Nathan Kraft

Dr. Tanner

Data 200

4-18-21

Final Project

Formula 1 is a racing league that travels across the world to compete and show who has built the fastest car over the season. There are 10 teams, each team having 2 drivers per team. At least, that is how it has been over the past few years. There are certain point amounts given out for each placement a driver earns within the race. The total amount of their points earned over the year decides where they end up regarding the Driver’s championship. The points combined from the two drivers from each team decide what placement the constructors earn for the Constructors championship.

The goal for this project is to try to predict which drivers would finish each race in a podium position, this meaning they finish within the top 3, at the end of each race. This is interesting to predict, since you can have drivers that may be in the best car, they get a grid place penalty, and then start the race at the back of the pack, and still manage to finish in the top 3.

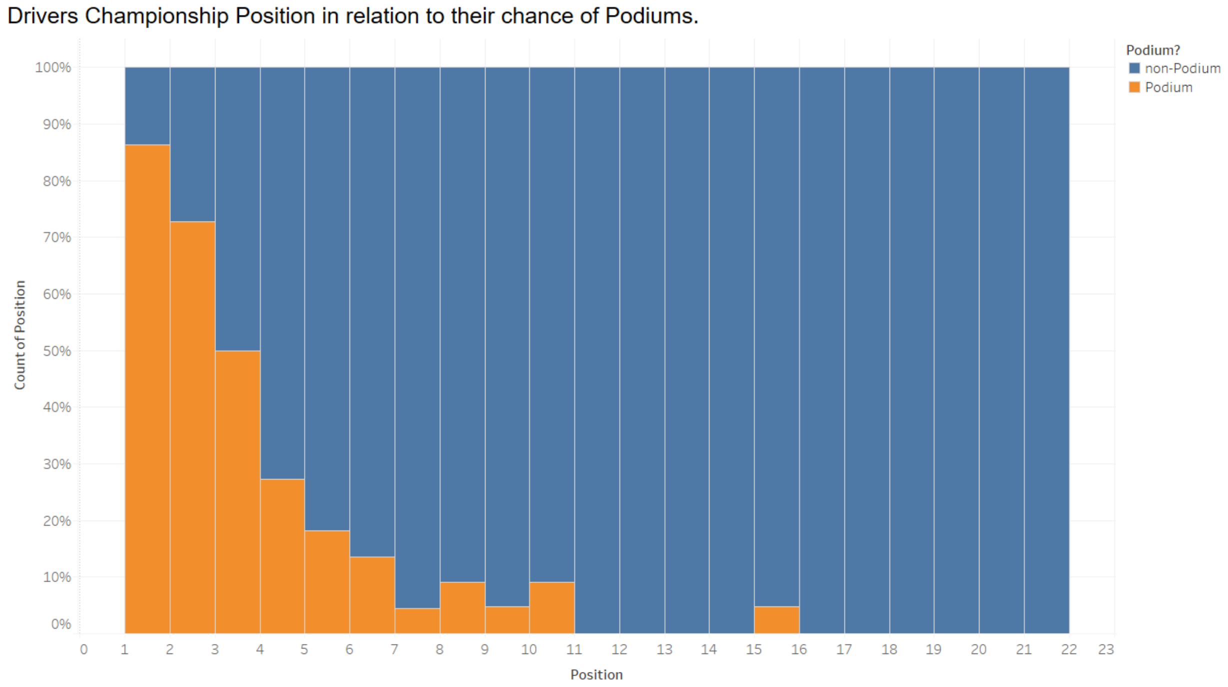
I am planning to use a driver’s current point standings in the Drivers’ Championship, as of that race, and their starting grid position as the main variables for predicting. These are also decided by other factors as well, so they are technically more involved, but not necessarily being investigated further by me. These variables are important because they tell how the driver has been performing throughout the season so far, what their current point standings are, and then how they performed in qualifying which decides their starting placement- not including penalties- for the main race, being their grid placement.

The data I am using for this project is originally collected by the FIA, *Federation Internationale de l'Automobile,* which is the main governing body of racing in the world and is a big part of governing the regulations set for the sport. It is collected as each race happens within a season, meaning as the race finishes, the FIA record what order each driver finishes, gives them their points for the finishing positions and keeps a record of any penalties they might have earned.

I believe that this is a reliable source. I have found nothing disproving any of the data found from my source. I found it from a database which I believe took it from the actual Formula 1 website. I found my data from ergast.com I will have a link for the main API page for the data in the Works Cited Page.

Where they are within the points standings at the time, and how they are doing throughout the season will show how often they tend to be in the top 3, being a podium. What position they start the race in has a decent shot of showing what position they finish in, it tends to be very close if not the same as their starting position, disregarding any random instances where a good car happens to start at the back of the pack because of a grid penalty.

The following graphic (*Figure 1)* shows how their Drivers Championship Position relates to their likelihood of achieving a podium. The people achieving podiums are orange on the graph. You can tell that the higher their position in the championship, the more likely they are to achieve a podium, which is what is believed, since the more podiums you have the more points you earn, so you will climb in the championship standings.

