Analyzing the Effects of Weather in Formula 1

Author: Abhi Gudimella

Introduction

In this project, I used the FastF1 library, which contained data about every Formula 1 race since 2018. The details of what variables I had access to is annoyingly complicated and tedious, since there are over 100 features combined across each dataframe.

I focused on the effects of weather between the Japanese Grand Prix in 2023 and 2024, which was hosted at the Suzuka race track both years. I picked these two specifically because there were no driver/team changes between 2023 and 2024. Therefore, the only variables changing would be the weather conditions. In 2023, the Grand Prix was hosted on September 23rd, 2023 when Japan experienced warm sunny weather and clear skies. On the other hand, in 2024, the Japanese Grand Prix was hosted on April 6th 2024, when the weather conditions are colder and cloudier.

My variables mainly consisted of weather data of the track (air temperature, humidity, track temperature etc.) and telemetry data of each car (lap times, sector times, speeds etc.). This data is recorded by the API used by the FastF1 Library at regular intervals throughout the race weekend. The interval seems to be around once a minute but varies. Therefore, although the library provides weather and telemetry data using different data frames, I was able to sync it up using the timestamps when these were recorded. The data is also synced up to each car, driver, and lap.

Using this library, I was able to pull data for the Japanese 2024 Grand Prix, Japanese 2023 Grand Prix, and in fact, data for the entire 2023 season. The details of how the 2023 season dataset was used are mentioned later. All 3 of these data frames consisted of the same features that were mentioned above.

This project is important since understanding the effects of weather will help engineers in the industry understand how to modify and optimize their car for races which are held all throughout the world all throughout the year in every season. Additionally, the project also dives into the ins and outs of understanding model performance in the context of Formula 1. Since margins in Formula 1 are so precise and differences between teams are next to nothing, a model's performance needs to be highlighted a lot more than models in other contexts.

Question #1: Can we predict performance for the 2024 Japanese Grand Prix using only the data from the 2023 Japanese Grand Prix?

Methods

To answer this question, I trained a linear regression model to use data from the 2023 Japanese Grand Prix to model the performance of the 2024 Japanese Grand Prix. The goal was that an analyst in an Formula 1 team, heading into the 2024 Suzuka race would know the weather forecast. Therefore, they would try to build a model using the data they had from the earlier year and run it on the weather data they saw forecasted.
In terms of pre-processing for this model, I separated the predictors into categorical variables and continuous variables. Refer to the code to see which variables are which because there are a lot of variables! The continuous variables were z-scored and the latter one-hot-encoding to scale them all the same for the model.

Results

The Linear Regression model returned the following performance metrics.

Train MSE : 9.24890541776849e-11
Train MAE : 7.640473423415946e-06
Train MAPE : 7.722713625006928e-08
Train R2 : 0.9999999999671673

Test MSE : 0.0011828715739081108
Test MAE : 0.023699127591595464
Test MAPE : 0.00023986268972415556
Test R2 : 0.999938594903639

Discussion

First, it should be noted that something that is really nice about the way my data was set up was that I was able to divide my data into present data (from 2023) and future data or test data from (2024). Therefore, the model is trained on 80% of the 2023 data. Therefore, the train metrics above pertain towards the rest of the 2023 data. More importantly, the test data and test metrics pertain to the 2024 data.
Obviously the R2 score is really high, which would mean there is a strong correlation between the 2023 and 2024 data. However, this is to be expected and R2 is actually not a good

measure of the model's quality. Since Formula 1 is won and lost by mere seconds and sometimes even less the data is very closely packed together. In other words, absolutely speaking, there is little to no difference between lap times across both years. However, looking specifically at the mean absolute error, we see that the test data has a worse MAE by a factor of over 1.000,000. Moreover, over 50-60 laps, being off by a factor of 1,000,000 on your predicted lap time for each lap is pretty much a guaranteed loss in terms of building and picking pit stop strategies for the race.

