ORIE 3120

Final Milestone- Data Analysis

Our group is performing data analysis on Formula One, looking specifically at tire performance. We used the FastF1 API, which has data from 1951-2022, including race, qualifying, and practice sessions, to (individually) perform an analysis on our questions selected.

Question: Can tire compound type be predicted from race data?

The goal of this question is to attempt to identify the tire used to set a lap during the race given only the number of laps completed on this tire, which race stint this tire represents, and the average lap speed during this race stint.

To answer this, we chose the 2021 Russian Grand Prix. This is because this race was extremely tire dependent, as it started raining about 7 laps before the race finish. It forced teams to consider on a lap-by-lap basis whether or not they should pit for the intermediate (green) tires, as they were better suited for the changing conditions. Race strategy is difficult to develop because it depends not only on your car's performance and ideal outcome but also on what other teams and drivers are doing. For example, if the leader is considering pitting for intermediates, they must consider not only whether they would be faster on the intermediate tires, but also whether they will be fast enough to catch up and make up for lost ground to the people behind them who did not pit. Hence, this race is good for developing a model to predict tire choice since tires played a crucial role in lap speed and final race position.

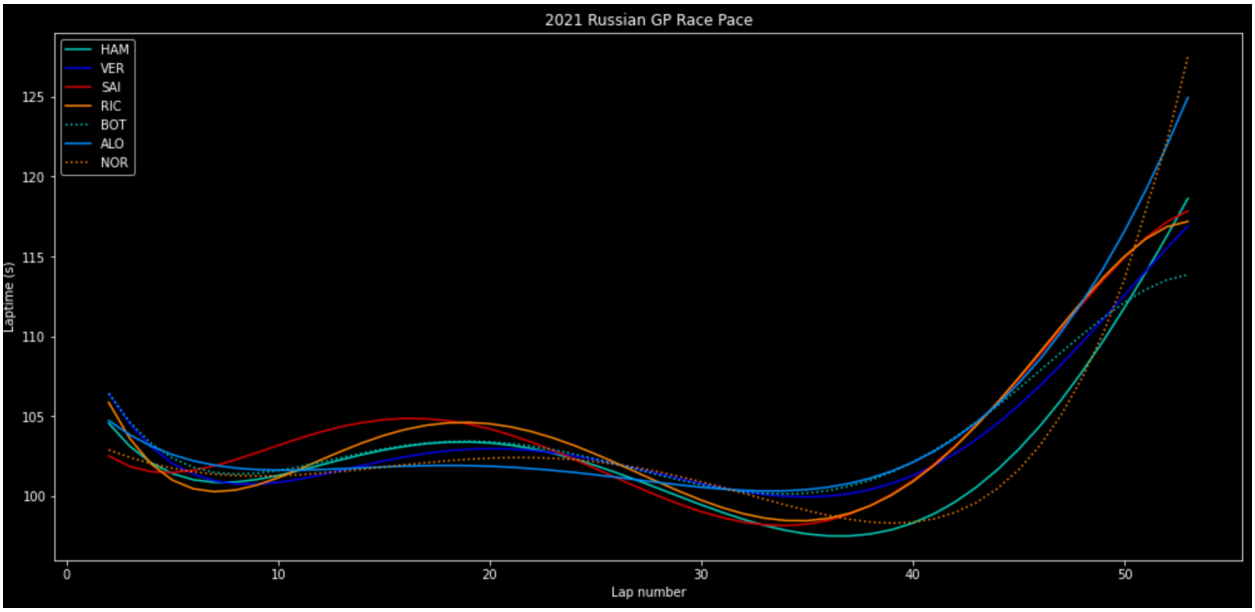

Figure 1: 2021 Russian Grand Prix Race Pace

Figure 1 shows the race pace for the top 7 finishers during the race. This is an important context because Norris (NOR- dotted orange line) was in first for most of the race, and was the fastest for much of it, as shown by his lap times being lower for most of the race. However, as mentioned, when the weather changed, he and his team had to consider not only whether he would be faster switching to intermediates, but also if the others behind him would pit as well, or he would have to catch up to make up for lost positions. Unfortunately for him, he pitted later

than his rivals and lost a lot of time because the hard compound tires no longer had grip on the rainy track surface, as reflected by his extremely high lap times at the end of the race. This visualization helps contextualize how tire strategy affects race performance in mixed conditions. Those with the fastest pace (lower on the y-axis) towards the end were those who pitted at the right time– when the intermediate tires became the best tire to use (BOT, VER, RIC). Everyone's race pace slowed towards the end due to the rain, but some faster than others. The next figure gives a clue as to what the optimal strategy was.