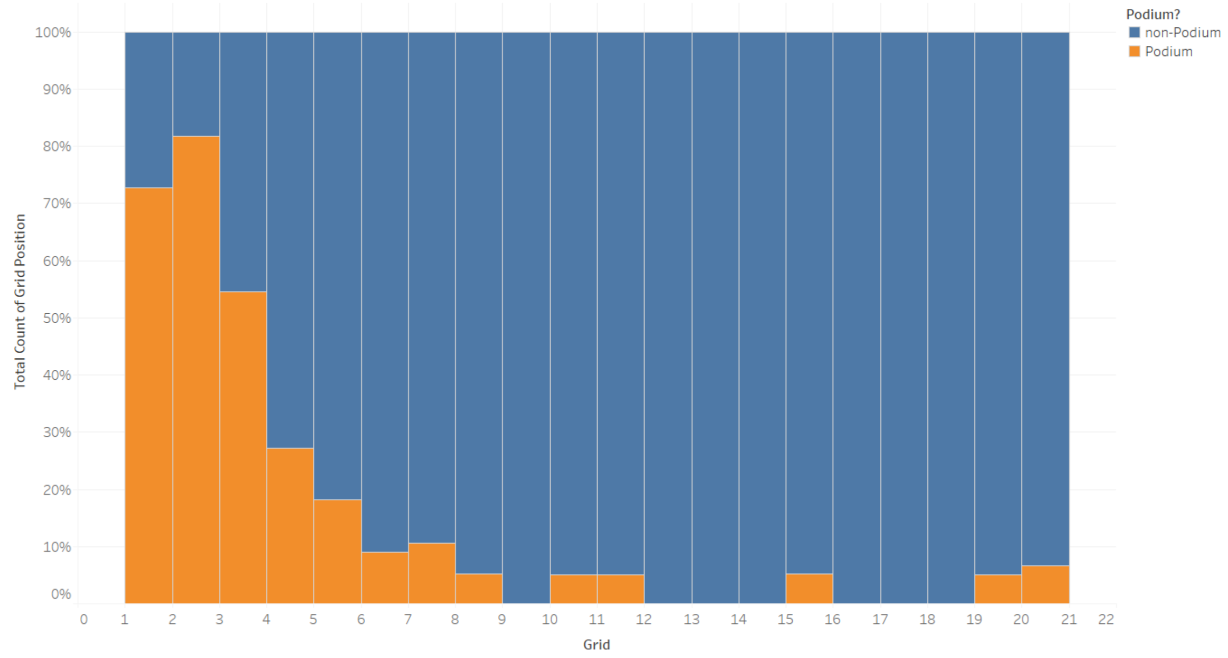


Drivers Championship Position in relation to their chance of Podiums.

The trend of % of Total Count of Position.X for Position.X (bin). Color shows details about Podium?. The view is filtered on % of Total Count of Position.X, which keeps non-Null values only.

Figure 1

The following graphic (*Figure 2)* shows the relationship between the starting Grid position and the number of podiums from that grid position. This shows that the higher in grid position the driver starts, the more likely they are to achieve a podium. This would be expected since the higher in the order they start, the more likely they will finish in the top 3. That and being higher in the grid order means they are more likely to perform well and put in faster laps than the lower grid placements.



Grid position and relation to their chance of podiums

The trend of % of Total Count of Grid for Grid (bin). Color shows details about Podium?. The data is filtered on Grid, which ranges from 1 to 20. The view is filtered on % of Total Count of Grid, which keeps non-Null values only.

Figure 2

I made a confusion matrix to try to investigate what the correlation is between my two variables and the chance of a driver getting onto the podium within a race being my classification variable. The matrix takes the prediction of the model, and then tests to see if the model is correct. I found weights and percentiles that worked well and were able to predict results well.

The best weights that I found for Grid position and Drivers Standings position respectively were (1,4). This gave the highest accuracy with close to the highest area under the curve. I then narrowed it down to 15th percentile being the best one to use. It has the closest predicted podiums to the number that happened. The model predicted a little less podiums than happened and it had a good number of non-podiums predicted that is close to the correct amount. This was the only combination of weights and percentile/threshold with a high accuracy and a low number of false positives happening. Figure 3 below shows the numbers of the model.

	Actual Podium	Actual non-Podium	Total Predicted
Predicted Podium	43	13	56
Predicted non-podium	23	315	338
	66	328	394

Figure 3

It had the highest accuracy of all of them at 0.9086; it also had a true positive rate (number of predicted podiums divided by the number of actual podiums) of 0.6515, and a false positive rate (number of falsely predicted podiums divided by the number of non-podiums) of 0.0396. I found that you could get a higher true positive rate if you went to a higher percentile, but you would also increase the false positive rate along with it and increase the number of podiums that you would predict and exceed the number that would be possible within the season.