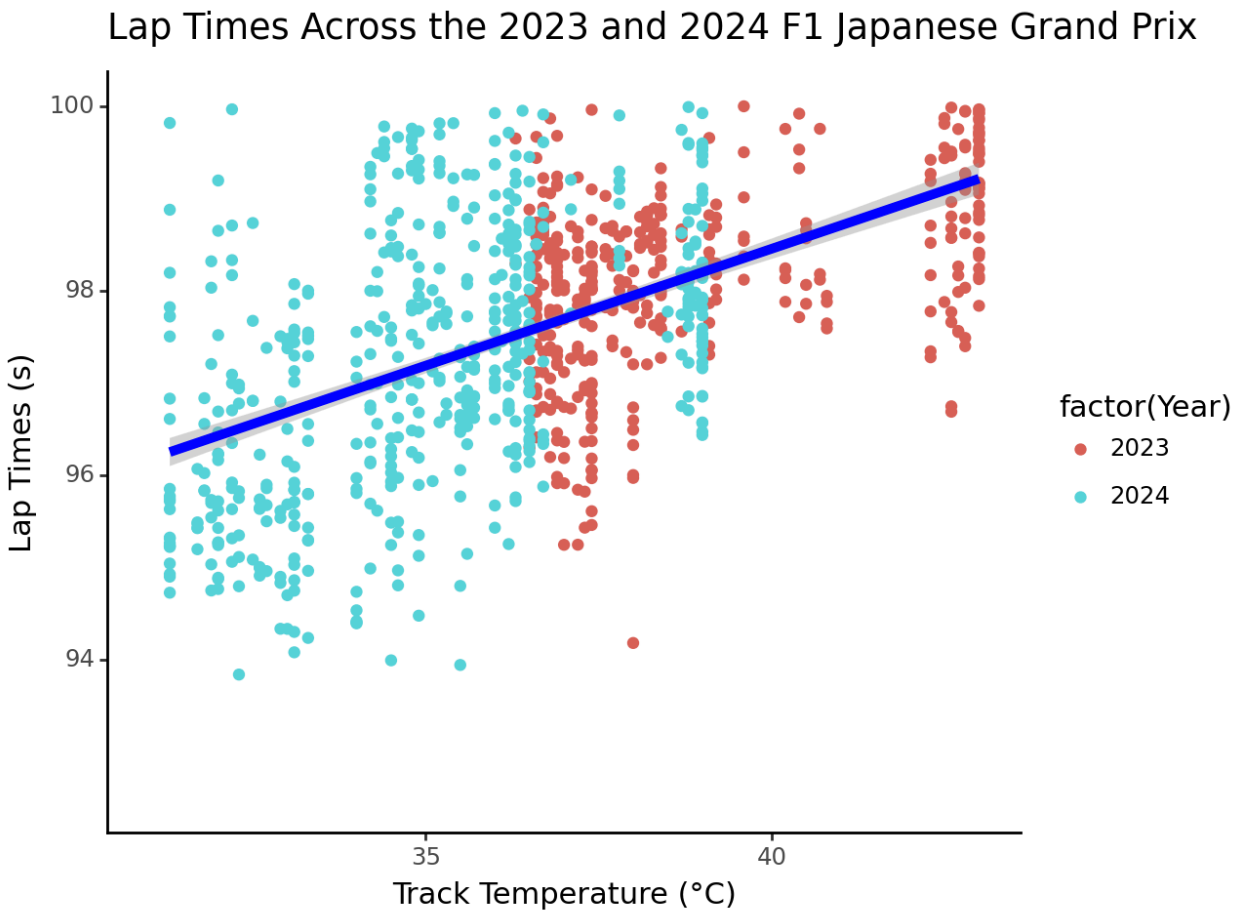

Figure 1: A scatter plot with a linear regression fit to it that shows Lap Times across the 2023 and 2024 Japanese Grand Prix. It has Track Temperature in Degrees Celsius on the X-Axis and Lap Times in seconds on the Y-Axis. The data fits well to the line and is positively correlated, but is split almost down the middle with 2023 having higher temperatures than 2024, which has lower temperatures and slower lap times.