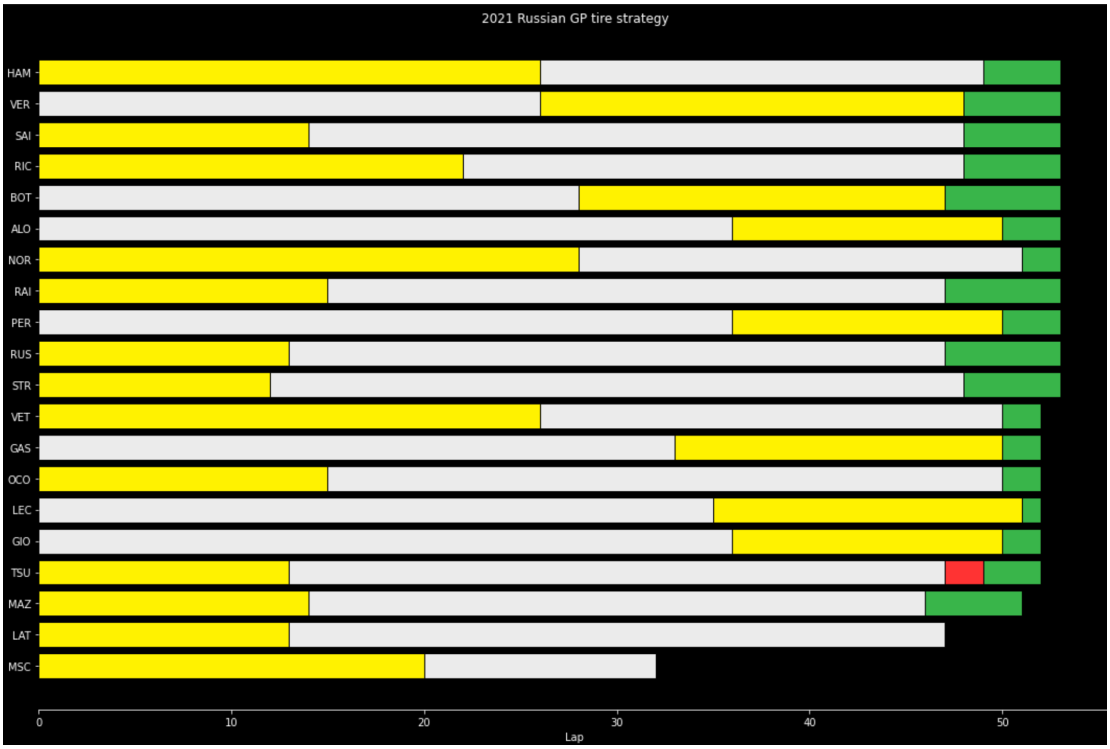

Figure 2: 2021 Russian GP Tire Strategy

Figure 2 is a visualization of the tire strategies each driver employed is also useful and it shows exactly when each driver pitted. Bottas (BOT), Verstappen (VER), and Ricciardo (RIC) were among the first to pit among the top drivers, and this paid off, as they finished second, fifth, and fourth, respectively.

Analysis

For the data analysis, we attempted to create a model that would successfully predict tire choice given the total number of laps completed on a certain tire, which stint during the race it is, and the average lap speed during that stint. A stint is effectively the number of different tire sets used in a race so far. If it's their third stint, that means these are their third set of tires and they have pitted twice. To do this, we used a multinomial logistic regression, with the permission of the instructor. This is because we needed to use a classification model, but since there are more than two options for tires, a logistic regression would not suffice. A multinomial logistic regression works like a logistic regression except it can classify 2 or more outcomes.

For the training set, we used data from the two free practice sessions held ahead of the race. Usually, there are three practice sessions, but the third was canceled due to torrential rain. For the X_train, we used the total number of laps completed on a certain tire, which stint during the race it is, and the average lap speed during that stint. The y_train was the type of tire used each lap. In total there were five possible tire choices: soft (red), medium (yellow), hard (white) , intermediate (green) , and wet (blue). Wet tires were never used during this race weekend, and so do not factor into the model.

The test data was the actual race lap data. For X_test, we used the total number of laps completed on a certain tire, which stint during the race it is, and the average lap speed during that stint. After completing the regression on the training set, we predicted using the test set, and then recorded the predicted tyre compound next to the actual compound, and made two plots. The first represents the actual tire compound used and the second is the predicted tire compound.

