Formula 1 Analytics

Yash Vardhan Soni
Department of Computer Science and Engineering
PES University
Bangalore, India
dhan9576@gmail.com

Ayush Praveen
Department of Computer Science and Engineering
PES University
Bangalore, India
ayushpr.ap@gmail.com

Anirudh Chandrasekar
Department of Computer Science and Engineering
PES University
Bangalore, India
ani16may2002@gmail.com

Pracheth Thamankar
Department of Computer Science and Engineering
PES University
Bangalore, India
pracheth.thamankar@gmail.com

***Abstract*—Formula 1 is the greatest racing spectacle on the planet, with more than half a billion fans following its unrivalled mix of guts, grit and glory world wide. With the amount of data being captured, analyzed and used to design, build and drive the Formula 1 cars is astounding. It is a global sport being followed by millions of people worldwide and it is very fascinating to see drivers pushing their limit in these vehicles to become the fastest racers in the world! Since 2018, F1 has seen a steady increase in average viewership per race in the U.S. — from half a million in 2018 to almost 1.5 million in the 2022 circuit. The 47 percent increase from 2021 alone has helped spur F1's financial success. This paper focuses on using statistical analysis methods to gain insights on the various factors that affect lap time and driver performance. Exploratory Data Analysis is performed on a broad spectrum of features such as tyre compounds, time in pit stops, pole position, constructor standings, gear shifts on tracks, etc. We intend on running a correlation analysis between drivers and tracks to obtains each drivers best performance on different tracks.**

***Index Terms*—Formula One, Data Analytics, Pole Position, Best F1 Driver**

I. Introduction

The sports market has exploded in the 21st Century with the introduction of new media formats such as streaming broadcasting and social media. With the growth of sports and the monetary value associated with it sports analytics and sports team management has become a popular topic.

Game outcome prediction plays a major role in sports performance analysis. Game prediction significantly influences many parts of the sports market such as viwerbase,team management and strategy and also sports betting.

Formula One (also known as Formula 1 or F1) is the highest class of international racing for open-wheel single-seater formula racing cars sanctioned by the Fédération Internationale de l'Automobile (FIA). The World Drivers' Championship, which became the FIA Formula One World Championship in 1981, has been one of the premier forms of racing around the world since its inaugural season in 1950. The word formula in the name refers to the set of rules to which all participants' cars must conform.A Formula One season consists of a series of races, known as Grands Prix, which take place worldwide on both purpose-built circuits and closed public roads. [1]

The Formula 1 Championship is held annually with each season comprising of 20 races held in circuits around the world. Each race has ten teams with two drivers per team, competing against each other. Each team is responsible for building and tuning their cars, planning out strategies, training drivers, etc.
In March 2007, F1 Racing published its annual estimates of spending by Formula One teams.The total spending of all eleven teams in 2006 was estimated at $2.9 billion US. This was broken down as follows: Toyota $418.5 million, Ferrari $406.5 m, McLaren $402 m, Honda $380.5 m, BMW Sauber $355 m, Renault $324 m, Red Bull $252 m, Williams $195.5 m, Midland F1/Spyker-MF1 $120 m, Toro Rosso $75 m, and Super Aguri $57 million. [2] Given the monetary value involved in F1's logistics, team planning, technology and training, analysing performance and finding ways to make racing more efficient has become critical. This has paved way for Data Analytics in F1.
Data analytics has become an integral part of the F1 environment.Formula 1 is one of the most data-driven sports in history. The data analytics in F1 is based on telemetry.Telemetry is the collection of measurements and other data at remote or inaccessible points and their automatic transmission to receiving equipment for monitoring. For example the sector times for each track, tire temperatures, braking points, etc. Telemetry started to be used in it's modern form in the late 80's. Each car has from 150 to 300 sensors. The number isn't exact because from track to track they add and remove sensors. In Formula 1 racing, even the most subtle improvements can lead to victory. These improvements not only include fine tuning the F1 cars, but also training the drivers and the pit crew.

"Formula 1 is incredibly data-rich," says Dan Keyworth, director of business technology at McLaren Racing, one of the leading F1 teams of all time."Although we are a race team, we're really a technology business at heart."

