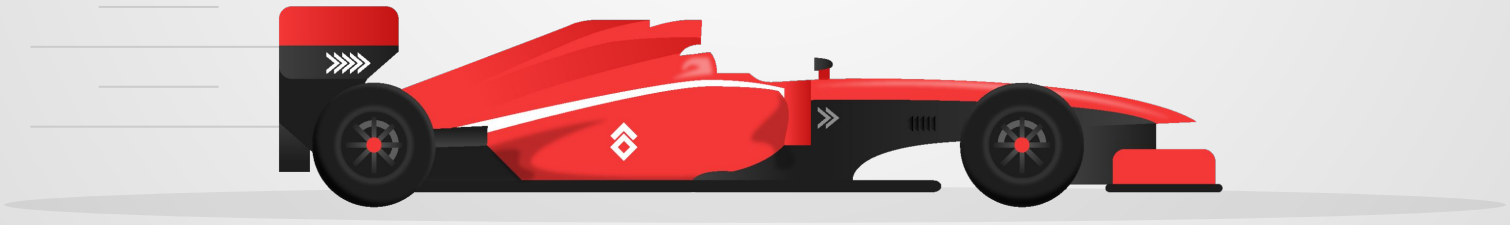
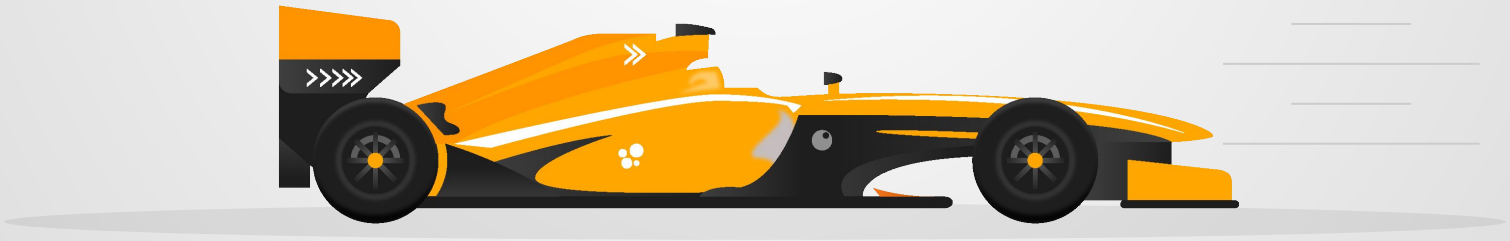
FORMULA 1

By Caitlin Nguyen



INTRODUCTION

- Formula 1 Driver Performance dataset
- Source: Kaggle
- Current data as of the 2023 Bahrain Grand Prix



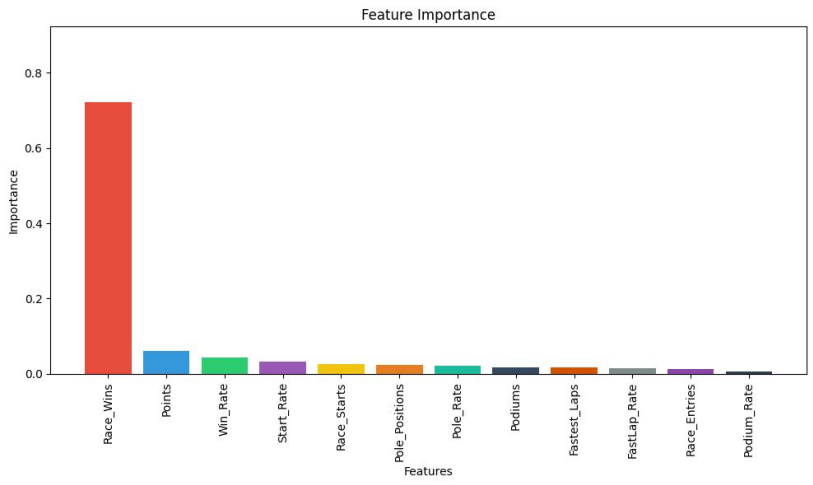
Dataset Overview

Key Features:

