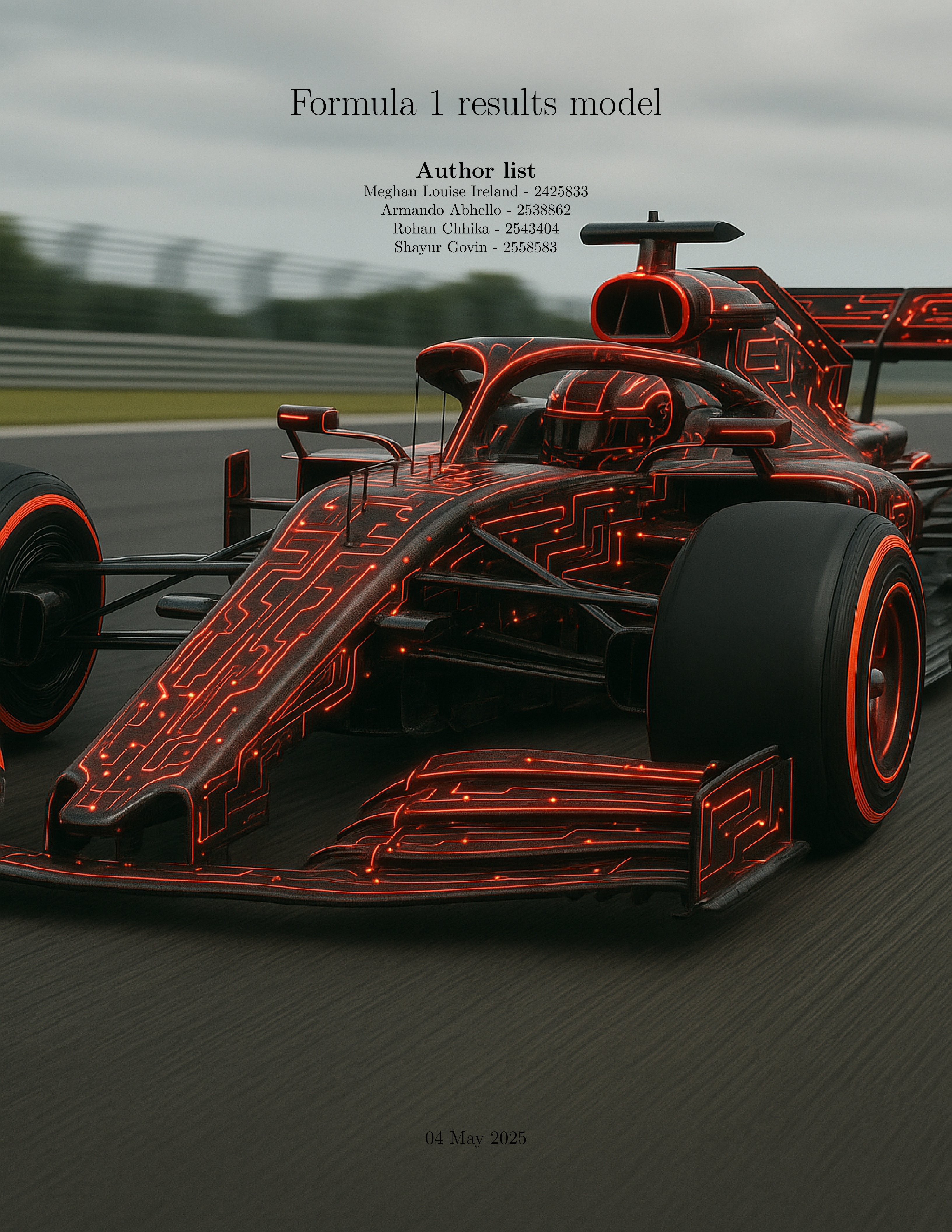


Formula 1 results model

Author list

Meghan Louise Ireland - 2425833
Armando Abhello - 2538862
Rohan Chhika - 2543404
Shayur Govin - 2558583

04 May 2025

1 Introduction

This report outlines the analysis of a Formula One data set using machine learning techniques. The dataset encompasses various aspects of Formula One racing, such as driver performance, race results, lap times, weather, and circuit characteristics, which provide a rich foundation for exploring patterns and predicting outcomes.

