**Formula 1 Tire Predictor Using Neural Networks and LSTMs**

Karina & Abhi

***Project Objective:***

This project aims to enhance race strategy planning and decision-making through machine learning models that predict lap times and tire strategies so teams can optimize their performance, minimize pit stop times, and gain an advantage in Formula 1 races.

***The Data:***

The data was obtained using the FastF1 Library, a Python-based tool that allows interaction with publicly available Formula 1 timing, telemetry, and weather data. The FastF1 library utilizes the Ergast API that provides a lot of information, including lap-by-lap timings, telemetry data (ex: speed, throttle, brake, gear usage etc.), and session metadata (event name, location etc.) and information about the weather (track surface temperature, air temperature, humidity etc.). Additionally, it offers information on tire compounds, stint durations, and pit stop events, which are crucial for analyzing and predicting tire strategies.

The dataset preparation process began by querying specific race sessions from the library. These sessions were selected based on relevance to the project goals. Since the goal was to focus on lap times and tire strategy, with a particular focus on extracting detailed stint information and lap-by-lap telemetry data, the data comprises information about every lap for every driver from 2020 to 2024. Since the data was pulled before the end of the 2024 season, the dataset consists of all but one race from the 2024 season. The raw data, once retrieved, was organized into two structured pandas DataFrames–Lap data and Weather data.

Data cleaning and pre-processing were key in transforming the raw data into a usable format. Steps included handling missing or incomplete telemetry values, aligning data points across multiple sessions, and encoding categorical variables such as tire compounds. Certain transformations were applied, such as normalizing telemetry values and converting timestamps into the number of seconds from the start of the race for analysis. The final dataset (made by combining the lap and weather data) was designed to capture a race's strategic profile, featuring columns for lap number, driver identifier, tire compound, stint duration, average lap time, and pit stop deltas. This dataset provided a comprehensive view of driver behavior and strategy execution, making it well-suited for this project.

***Exploring the Data:***

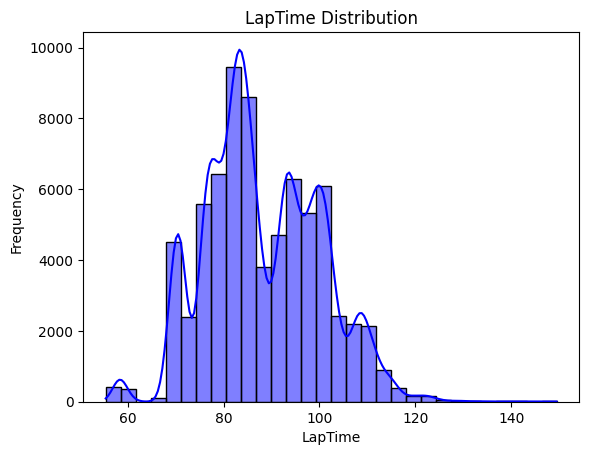


Figure 1: Frequency of LapTimes

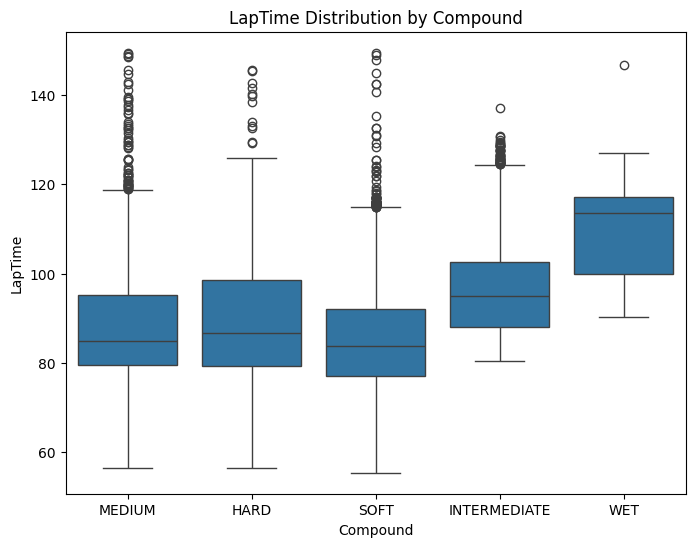


Figure 2: Laptime Distribution by Compound (Tire Type)

***Models and The Questions We Want to Answer:***

We split the project into two parts. The first was to predict lap times and discuss the effects of tire compound and weather in Formula 1. This was answered using a Neural Network. The second was to predict tire strategy (i.e., when to do a pitstop for a driver and what tire compound to put on the car) throughout each race. This was answered using an LSTM.