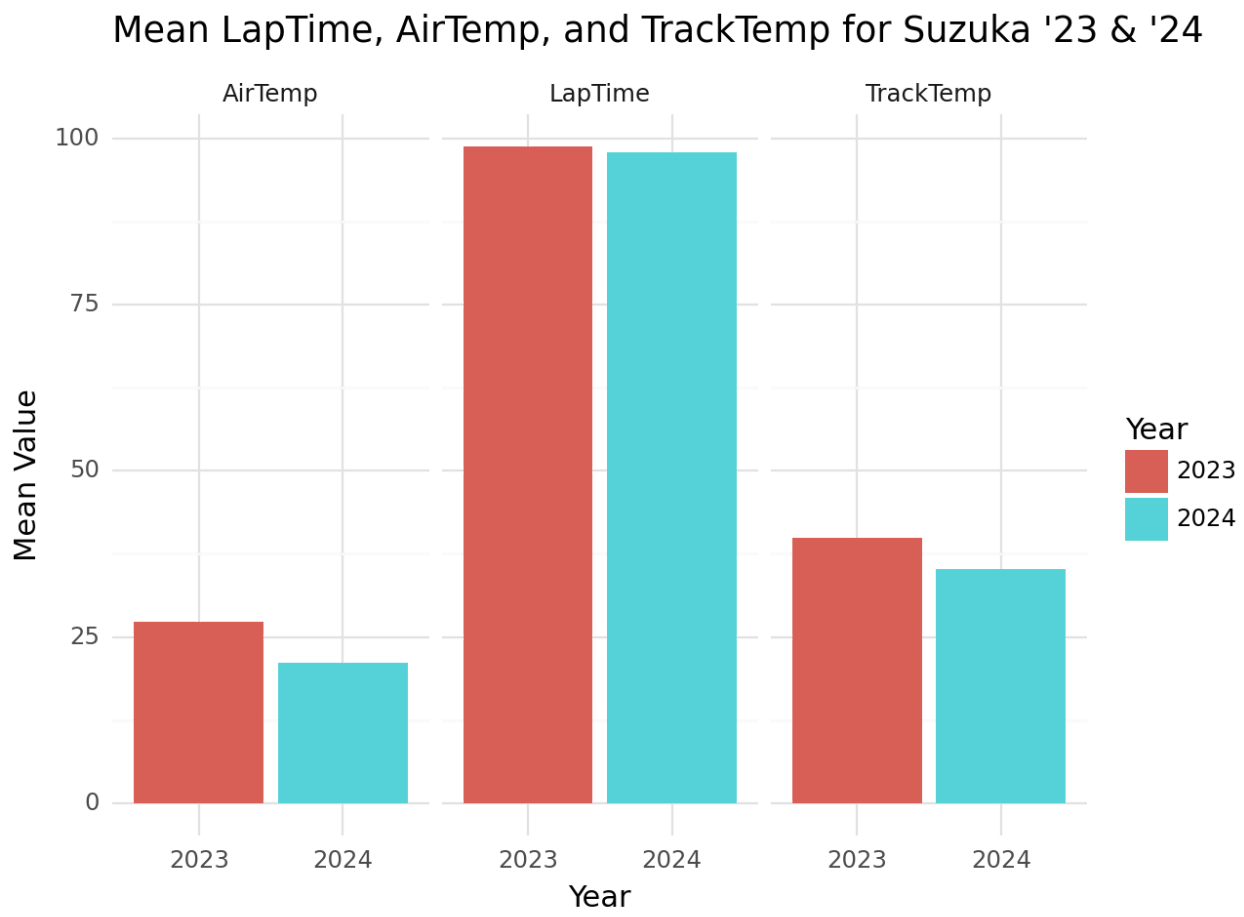

Figure 2: This bar graph shows the differences in mean air temperature, track temperature, and lap times across 2023 and 2024 at Suzuka. 2023 shows a significant increase in temperatures and an almost 1 second increase in lap times than 2024.

Question #2: Can we find other tracks similar to Suzuka that we can use to predict performance for the race in 2024?

Methods

Initially, we trained our lap time predictor on Suzuka 2023. I wondered if there were other tracks more similar to Suzuka 2024 given that the weather had now changed. In order to move forward, first I had to generate a dataset that was based on tracks rather than laps and drivers. Therefore, using pandas, I was able to group each track together. Therefore, I could find the mean, median, standard deviation, and other metrics for each track by itself. Then, I could train a clustering model to find the closest track to Suzuka 2024.