2 The Data

2.1 Data Source

The dataset used to train our model was retrieved from the **OpenF1 API**, which provides comprehensive real-time and historical Formula 1 data. This includes detailed information for all practice sessions, sprints, qualifying rounds, and races across a season. The table below summarizes the specific API endpoints used to construct our dataset:

API Endpoint	Description
`https://api.openf1.org/v1/drivers`	Retrieves information about each driver, such as name, number, and team affiliation.
`https://api.openf1.org/v1/laps`	Provides lap-level data used to determine final race positions and compute average lap durations.
`https://api.openf1.org/v1/sessions`	Lists all sessions in a season, including metadata like the session key and meeting ID.
`https://api.openf1.org/v1/weather`	Supplies weather conditions for each session, which may influence race performance.

Table 1: OpenF1 API endpoints used to build the model dataset

2.2 Data Content

The raw data from the API consists of data from various F1 sessions, including Practice 1, Practice 2, Practice 3, Sprint Qualifying, Qualifying, Sprint Race, and Race from various Formula 1 circuits. A sample of this information can be seen in Fig 10. For each session, the following data is contained:

Table 2: Structure of API Output for a Race Session

Position	Driver Name	Driver Number	Time (s)
1	First Last	#XX	123.456
2	First Last	#XX	123.789
⋮	⋮	⋮	⋮
20	First Last	#XX	130.000

Session Metadata
Meeting Key
Session Key
Track Temperature (°C)
Air Temperature (°C)
Pressure (hPa)
Humidity (%)
Wind Speed (m/s)
Wind Direction (°)
Rainfall (mm)

3 Objective

The primary goal of this project is to develop a machine-learning model that accurately predicts the average lap times of Formula One drivers using session-specific and environmental metadata. These features are sourced from the OpenF1 API and include variables such as air temperature, humidity, driver identifiers, and race session types. An additional objective is to explore the feasibility of classifying the final race positions using these features, which can be used for various purposes such as predicting race outcomes, predicting driver's and constructor's championships, and creating power rankings of the top drivers. This presents a useful and practical stretch goal for the project. Due to the unpredictable and volatile nature of F1, this will be no simple task. As such, we will be observing our accuracy with margins of error, particularly 1s, 2s, 5s, and 10s errors.

4 The Model

4.1 Model Selection

We implemented a feed-forward neural network for this regression task. Its ability to capture complex, nonlinear relationships within the data made it an ideal choice. Each version of the model contains multiple fully connected layers with ReLU activations. Later versions include dropout layers to reduce overfitting.

4.2 Version 1: Baseline Model

File: `f1-model.py`

- We built this initial model by keeping things simple and focusing on the 2024 season.
- We applied one-hot encoding to the categorical fields: `session_type`, `year`, `driver_number`, `driver_name`, `position`.
- We then fed this data into a 3-layer Multi Layer Perception with 64 neurons in our first hidden layer and 32 in the second.
- We employed MSE as our loss function.
- This version was primarily used for establishing a baseline and exploring our data so we did not normalise or transform our target, nor did we consider validation data.

4.3 Version 2: Enhanced Model With Data Splitting

- We introduced log transformation on `average_lap_time` - our target. and introduced a dropout rate of 0.2.
- We also split data into 70% training, 15% validation, 15% testing.
- Finally, we added early stopping after 10 epochs without significant change.
- This model saw significant improvement and contains much of the framework of our final model.

4.4 Version 3: Primary Optimized Model

- We adjusted our target normalisation to Standard Scalar and reversed it post-prediction.
- We further introduced SmoothL1Loss for better handling of outliers due to DNFs or car issues/damage.
- We further reduced batch size to 15 and added a learning rate scheduler to adjust the training dynamics.
- We finally tweaked some more of the hyperparameters introduced in the previous model, such as adjusting the dropout rate to 0.1) early stopping to 15 epochs.

5 Model Training

The models were trained using PyTorch for up to 100 epochs, with early stopping based on validation loss performance. Changes between versions can be seen in Figures 7 and 8.

Version 2 Highlights:

- Log scaling helped massively here in normalising the target.
- We say that our loss plateaued around epoch 70 and we decided that further adjustments were necessary.

Version 3 Highlights:

- After tweaking hyperparameters, we achieved a much smoother loss curve and lowered final validation loss significantly.
- Our scheduler poses a challenge to implement but aided greatly in model generalization and training stability.

6 Results

6.1 Error Metrics

The model performance is evaluated using the following metrics:

- **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and actual values.
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
- **Root Mean Squared Error (RMSE)**: The square root of MSE, providing error in the same units as the target.
$$\text{RMSE} = \sqrt{\text{MSE}}$$
- **Mean Absolute Error (MAE)**: The average absolute difference between predicted and actual values.
$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Here, y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of data points.

