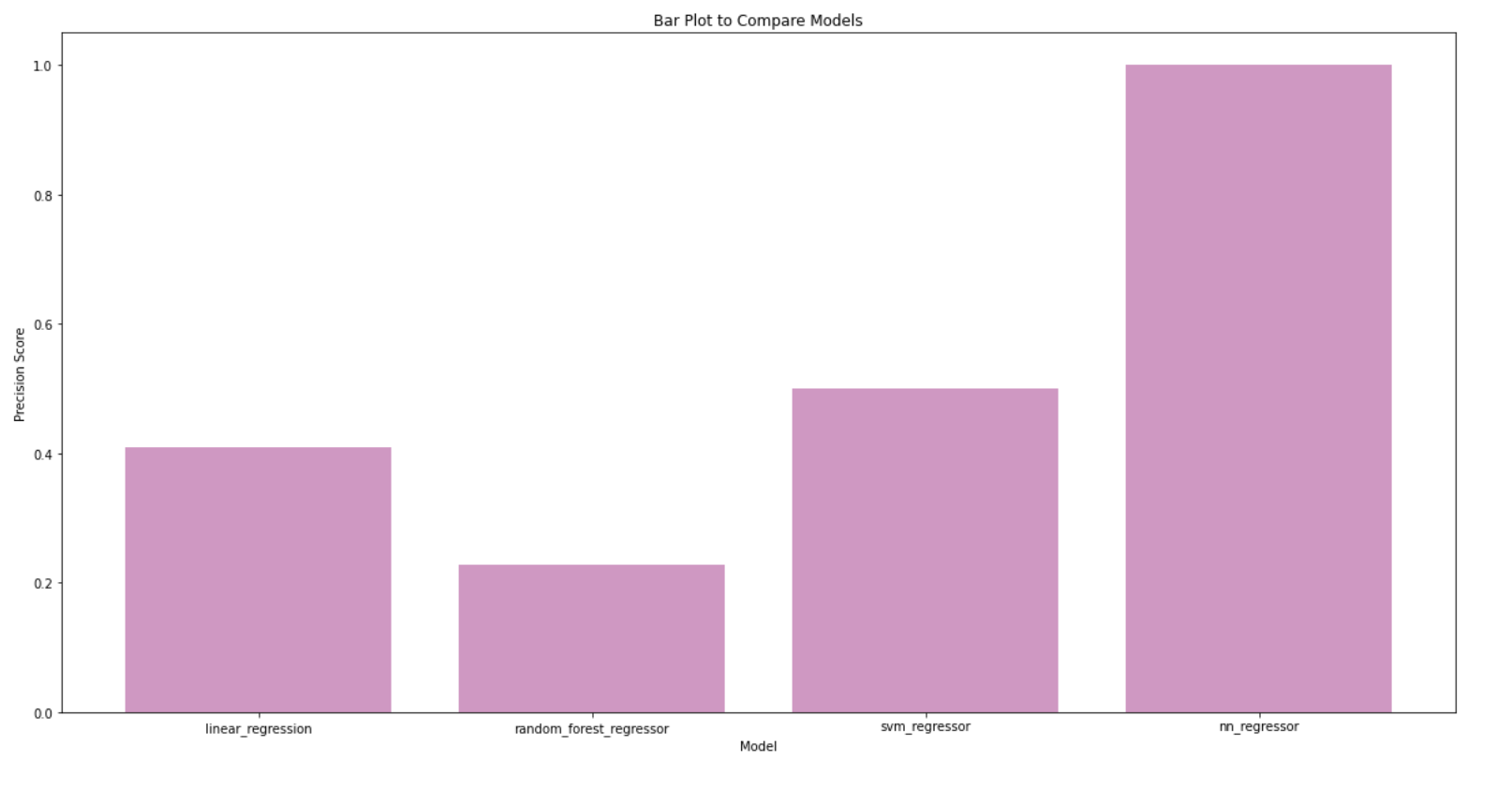
Model Performance Metrics:

| **Metrices** | **Linear Regression** | **Random Forest Regressor** | **Support Vector Machine** | **Neural Network** |
| --- | --- | --- | --- | --- |
| R² | 0.64 | 0.95 | 0.64 | 0.70 |
| MSE | 13.52 | 1.74 | 13.52 | 11.22 |
| RMSE | 3.67 | 1.32 | 3.67 | 3.35 |

Models Prediction Score:

| **Metrices** | **Linear Regression** | **Random Forest Regressor** | **Support Vector Machine** | **Neural Network** |
| --- | --- | --- | --- | --- |
| Score | 0.31 | 0.31 | 0.45 | 0.98 |

Model Comparison Plot:

****

**Outcome**

**Takeaways**

**References**

<http://ergast.com/mrd/>

<https://towardsdatascience.com/formula-1-race-predictor-5d4bfae887da>

<https://medium.com/@willgeorge93/formula-1-championship-predictor-a-machine-learning-solution-a86efcb9298>

<https://jasonjpaul.squarespace.com/formula-1-data-vis>

<https://medium.com/@willgeorge93/formula-1-championship-predictor-a-machine-learning-solution-a86efcb9298>

<https://www.formula1.com/>

<https://scikit-learn.org/>