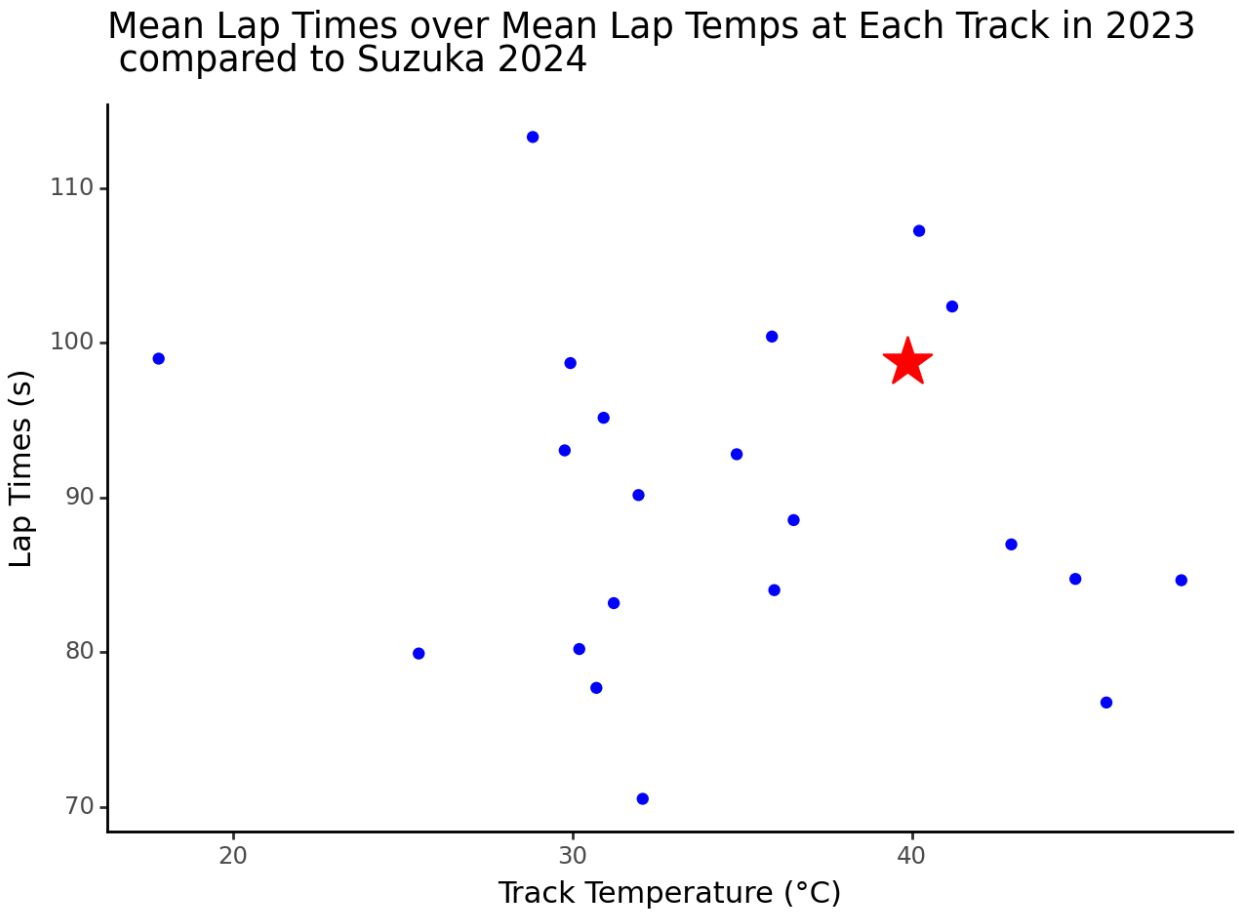

Figure 3: Here, the red star represents Suzuka 2024 and the rest of the blue dots represent all the races from 2023. With Track Temperature in the x-axis in degrees celsius and Lap Times on the y-axis in seconds, the graph is a scatter plot representing the relationship of each track in the 2023 season with respect to each other as well as Suzuka 2024.

I ran a KNN model (which finds the 'k' number of nearest neighbors) to find the closest tracks to Suzuka 2024 to help understand data from which track would be the best to use to train a model that predicts lap times for Suzuka 2024. The closest neighbor to Suzuka 2024 that this model returns should also be closest to the red star on the graph.

The pre-processing for the KNN model was just z-scoring all the continuous variables because the new dataframe I created only had continuous variables.

Results

The model worked well to initially find the closest 4 neighbors, which I then trimmed down to the closest neighbor in order to avoid confusion. Surprisingly, we saw that the British Grand Prix 2023 was the closest neighbor to the Japanese Grand Prix 2024.

```
[156] nn = NearestNeighbors(n_neighbors= 1)
      pipe = Pipeline([('z', z), ('model', nn)])
      pipe.fit(tracks[predictors])

      distances, neighbors = pipe.named_steps['model'].kneighbors(pipe.named_steps['z'].transform(suzuka_2024[predictors]))

[160] suzuka_2024 = suzuka_2024.assign(neighbors = list(neighbors))

      #Assigning the closest neighbor
      neighbor_name = tracks.iloc[neighbors[0]]['Event'].tolist()[0]
      neighbor_name

      'British Grand Prix'
```

Discussion

Although one may expect that Suzuka 2023 would be the closest neighbor to Suzuka 2024 despite the weather, I very interestingly found that, in fact, the British Grand Prix from 2023, referred to as Silverstone 2023 from here on out, is a closer neighbor. Initially, this is very surprising and almost makes no sense.