6.2 Baseline Model (v1)

- Test MSE: 304.3537
- Test RMSE: 17.4457
- Test MAE: 12.1459
- Accuracy within ±1s: 4.46%
- Accuracy within ±2s: 13.39%
- Accuracy within ±5s: 31.25%
- Accuracy within ±10s: 56.25%
- No transformation of the target resulted in lower accuracy.
- This is expected as we still needed to learn a lot about the nature of our problem and consider the high volatility of the sample data.

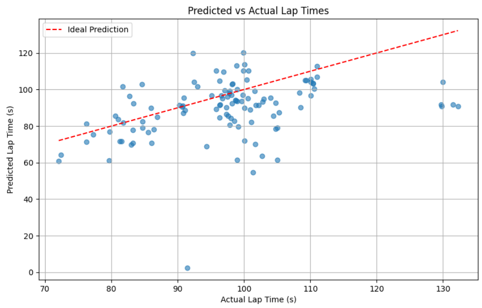


Figure 1: Predicted vs Actual Lap Times

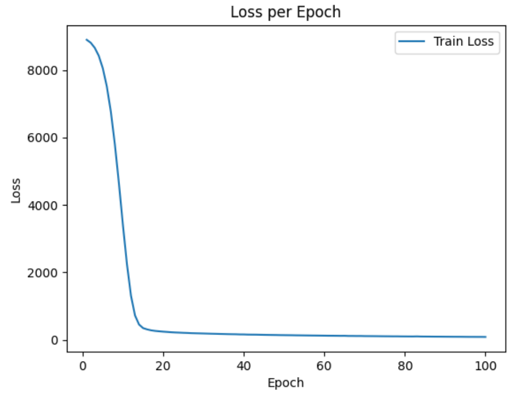


Figure 2: Loss per Epoch

6.3 Enhanced Model (v2)

- Test MSE: 46.3501
- Test RMSE: 6.8081
- Test MAE: 4.9420
- Accuracy within ±1s: 10.34%
- Accuracy within ±2s: 27.90%
- Accuracy within ±5s: 63.64%
- Accuracy within ±10s: 87.15%

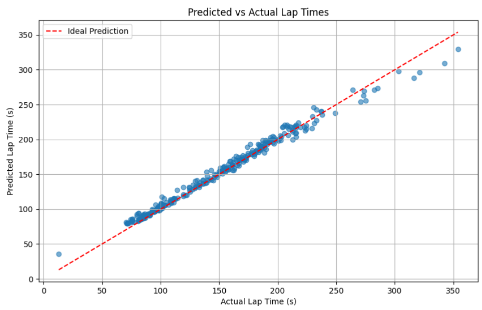


Figure 3: Predicted vs Actual Lap Times

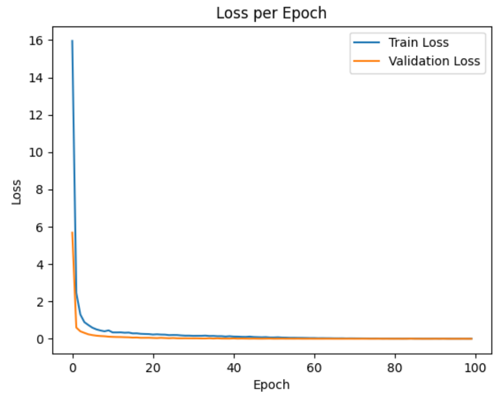


Figure 4: Loss per Epoch

6.4 Optimized Model (v3)

- Test MSE: 20.1799
- Test RMSE: 4.4922
- Test MAE: 2.5655
- Accuracy within ±1s: 28.21%

- Accuracy within ±2s: 54.86%
- Accuracy within ±5s: 90.60%
- Accuracy within ±10s: 98.12%
- Clearly outperforms earlier models due to superior preprocessing and training strategy.

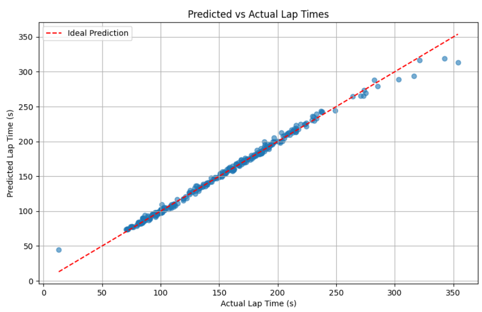


Figure 5: Predicted vs Actual Lap Times (Optimized Model v3)