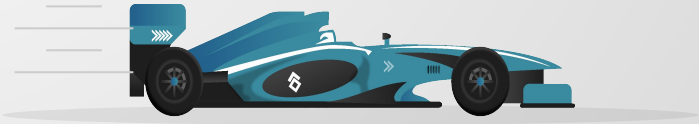
- Driver: Name and Nationalities
- Career Stats: Championships, race entries, starts, pole positions, wins, podiums, fastest laps, total points
- Performance Metrics: Pole rate, start rate, win rate, podium rate, fastest lap rate
- Career Span: Years active, champion status, points per entry
- Historical Data: Championship years, Driver's decade

01

When predicting the likelihood of being champions, which predictors are essential to improving the R2 when compared to a model that uses all the variables?

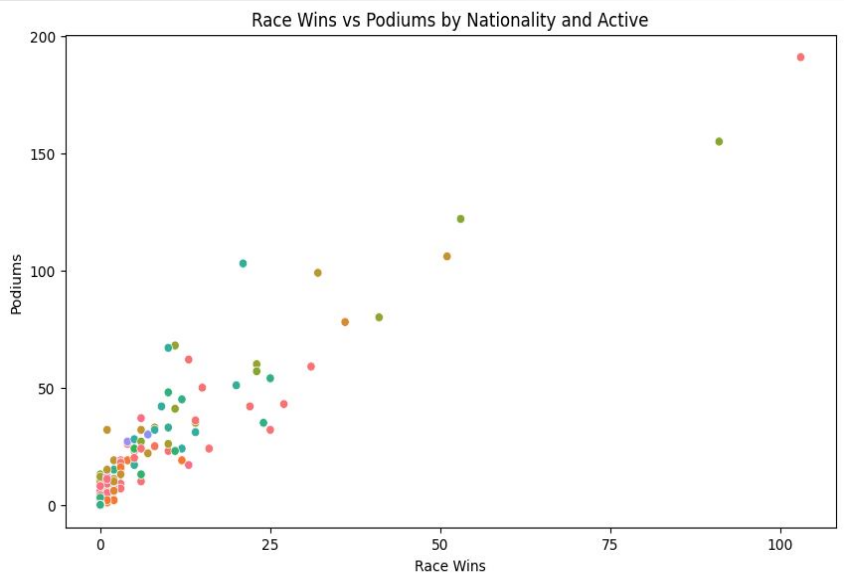


The predictors to the left are essential to improving the R2 from 0.29 to 0.85. They are the most important in predicting whether or not a driver would likely win championship.



02

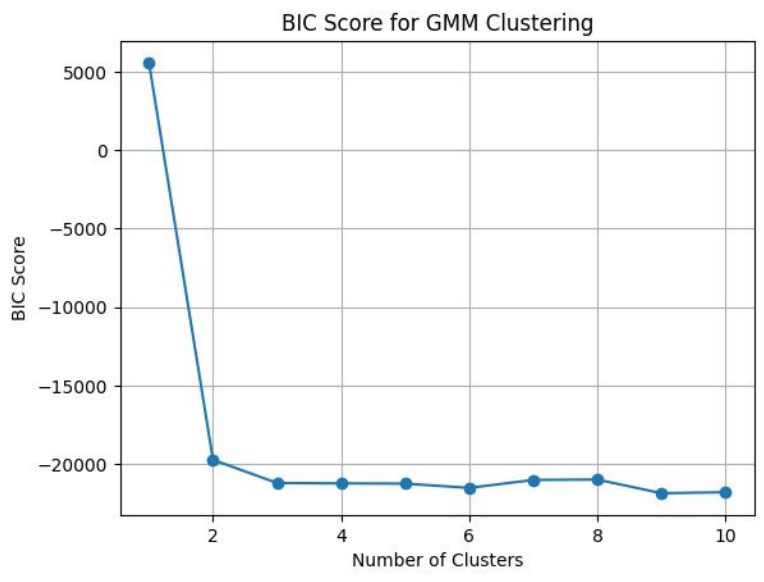
Considering race wins, podiums, and points, what distinctive clusters emerge among drivers, and what distinguishes these clusters in terms of their nationality and active status?



