Data-Driven Insights in Formula 1 using Machine Learning

Keane Coutinho,2021A7PS0080U

Agenda

- About Formula 1
- Objective
- Pre-Processing Dataset
- Models
- Experimentation
- Results
- Conclusion and Future Scope

Formula 1

- Formula 1 is the highest class of single-seater auto racing sanctioned by the Fédération Internationale de l'Automobile (FIA).

- 20 drivers and 10 teams compete each year for the World Championship

- The sport encompasses a rich set of data points including lap times, pit stops, weather conditions, and more.

Objective

Understand the various factors that influence the results of F1 race.

Predicting if a driver will finish on the podium, in points or no points

Comparing different models seeing which is most accurate.

Pre-Processing of Dataset

Initially all the different csv files were combined to form one Dataframe then the unwanted columns were removed.

Then using the Date of Birth of the driver and date of race the age of driver in days was found.

Using the number of DNFs caused by the driver and the number of races raced by the driver, the driver confidence was calculated.

Using the number of DNFs due to technical failure and the number of races raced by the constructor team the constructor reliability was calculated.

Models Used

Logistic Regression

Process of modeling the probability of a discrete outcome for the given input variables. Most commonly used to predict a binary outcome like true/false or yes/no

Decision Trees

Decision Trees, are a non-parametric supervised learning method used for classification but can also be used for regression. The model is created by learning simple decision rules inferred from features of the dataset.

Random Forest

A Machine Learning algorithm that combines the output of multiple decision trees to reach a single result. It can be used for both classification as well as regression.

Gaussian NB

Is a classification technique used in machine learning based on the probabilistic approach and Gaussian distribution.

EXPERIMENTATION

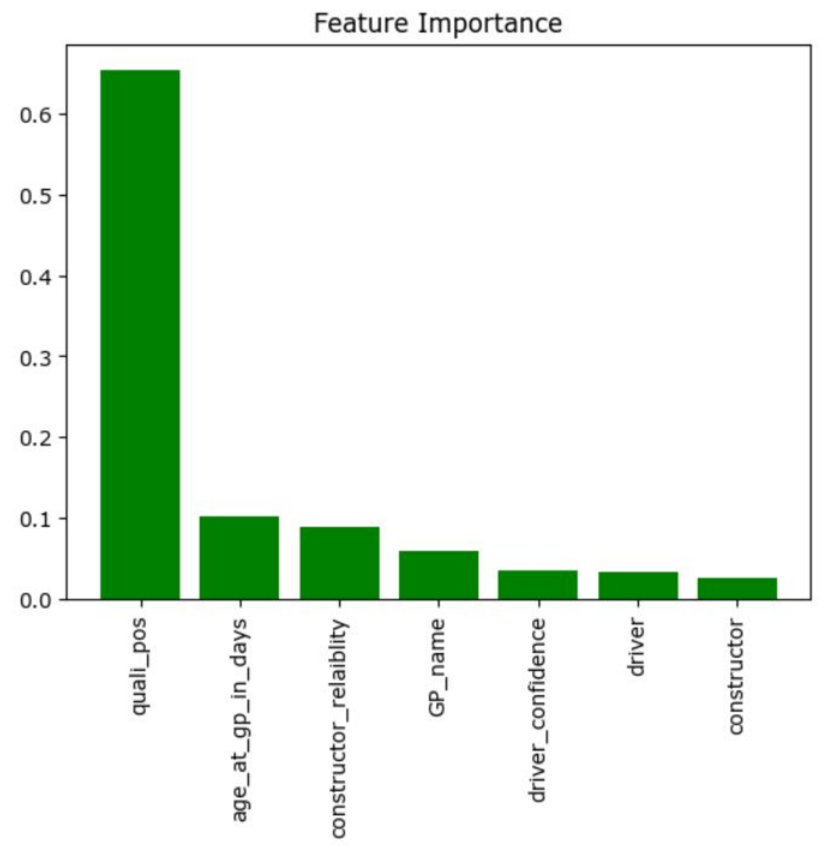
- All the text based data was label encoded before passing it into the model

- The measurement unit used was accuracy.

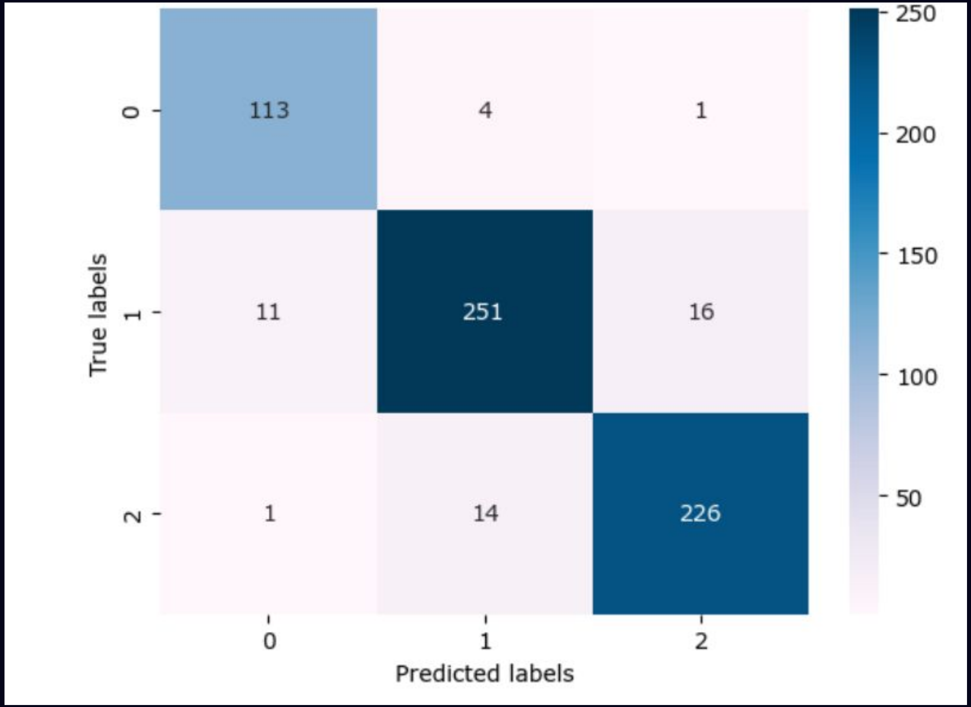
- All the 4 models were ran with 10- fold cross-validation.

Results

- Random Forest had the highest accuracy of 92.65%
- Logistic Regression achieved an accuracy of 83.75%.
- Decision Trees had accuracy of 88.93%
- Gaussian NB achieved an accuracy of 82.76%.



Feature Importance of Random Forest



Confusion Matrix for Random Forest

Conclusion and Future Scope

01

We successfully demonstrated the use of machine learning to predict Formula 1 race outcomes where Random Forest had the highest accuracy.

02

Features like weather which play a key role in determining the result of the race should be considered.

03

Also features like number of pit stops and pit stop times should be included in the dataset as this can be a decider between podium or not.

Thank you