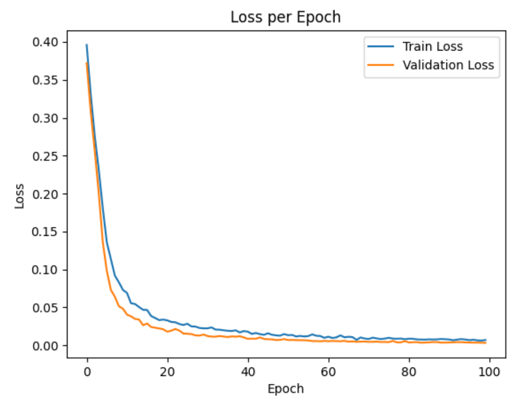


Figure 6: Loss per Epoch (Optimized Model v3)

A comparison of these models in terms of performance can be seen in the bar-chart and in radar-form with adjusted metrics below:

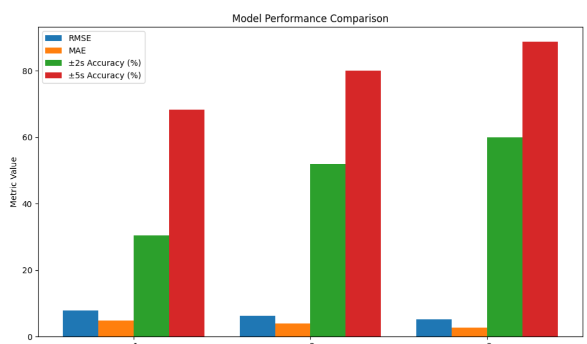


Figure 7: Bar Chart comparing the three versions of the model using key metrics

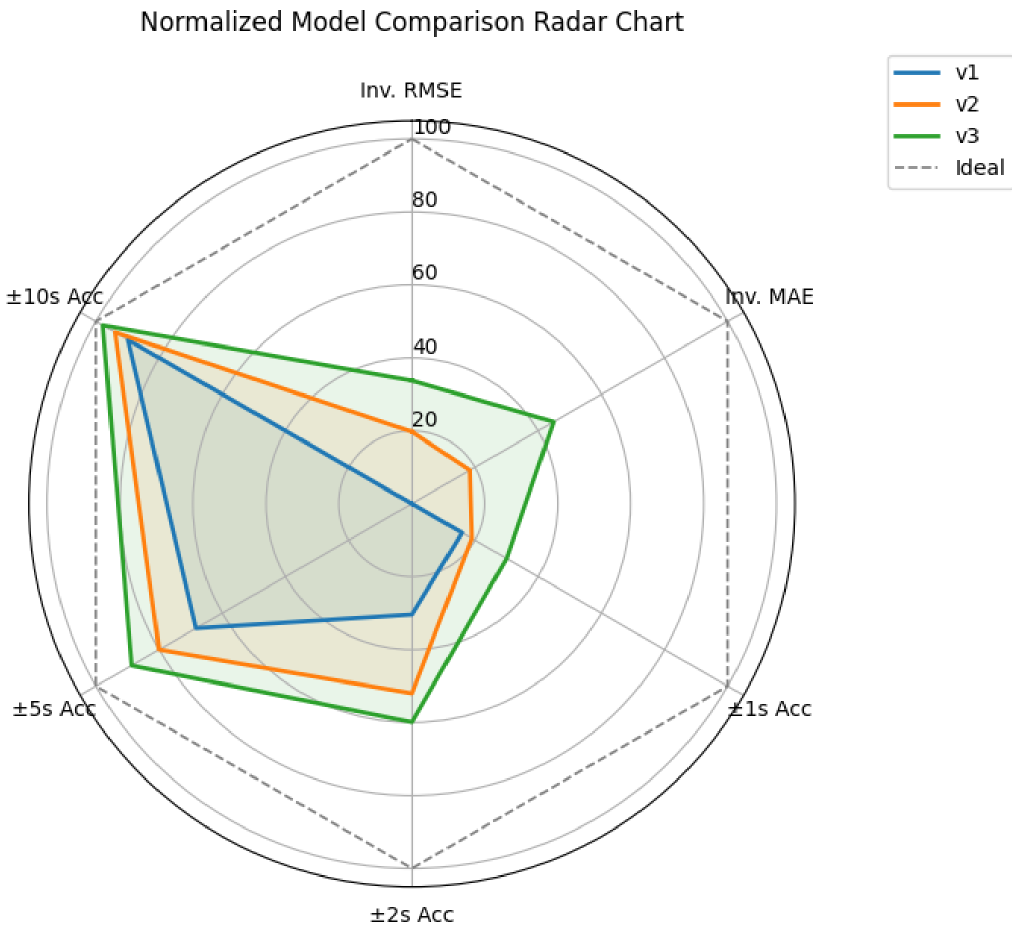


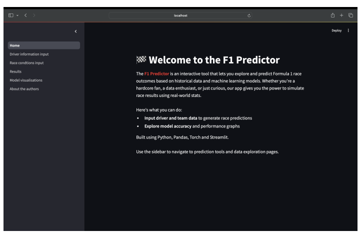
Figure 8: Radar Chart comparing the three versions of the model using key metrics