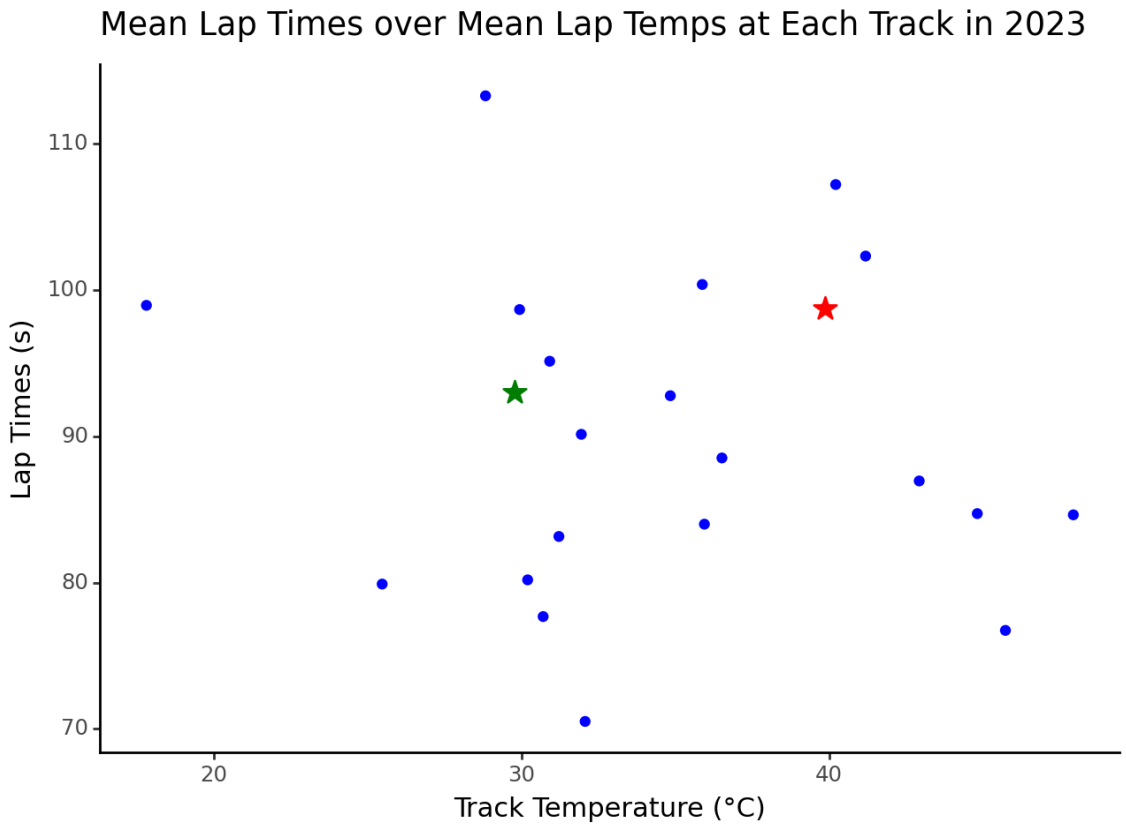

Figure 4: This is the same graph as Figure 3, but adds a green star decently far away from the red star (representing Suzuka 2024). In this graph the green star represents the British Grand Prix 2024, which was classified as the closest neighbor or Suzuka 2024 in terms of the data I have access to.

However, the results of this model helped me understand the importance of domain knowledge. Going back and looking through the F1 races from 2023, we can see that the weather during Silverstone 2023 is quite similar to what we saw in Suzuka 2024.

It should also be noted that a NearestNeighbors model may not be complex enough to group or cluster these tracks together. In the process of landing on a NearestNeighbors model, I had also implemented other clustering models such as KMeans, GMM, and DBSCAN. However, these did not churn out promising results and were predicting each track as a cluster in itself. Hierarchical clustering led to some useful results, being able to actually cluster each track into 5 groups, but also did not lead anywhere.