II. Review of Literature

A. Will the pole sitter be able to use their advantage and dominate the race from the start to the finish line ?

This paper [3] seeks to analyse the effect of Pole Position [4] on the outcome of the race through Formula One's history A front position on the grid, or the so called Pole Position is considered to be the best. It gives the driver a head start to the first along with some significant benefits such as flexible boxing strategies, 'clean air' which affects aerodynamic efficiency of the car and the safety of avoiding mid-field collisions. The authors considered that implementing OLS is inappropriate since the outcome variable 'Final Position' is a limited dependent variable and varies between 1 and 24. Therefore it cannot be treated as a continuous variable. The paper uses two models the Logit model and the Probit model to determine a dummy variable (Will the Pole sitter win(or not win the race)) and the Poisson model to determine the value of the limited Dependent variable. The authors analysed these response variables against several potential explanatory variables such as the driver's skill, performance of the constructor, rain on race day, as well as track characteristics The paper takes into account that over the course of Formula 1's history the procedure of awarding pole sitter position to drivers has changed many times. In general, before the main event, drivers participate in qualifying sessions which involve consequent knockouts and a final timed trial to determine the driver that wins the pole sitter position. The fastest driver/constructor pair would usually get the pole position. But other factors such as fuel usage, time into which car boxed, etc also have an effect.

The paper finds that the effect of pole position is statistically significant at the per cent level (p¡0.01) and positive. Also a position further from the start reduces probability of winning the race (p¡0.01). Rain does not seem to effect the probability of winning. The paper highlights an interesting trend, the effect of Pole Position is increasing over time. In the 1950s the effect is insignificant, from the 1970s the factor becomes statistically significant (p¡0.05). The effect is largest during the 2000s. However due to change in regulations there are slight variations in the effect of holding Pole position at start of the race. The paper also finds that 5 drivers (Michael Schumacher, Aryton Senna, Alain Prost being 3 of them) have a consistently high win probability. This probability also has a life-cycle affected by the drivers general skills. The model used explains 29 percent of the variation in the winning variable . Also the pole position has a significant effect on finishing position (p¡0.01).

The paper concludes that Yes, the pole position does provide a significant advantage over the other drivers in the grid. This advantage is about two positions at the finish line or about a 10 percent point higher probability of winning the race. The paper captures the effect of features such as the driver's ability, constructor and track characteristics and the effect or the lack thereof of rainy conditions.

B. Who is the best Formula 1 driver?

This paper [5] is the first to try to evaluate the true talent of a Formula 1 driver by separating it from the performance of his car. The authors found that most rankings today represent a simple sum of metrics. However, 20 of achieved points and do not reflect a driver's true talent. The racing position of every Formula 1 driver is a function of several factors such as their individual talent, the quality of their cars as well as other race-specific variables, such as weather conditions, characteristics of the track and home advantage, among others.

The authors define the dependent variable

$$Y_{ij}$$

which represents the classification of a driver i in race t. It is found that points is not a good choice to choose as the dependent variable since points gives us no information on the difference in skills of drivers. Also, the number of points awarded has changed over the course of F1's history hence it is unreliable.Racing times was found to depend on racing strategies and hence is not a good dependent variable.Also in later stages of a race advanced drivers tend to slow down on purpose to hold their positions.Similarly the authors eliminated some other prospective dependent variables based on the fact they were influenced by the team and not just the driver. The authors developed the following functional form for the linear regression model

$$y_{it} = \alpha_i + \gamma_{s,i} + X_\beta + U_{it} \tag{1}$$

where α_i is a dummy variable capturing quality of driver i and γ_{si}, represents car-year- specific effects. X is the design matrix of the other control variables and β its corresponding coefficient vector. The design matrix contains 1291 dummy variables of all drivers and all cars plus additional control variables.

stands for the error term. By using linear regressions and controlling for driver and car dummies the authors separate the talent of Formula 1 stars from what their car contributes to success. Michael Schumacher has had the most absolute wins and is among the TOP-10 drivers. However, he is not the top ranked driver. The best Formula 1 driver ever is Juan Manuel Fangio.