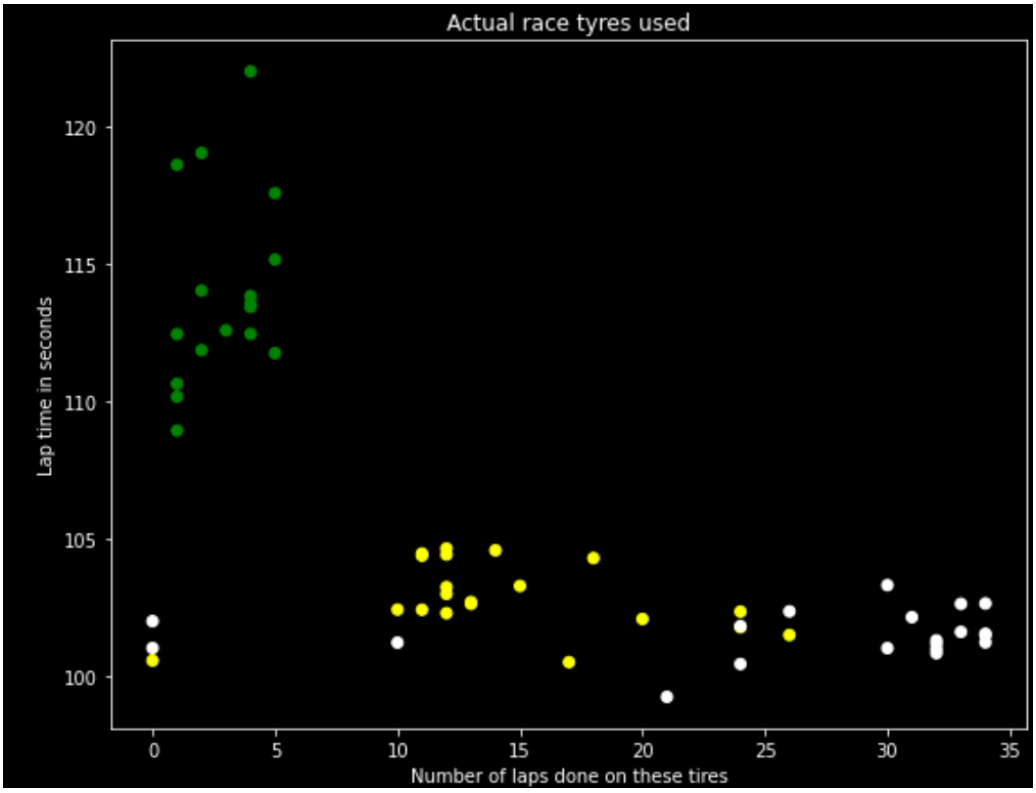

Figure 3: The distribution of tires used during the actual race

Figure 3 represents the actual tire compound used. These are the actual tires used for each lap during the race. Some things to note are that there is some clear clustering between the medium (yellow), hard (white), and intermediate (green) tire groups. This makes sense given some race context. In the bottom left corner are a handful of medium and hard compounds that were used for only a couple of laps. These are a group of drivers who had to pit regardless of the weather because their strategy utilized a long first stint. When they finally pit, the conditions were not rainy enough to warrant gambling on intermediates, and so they switched to medium or hard tires. Unfortunately, the conditions changed rapidly, and they pitted almost immediately again (onto intermediates) due to worsening conditions. In the bottom middle there is a cluster of

medium (yellow) tires. This makes sense, given medium tires are meant to last 10-20 laps. On the bottom right is a cluster of hard (white) tires, since hard tires last 25-35 laps usually. In the top left is a group of intermediate (green) tires. This is the most distinct group, with no overlap with the other tire compounds. This is typical of intermediate tires, as they are designed for mixed/ rainy conditions, and so are slower but provide the driver more control. They were only used for the last 5-7 laps of the race since that's when the rain began.