***The Neural Network to Predict Lap Time:***

To predict lap times, the model used both tire-related and weather-related features. Tire-related features include Compound, the type of tire, TyreLife, and FreshTyre, which indicates whether the tire is fresh (0 = No, 1 = Yes). Weather-related features include AirTemp, Humidity, Pressure, and Rainfall (0 = No, 1 = Yes), as well as TrackTemp and WindSpeed.

The data was cleaned and pre-processed to prepare it for the neural network. First, missing values were removed, and the first column was dropped because it contained values starting from 1, 2, and 3, which were removed and adjusted to start from 0 for consistency. Next, the categorical feature, ‘Compound, was encoded ordinally, the continuous variables were standardized, and the data was split into training and testing sets, with an 80/20 split.

The model uses a feedforward neural network that is well-suited for tabular, structured data without sequential dependencies. To balance performance and complexity, the model includes four layers: one input layer, three hidden layers (with 128, 64, and 32 neurons), and one output layer. The network was trained for 100 epochs, during which the loss steadily decreased as the model converged. The Adam optimizer was used with a learning rate 0.0001, which allowed the model enough time to reach convergence without overshooting.

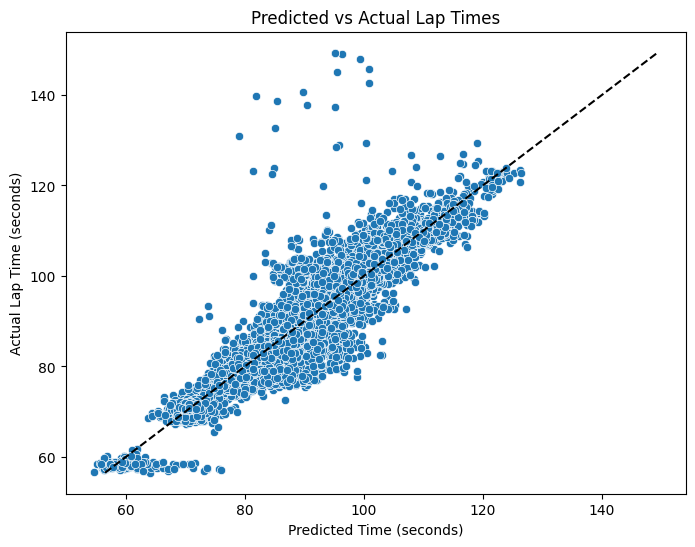


Figure 3: Model’s Prediction for Lap Time versus Actual Lap Time

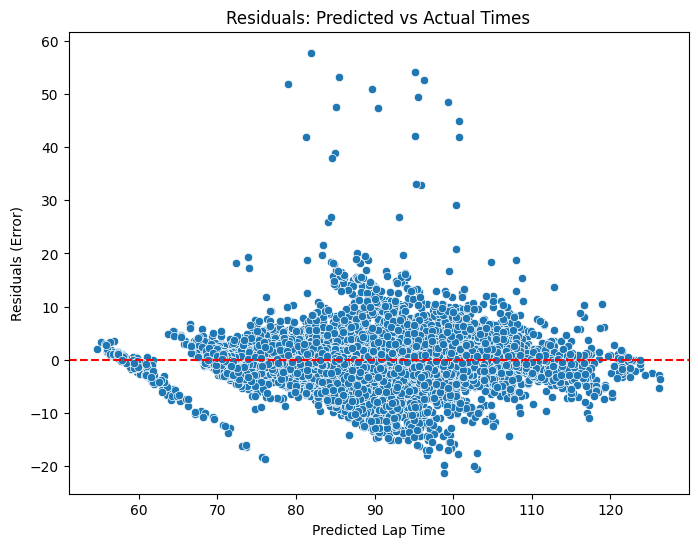


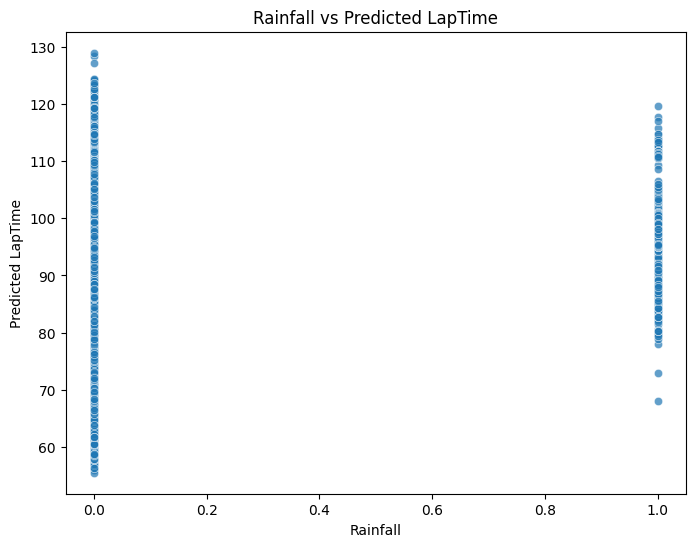
Figure 4: The Residuals of Predicted Lap Time versus Actual Lap Time

The model's performance was evaluated using both visualizations and metrics. In Figure 3, which shows predicted lap times versus actual lap times, most points align closely around the x = y line, with only a few outliers. This indicates that the model does a good job of capturing the relationship between input features and lap time.

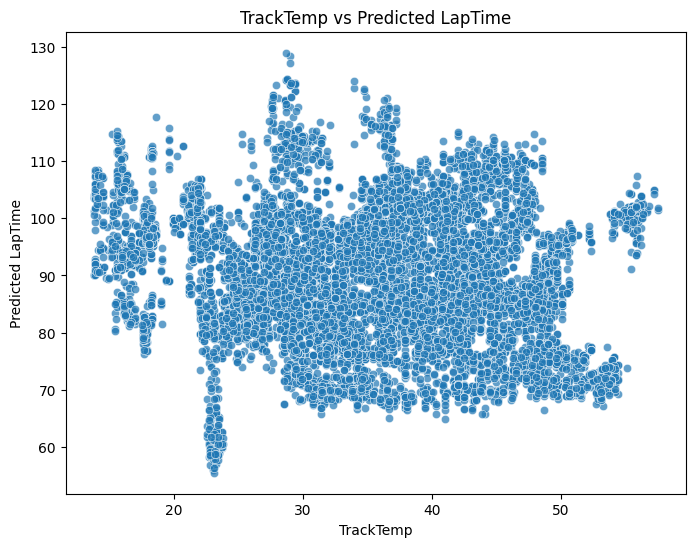


Additionally, the training and testing MAE (Mean Absolute Error) are both around 2.6 seconds, meaning the model's predictions are typically off by about 2.6 seconds. The test set's RMSE (Root Mean Squared Error) is approximately 4.04 seconds, representing the average prediction error. These results suggest the model performs well overall, with predictions quite close to the actual lap times.