Their analysis shows that Formula 1 data are not only of interest when trying to evaluate a driver's talent or in order to establish a world ranking. Additional economic and non-economic applications can be envisioned. By analyzing changes in the rules of Formula 1 driving we could quantify incentive effects. The analysis of dropouts also provides interesting insights on risk-taking.

III. Dataset

The Formula 1 World Championship dataset [6] contains 14 csv files. This data has been collected through 71 seasons of formula 1 from 1950 till the latest 2021 version. This data has been consolidated from ergast Developer API. It contains data about the circuits, constructor results, constructor standings, constructors, driver standings, drivers, lap times, pit stops, qualifying results, races, race outcomes, seasons, sprint results and status of the driver at the end of the race. These data values have been pre-processed by various cleaning and imputation methods.

IV. Data Analysis

We merged certain dataframes and formed a single dataframe for the analysis. The race outcomes, drivers, driver standings, constructors, and status were merged. There were three columns with missing values namely time taken to finish the race in milliseconds, fastest lap, maximum speed. We filled the missing values in the fastest lap with 0 as we can not assume on which lap they could have gone the fastest. For the other two columns we filled the missing values with the mean of their respective columns.

While checking for the skewness of the columns and found that wins, points and race time in milliseconds show high positive skewness while fastest lap and maximum speed have high negative skewness.

Next, to deal with outliers we specified the Inter-Quartile Range and considered only those data points which lie within the range. To find the correlation between the features, we plot a correlation heatmap[1].

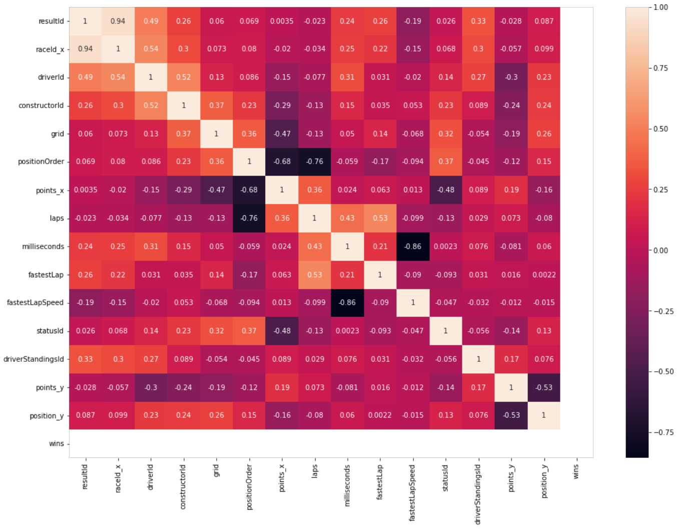

Fig. 1. Correlation Heatmap

As the correlation plot suggests, drivers who have greater maximum speed finish races faster which is indicated by milliseconds having a negative correlation with fastestLap-Speed. It is observed that drivers with a higher position order (numerically low) have completed more laps as our data contains lap wise information, indicated by a negative correlation. It is also observed that drivers with a better grid position(numerically lower) at the start of the race tend to score higher points which is indicated by a negative correlation between the columns. It does not have a strong correlation though as there might be other factors affecting race outcomes.

Acknowledgment

We would like to acknowledge our Data Analytics Course Professor Dr. Gowri Srinivasa for encouraging and providing us with this opportunity to get a hands-on experience in the field, and guiding us along the way. We would also like to acknowledge our assistant professors who have prepared the course content and also the teaching assistants who have been constantly providing resources to practice the learnt concepts.

References

[1] https://en.wikipedia.org/wiki/Formula_One

[2] Budgets and Expenses in Formula1. F1scarlet. Retrieved 30 August 2015.

[3] Wesselbaum, D. and Owen, P.D. (2021), The Value of Pole Position in Formula 1 History. Australian Economic Review, 54: 164-173.

[4] https://f1experiences.com/blog/f1-glossary-a-z-most-commonly-used-terminology

[5] Stadelmann, David. (2007). Who is the Best Formula 1 Driver - An Econometric Analysis (Wer Ist Der Beste Fahrer in Der Formel 1? Eine Ökonometrische Analyse). SSRN Electronic Journal. 10.2139/ssrn.1017292.

[6] https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020/metadata?select=lap_times.csv