- Left Scatter Plot: outliers/noise points which are drivers who have won more than a few races
- Right Scatter Plot: dense clusters represent the majority of drivers who either had or had not won races or stood on the podium.
- No clear cluster in terms of nationality or active status

03

Can we cluster drivers based on their performance metrics (e.g., Win Rate, Podium Rate, FastLap Rate) to identify distinct profiles of drivers?



- This plot informed the decision on the number of clusters. An 'elbow' in the BIC score graph typically indicates the optimal number of clusters. In your case, the BIC score levels off after two clusters, suggesting that two is the optimal number of clusters for these data.

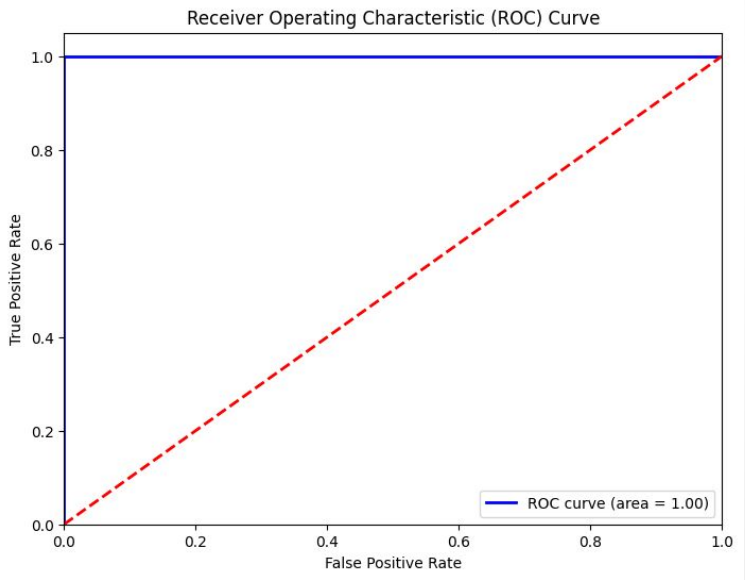
03

Can we cluster drivers based on their performance metrics (e.g., Win Rate, Podium Rate, FastLap Rate) to identify distinct profiles of drivers?

- The drivers are distributed among the clusters. The clear separation of drivers into different clusters indicates distinct profiles of performance.

04

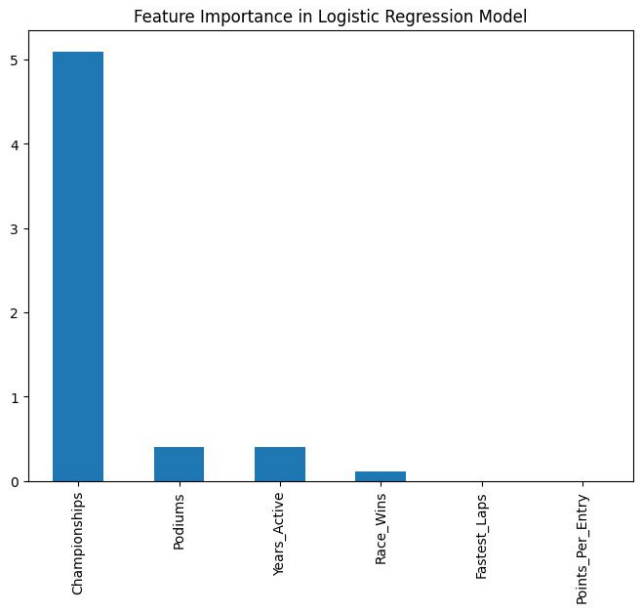
Which variables are the strongest predictors of a driver becoming a champion?



- The AUC of 1.00 means the logistic regression model has achieved perfect classification. The model has a 100% true positive rate (sensitivity) and a 0% false positive rate, indicating it can perfectly distinguish between champions and non-champions.

04

Which variables are the strongest predictors of a driver becoming a champion?



- The most significant predictor is the number of championships won, indicating a strong link between this feature and a driver's champion status. Other factors like podium finishes, years active, race wins, fastest laps, and points per entry also contribute to the prediction but to a lesser extent.