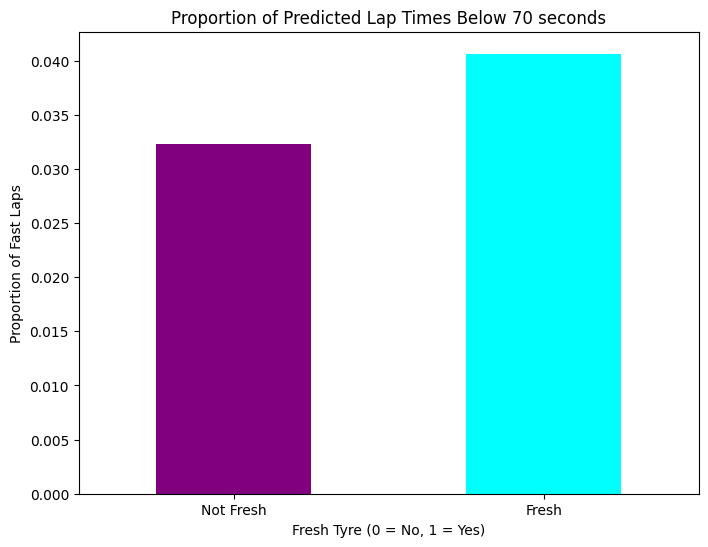
*How do features like tire type and weather affect the lap time?*



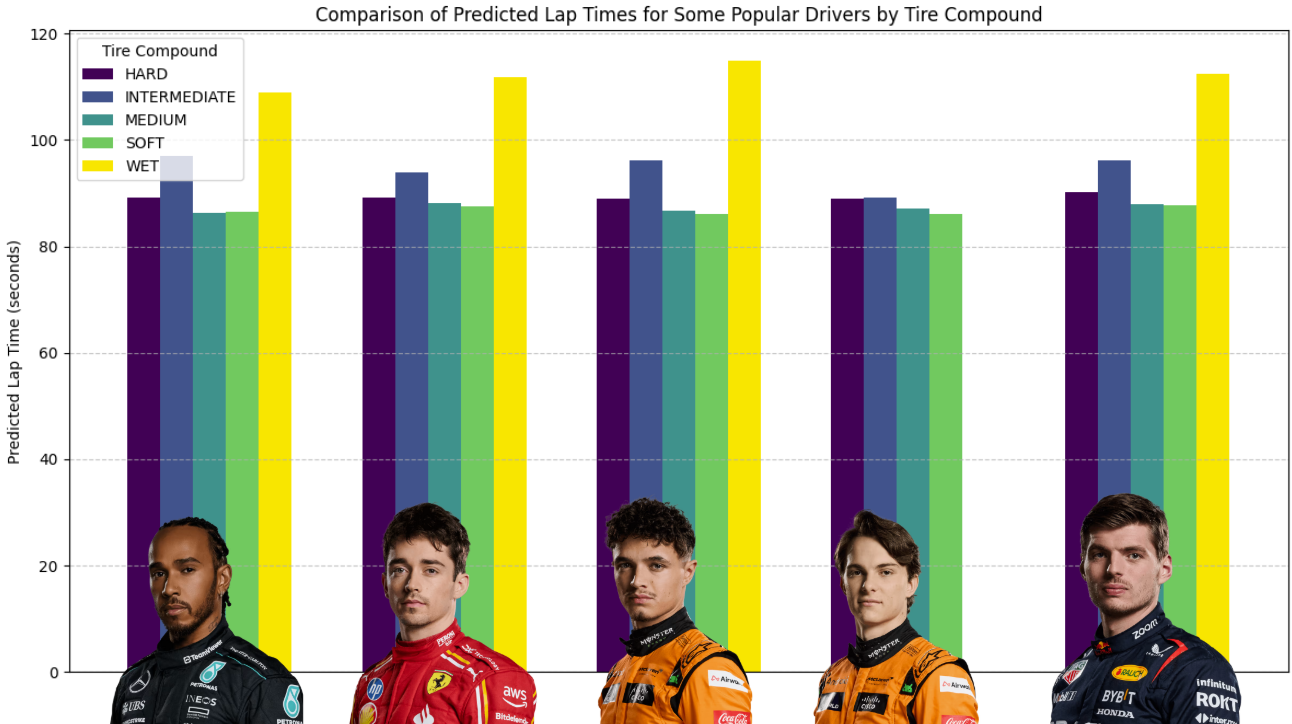
The graph comparing the rainfall to predicted lap times shows that laps completed during dry conditions have shorter lap times compared to when it is raining. This suggests that rainfall significantly slows down performance on the track.