It is most important for falsely predicted podiums to be low, because then we aren’t predicting someone getting a podium, and earning lots of points, when they don’t. Predicting a driver to not podium when they do podium is a problem, but it’s less severe than predicting a person to podium when they don’t get it. I have also found that most of the false predicted podiums were on drivers that didn’t typically podium and then were able to for that race.

I would say there are a lot more variables that could be investigated for this. We could investigate earlier years, or even how they performed when they were in different leagues. We could investigate their fastest laps more closely and how consistent they are. Another thing that could be interesting to investigate is how well the team they are driving for has performed over the last few years. All of these could be variables to investigate if I had more time, but it also could be tough to investigate these variables with the tools I have.

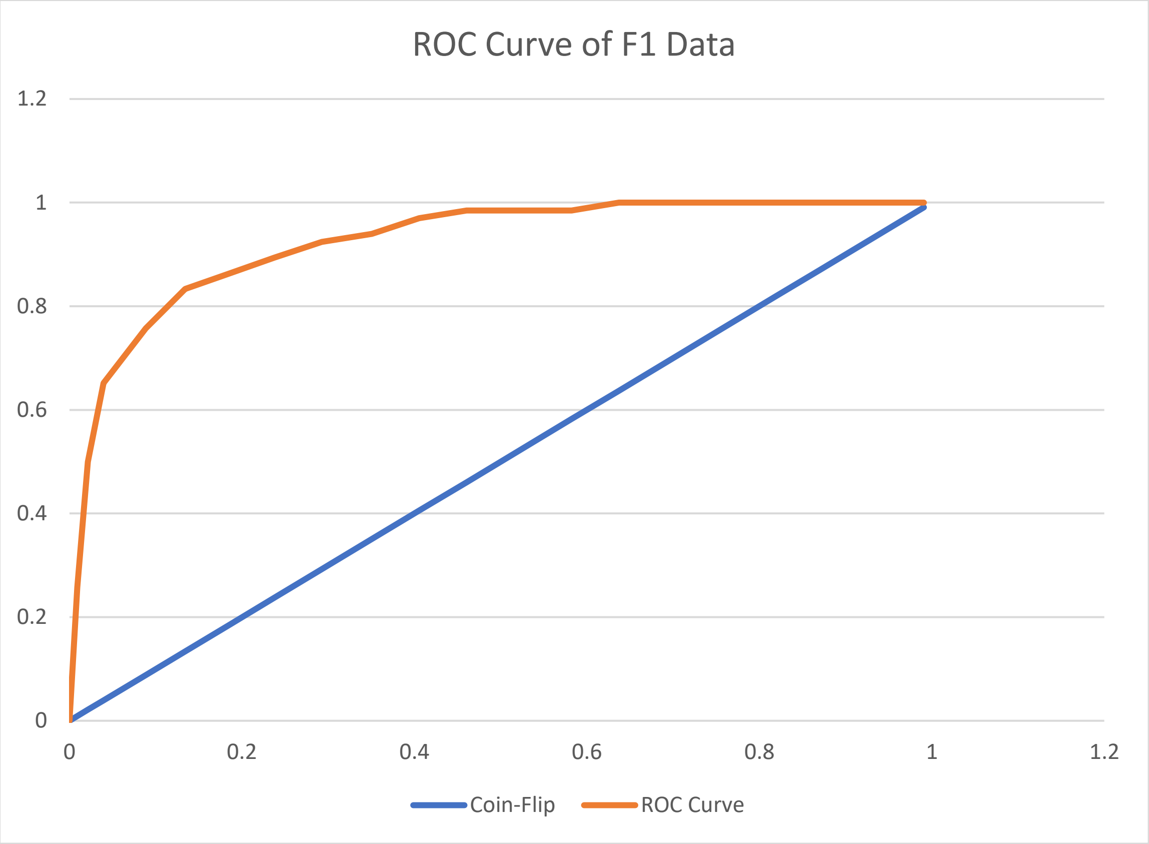


Figure 4

The ROC curve (*Figure 4*) shows how the Confusion Matrix is very accurate, at least compared to a Coin Flip. It shows that the matrix is accurate at getting true positives at the beginning, and then it starts to even out and eventually match up with the coinflip. It shows that the variables of grid position and what position they are currently in the standing have a high effect on if they happened to podium that race. The best we found was at 0.15 percentile that is where we found the predicted number of podiums to be closest to accurate number of podiums, and it was also the one with the highest accuracy. You can see this by looking at figure 3 and looking closely at 0.15 percentile.

In closing, I believe that I have proved that these two variables are very effective when predicting possible podiums. They both have high correlation with the classification variable and are also consistent for each race, when it comes to the certain drivers I am investigating, at least in 2021. Through investigating the data and forming the matrix and graphics from said data, I have shown that these have strong correlation and that there is much more to investigate with the data, as well.

Works Cited

“Database Images.” Ergast Developer API, 5 Mar. 2022, http://ergast.com/mrd/db/#csv.