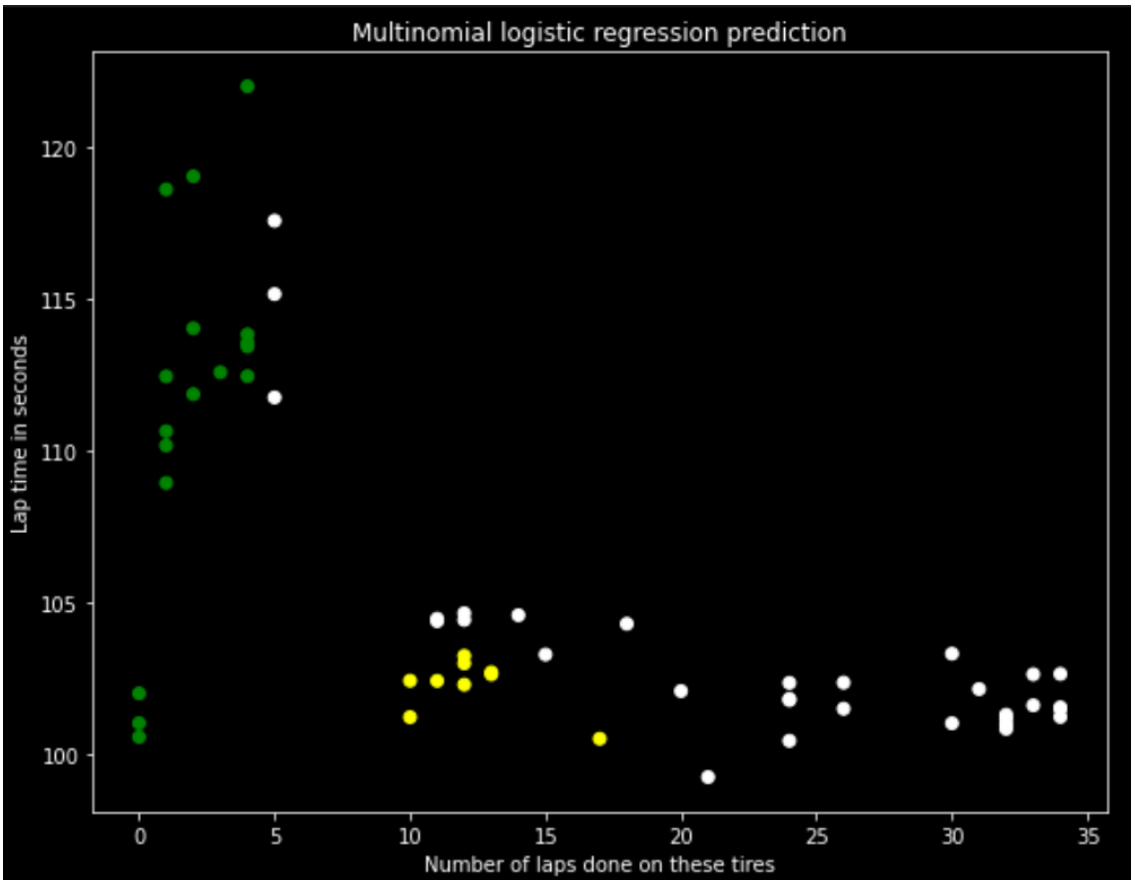

Figure 4: Multinomial logistic regression prediction of tire choice

This is our model for what type of tyre we predicted. As a review, this figure shows predicted tire compound choice. It uses tire life, which number stint this is, and the average lap time during that stint.

Looking at the plot, it is not perfect, but it does a respectable job at predicting the tire type. Most notably, it has a similar clustering to the actual tires used during the race. The top left is intermediates, the middle is all the mediums, and the right is the hard tires. It struggles to predict the bottom left and middle center. This makes sense considering the variables used in training the model and the context of the race. The bottom left would be difficult to predict because the tire life variable suggests that the intermediate tires were only used for a few number of laps. During the race, the bottom left tires are medium and hard compound because of failed race strategies, hence are effectively outliers. The middle center is difficult to predict and reflect a shortcoming of the model, as the training data likely suggested that hard and medium tires had

some overlap in terms of number of laps done, but that the hard tires were slightly slower. Hence why the bottom middle is correctly identified as medium tires but the center middle is mislabeled as hard tires. Overall, however, the multinomial logistic regression does a good job of predicting the tire choice.

Reflection

If we were to iterate on our analysis and consider some limiting factors in our model, the most important things to consider in creating an improved model are the variables not included. We included tire life, stint number, and average lap speed during each respective stint. Unfortunately, there are variables that should be considered but cannot because they were not included in the dataset. For example, things like track temperature matter, as each tire has an optimal track and tire temperature. If the track is too cool, teams are less likely to use certain compounds. This helps explain why soft tires were not a factor during the race– the optimal temperature was not matched by the track temperature, so it was avoided. As rain fell, not only did the track become slippery, it became colder. This would be useful to track and would influence our model positively. Related, weather forecasting would be useful for the model, as knowing when the forecast predicted rainfall would begin impacted the actual race strategies but is not reflected in our model.

Additionally, things like driver technical ability and race IQ are important in race strategy but not measurable. For example, Lewis Hamilton is known for being a great driver in mixed conditions, so as conditions worsened, he was still able to maintain pace and close the gap to his rivals. Additionally, he asked his team over the radio to consider pitting intermediates. His feel for the car suggested that the conditions were getting unbearable in hard tires, and so it would be smart to pit. He pitted while in second, and he initially lost places to drivers who did not, but as their race pace worsened due to heavier rainfall, he got quicker and finished first. Driver ability and race IQ are impossible to measure but play an important role in the race.

Overall, however, the model does a good job of predicting which tires are used, and most importantly correctly identifies the overall cluster and distribution of tire compound.

Conclusion:

This milestone attempts to answer questions about tire performance and strategy. To do this, we used multinomial logistic regression. Our model accurately predicted the tire choice given stint number, tire life, and average lap time for every stint.