Analyzing the track's temperature and predicted lap times shows that laps completed when the track temperature was between 22-23 degrees Celsius resulted in the shortest lap times. This suggests that this temperature range is optimal for performance.



When applying a threshold of 70 seconds to define a "fast lap" and examining the proportion of fresh versus non-fresh tires, the results indicate that fresh tires are more common in fast laps. This shows the advantage fresh tires provide in getting shorter lap times.



When comparing the tire compound types of five of the most popular drivers (based on the formula1 website), it reveals that soft and medium tires consistently have the shortest lap times for all the drivers, while wet tires result in the longest lap times.

Overall, the model performs well in addressing the question we aimed to answer and provides valuable insights into lap time predictions. However, there is room for improvement by developing separate models for each event, as variations in lap lengths between events may impact the model's accuracy.

***The LSTM To Predict Best Tire Strategy:***

LSTMs are well-suited for time-series data due to their ability to capture long-term dependencies and sequential patterns throughout any sort of time-series data. Therefore, we chose to employ an LSTM to predict tyre compounds and pit laps. By processing all the features available in the dataset the LSTM aimed to uncover the underlying patterns that inform strategic decisions. In the real world, this might mean something more visual, such as the effects of the weather or wear and tear of the tyres. This approach is a data-driven (haha, because race cars!) way to forecast the intricate strategies employed by teams during a race. Continuous data (such as lap times and speeds) was z-scaled, while categorical data (such as the tyre types) was one-hot-encoded.

An initial model that was trained on minimal features describing the car’s telemetry and weather struggled to do well. After some investigation, it was found that the model wasn’t really the issue, but it was the data it was being fed. The organization of the dataset used for this project introduces a significant challenge when employing an LSTM network for prediction. The dataset is structured lap-by-lap, with all laps for a single driver in a specific race grouped together, followed by the laps of the next driver in that same race, and this pattern repeats across all drivers and races. While this format ensures that all relevant data is captured and arranged sequentially, it poses a problem for the sequential structure that LSTMs rely on. When sequences span the boundaries between different drivers or races, the LSTM unintentionally incorporates data that should not logically be part of the same time series. This overlap occurs because the sliding window does not inherently recognize the data being segmented for different drivers or races.

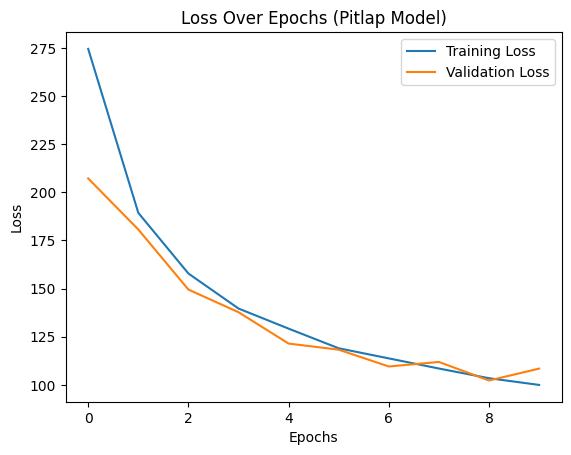
These issues degrade the model’s ability to discern meaningful patterns. The model assumes that every window of data it saw all belonged to one driver at one race. This violates the underlying assumption of the model and can lead to unintentional patterns being recognized, as the dynamics of one driver’s performance or the conditions of one race are unlikely to have any direct relevance to another driver or race. Moreover, this problem undermines the interpretability of the LSTM’s output. Predictions made on mixed sequences lack context since they are influenced by unrelated segments of data. Although the goal of the model was to make an accurate prediction about pit stops, it’s difficult to actually draw a meaningful conclusion from any prediction the model makes.

