Jacob Herbert and Ritwik Srivastava

INFO-I 399

1 May 2022

An Analysis of Formula 1 Race Data

Introduction

This paper dives into Formula-1 race data and tries to answer some questions about driver performance based on various attributes. Formula-1 has been the premier racing event in the world since its inception in 1950. These days, Formula-1 teams are highly dependent on large amounts of data, collected during the race, to improve the performance of their cars and drivers. This led to this research as it would be interesting to find out whether some basic race metrics like a driver’s starting position, average lap time, total race time, and any other metrics affect the driver’s finishing position. This would help us examine whether these basic metrics can help a team improve their cars, or whether they need more in-depth data about the car.

The dataset for this project was found on Kaggle and it contains 14 csv files. For our analysis we restrict ourselves to 4 files as these files have all the necessary data.

Data Questions

The two questions we want to answer are:

1. Does the grid position (the starting position of a driver) impact their final position?
2. Is lap time a good metric to assess a driver’s performance?

Methods

While we want to know whether grid position has a predictive effect on final position, we also want to create a model which has adequate predictive value. To do this, we first conduct a linear regression, giving those who did not finish (a not small portion) a value equal to the last finisher in the race. Then, we test a model identical except for the exclusion of all values where the driver did not finish the race.

We then add variables to the linear model which we think will enhance the predictive ability of the model. Such a process involves some degree of experimentation where variables are included or excluded dynamically to the linear model. To decide what variables to include, we conduct exploratory analysis on the data to see where variables correlate and how they are distributed. We also sought to include data from other tables by joining them.

After we do this to the linear model, we conduct the same process on the logistic regression model. At first, we define our binary variable as whether a driver won or not. Once we found a satisfactory model for that particular outcome, we expanded it to find whether the model can predict whether a driver placed in the top three (podium) for a particular race.

For all processes, we separate the data into an 80% training set and 20% testing set for both predictive and outcome variables.

Q1 and analysis

As previously mentioned, our primary goal of this project was to determine the effect that grid position has on finishing position. We began with the simplest possible regression, a linear model with grid position as the only input variable. We first conducted the analysis including those who did not finish the race and giving them the same position as the last finisher in the race. This yielded a mediocre model with an R-squared value of only .144, meaning it only explained about 14% of the variation in the model. When completely removing the observations where a driver did not finish, the model becomes far more predictive, with an R-squared value of .416. We think this jump in predictive ability is due to the fact that reasons for not finishing a race, such as crashes or breakdowns, are largely unrelated to performance and occur roughly equally to all drivers. This means that giving the drivers a negative result for a DNF unnecessarily muddies the waters.

While this was not a bad result for a single variable linear model, there is still much more explaining to do. To explore this, we added the rank of a driver’s fastest lap in a given race, which did improve the predictive ability, increasing the R-squared value to .524. However, because this information would not be available at the desired time of prediction–the beginning of the race–we decided to explore further and exclude this variable going forward. Instead, we opted to include variables from the driver standings data. We initially included the number of wins a driver has in a season, but it did not meaningfully improve the predictive value, likely because the data was far too sparse. Instead, we replaced it with a far more detailed data point: season points. Placing anywhere in the top ten finishers in a Formula 1 race nets a driver a certain number of points, with the most points going to the winner. At the end of the season, the driver and team with the most points wins the championship. When we include points, the R-squared jumps to .538, quite a notable jump given the general chaos of the sport and that these two variables are quite correlated with a correlation coefficient of -.551. Overall, we found this to be the best linear model, concluding that both points–a proxy for season-long performance–and grid position–an indicator of recent performance in qualifying–are significant in producing a model that predicts where a driver will place at the end of a race. This is especially notable because, for most of the season, the points data is also quite sparse and the metric becomes more informative over the course of a season. As a next step, it would be useful to analyze the hypothesis that, because of this increase in the informative nature of the model, it will become more predictive over the course of a season.

Now that we have analyzed the linear regression model’s ability to predict Formula 1 placement given data available at the beginning of the race. Now, we will use logistic regression to predict whether a driver will win a given race. To do this, we created a binary of the final position, with a 1 if the driver finished in first place, and a 0 otherwise. We then used the same predictor variables as the linear model–grid position and points. Using this data, we obtained a logistic model which predicted whether a driver won with an R-squared value of .965. Testing the accuracy of this model further with a created accuracy function, we discovered the model correctly predicted whether a driver won or not with 96.3% accuracy.

Given the high predictive value of the logistic regression model, we decided to expand the analysis to test whether it can predict if a driver places podium (meaning first, second, or third) in a race given the same predictor variables of grid position and points. To do this, we first had to create a binary for placing podium, with 1 being assigned to any driver who finished in the top 3 and 0 otherwise. The result of this analysis was a logistic regression model with an R-squared value of .771. Using the accuracy function, we find that it predicts whether a given driver placed podium or not with 82.5% accuracy–not a bad prediction. However, we did have some concerns about the true predictive value of the model being distorted by the fact that most drivers do not win a given race, thus meaning a 4% failure rate may not be much better than a simple educated guess.

Q2 and analysis

For the second question we again try to answer it by finding any correlation between the average lap time and a driver’s finishing position using the linear regression model. This gave us an even worse model as compared to using the starting grid position model, with an R-squared value of 0.047. This explains just 4.7% of the variation. We think that this model resulted in a low correlation because there is not a lot of variation in the average lap time of each driver. Therefore, we tried to use the total race time of each driver along with the average lap time in the linear model. But, this did not improve our model by a lot. We got an R-squared value of 0.048.

Since the linear model was not a good fit, we tried the logistic regression model. By using just the average lap time, we got an R-squared value of 0.11 which is double of what we got with the linear model. But this is also not a good enough value for prediction. When we introduce the total race time variable, we achieve the exact same R-squared value of 0.11.

Therefore, we can satisfactorily conclude that the linear and logistic models are not good models to test whether lap time is a predictor of driver performance. This is because the distribution of data is neither linear nor logistical. We might achieve better results using a decision tree model.

Results and conclusions

We cannot predict with much accuracy whether a driver’s grid position increases the chances of that driver winning the race.

The average lap time is not a good predictor of driver performance as there is very little correlation between the average lap time and finishing position.

Therefore, with our initial analysis, it is clear that basic metrics like grid position and lap time are not good indicators of driver performance and hence, teams must use much more and detailed metrics to improve the performance of their drivers and cars.

Paper and Project Contributions

Introduction - Ritwik

Question 1 and analysis + - Jacob

Question 2 and analysis + Methods - Ritwik

Results and conclusion - Ritwik

Both group members contributed to the code and analysis as well as the presentation slides.