Although I ran a Linear Regression model trained on the British Grand Prix trying to predict the lap times for Suzuka 2024, I don't feel entirely confident in its results. Here they are:

Train MSE : 3.8735381706952925e-11
Train MAE : 4.358219899180664e-06
Train MAPE : 4.684104385700185e-08
Train R2 : 0.9999999999663811

Test MSE : 0.004460229079233006
Test MAE : 0.06557504275380507
Test MAPE : 0.0006686444814603533
Test R2 : 0.9997684610887235

The Train Metrics refer to the model's ability to predict Silverstone 2023 and the Test Metrics refer to its ability to predict Suzuka 2024. The results are noticeably worse than those we got from the linear regression model in Q1 by a factor of almost 10,000.

Question #3: Does weather really play as big of a factor as other features such as drivers, cars, teams etc.?

Methods

To answer this question, I ran a Lasso Regression model that would help determine how useful each feature was. The preprocessing done was the same as in the previous two Linear Regression Models. Since we saw in the discussion for the previous section that the linear regression model in Q1 outperformed that in Q2, this Lasso regression model is trained on Suzuka 2023 to predict Suzuka 2024 like the former.

Further after running the model, I extracted the coefficients and respective penalties given by Lasso in an effort to understand which variables were prioritized.

Results

The model returned the following performance metrics.

Train MSE : 3.7838006631766865e-06
Train MAE : 0.0013506388249169795
Train MAPE : 1.3603881684941719e-05
Train R2 : 0.9999986567882244

Test MSE : 2.8475961903943417e-05
Test MAE : 0.0026362954341780645
Test MAPE : 2.5838414301817895e-05
Test R2 : 0.9999985217590622

Discussion

The Lasso Regression seemed to perform similar to the previous models with the metrics across the board being similar. This makes sense because even a Lasso regression model is a Linear Regression model with certain other complexities like penalized coefficients.
On that note, the model also returned the following penalized coefficients.

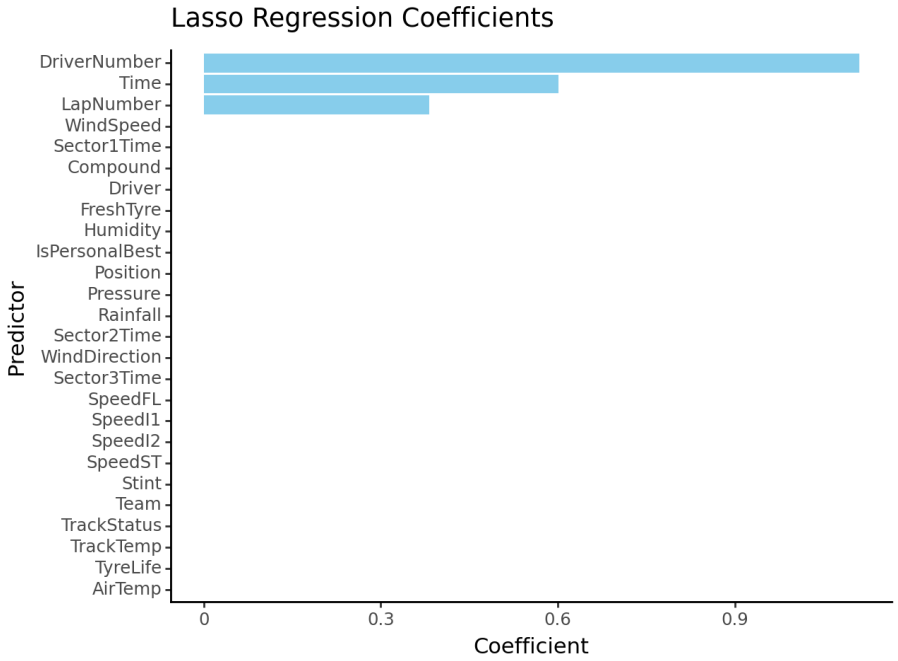

Figure 5: A bar graph showing the coefficients of each predictor (scaled by z-scoring). The only 3 coefficients with nonzero values are DriverNumber, Time, and LapNumber

The results from this model put a lot of things in perspective. The Lasso Regression zeroed out all coefficients to do with weather, showing that the weather played no significant role in

changing the cars' performance. This is a little jarring with the results we found in previous models, but paints an overall big picture by showing why all the performances were similar. There were no real variables changing and the only ones that did change (the weather) were zeroed out.

Reflection

This project was tough and I hated not being able to build good models. It seems like the models I know of are not complex enough to model something like Formula 1 especially without any first hand data.