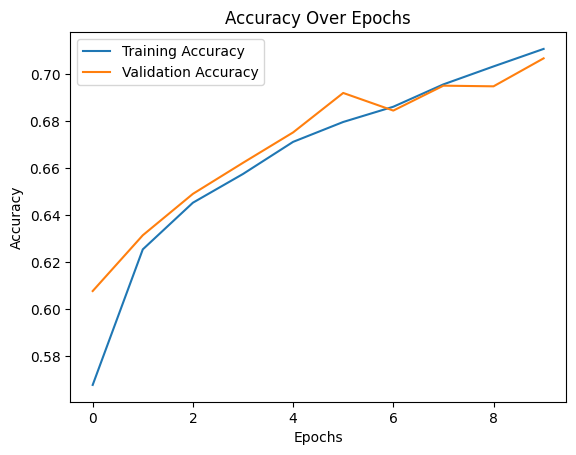
To address this issue, there needed to be another layer of preprocessing to ensure that the sequences fed into the LSTM make sense. Each sequence must consist solely of laps from the same driver within the same race. By maintaining the integrity of each time series, the LSTM can focus on extracting meaningful temporal patterns specific to individual drivers and races, leading to more accurate and interpretable predictions.

Then a second model was trained on data with the addition of a new feature into the dataset. The new feature was a driver-race key describing which race and what driver the data came from.

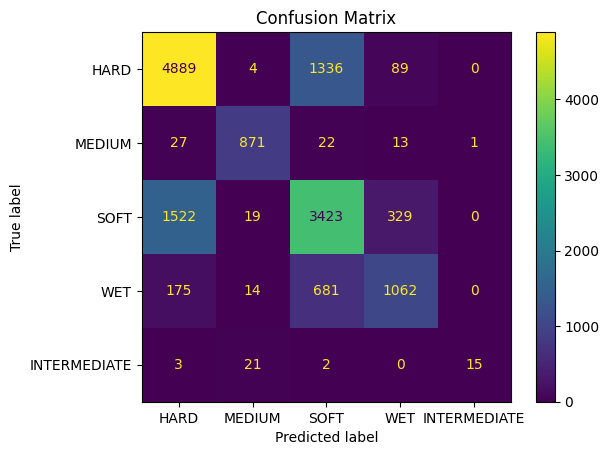


This model performed a lot better, yielding a loss (predicting the pit lap number) of 107 and an accuracy (while predicting the tyre compound) of 70%. Further investigation into the performance of both models also showed that there was no overfitting or underfitting, as seen in the graphs below.





Additionally, the tire compound predictor had some interesting elements. As seen below, the model seems to do exceptionally well at recognizing when it is time to put Hard tires on a car. It’s also good at recognizing when Wet tires should be put on. The latter seems rather obvious because it should be quite obvious when Wet tires are needed– in the rain. However, the former isn’t as simple to understand. One hypothesis we came up with was that the model wasn’t necessarily good at predicting when to put on Hard Tires. Instead, the model just recognized the fact that Hard tires are generally the most common race tires in most instances. Therefore, a model predicting Hard Tires 100% of the time would still be correct most of the time. This can be seen in the Confusion Matrix below; there are almost 6000 laps done on Hard Tires–a number to which no other tire compound comes close.



It’s also important to note that these models were only trained on 10 epochs each. Both models show room for improvement in their accuracies and loss, given more time and computing resources. In fact, at just 15 epochs, the models were able to obtain a validation loss of 94 laps and a test accuracy of 82%. However, due to the wildly large amount of data that the models are trained on, it takes about 2 hours for each iteration to fit the models on our computers. If these models were equipped with the million-dollar research and development that Formula 1 teams have access to, there would surely be an increase in the performance of these models.

Overall, the models were moderately successful at predicting the tire strategy for any driver. There are improvements to be made (such as the computing power suggested above). However, any significant improvement can only be made with access to better data. If this idea of using LSTMs to predict tire strategy was employed by teams who have access to data directly from sensors mounted throughout the car that come in live during a race, one would assume the models would perform better. This would be an interesting experiment in the future if such data were made available to the public.