7 F1 predictor mini web application

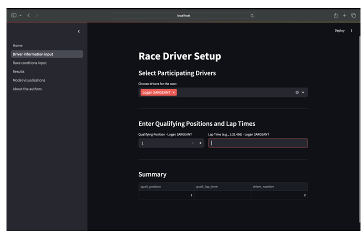
After improving our model, we consolidated the work into a mini web application that allows user input and predicts race results based on the provided data. The application works as follows:

- The user is prompted to enter the drivers participating in the race, along with each driver's qualifying average lap time and qualifying position.
- The user then provides information about the race conditions, such as weather and track length.
- Based on this input, the model generates predictions for each driver, ranks them by predicted average lap time, and tabulates the results accordingly.
- Additionally, the user can view a graph showing the training and validation error for that model run.

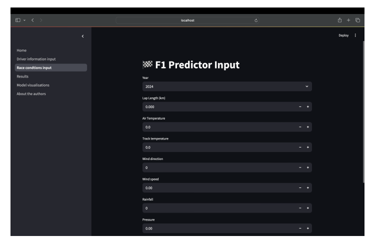
If no input is provided, the application uses default values to demonstrate what the results would look like under the default initial conditions.



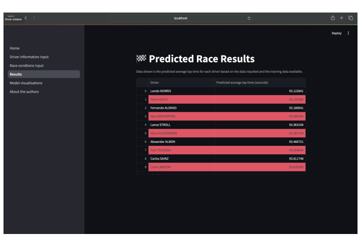
(a) Home Page



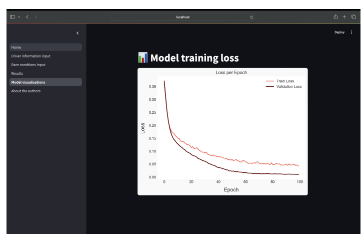
(b) Driver input



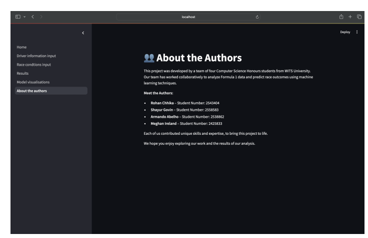
(c) Race condition input



(d) Results of the model



(e) Graph of error



(f) Author page

Figure 9: F1 predictor application

8 Future Work

Through out this project many strategies for how best to predict average lap time were attempted. These strategies involved not only the model used, but also the data and features thereof. There are still improvements which could be made to improve the prediction accuracy.

8.1 Adding more features to the training data

One improvement we could make would involve adding more features to the data set such as the drivers' standings and teams. We also found that the data retrieved from the API contained inconsistencies when it came to drivers who did not finish a race and issues where the data did not take into account time penalties. Although we attempted to remove data for a driver that did not set a time, this could be improved further by adding DNFs (Did Not Finish) and time penalties as features of the data set.

8.2 Exploring Other Models and Libraries

All models developed and evaluated in this project were implemented using the PyTorch library. In the future, we could explore and compare alternative machine learning frameworks such as TensorFlow, Scikit-learn, or XGBoost. Additionally, experimenting with different model architectures, such as Random Forests, or even combining multiple methods, may improve predictions.

8.3 Changing the target of the training data

Our data set had a target of average lap time, which was not always as accurate as we would have liked. One avenue to explore in future would be to try setting the target as position instead of average lap time.

9 Reflection and Conclusion

We see that the task was as challenging as it seemed at the start. By developing multiple iterations of a neural network that the predict average lap times for Formula 1 drivers, we see that our model is accurate within a second for nearly a third of all drivers and within two seconds for nearly two thirds. Considering regulation changes, variation in between tracks, weather conditions altering race-day pace, and adaptive pit strategies, high accuracy in a 2-second window for almost two thirds of the grid is a strong result.

Robust pre-processing and model optimization techniques allowed us to achieve an MAE of 2.7 seconds, showing a strong start for a small-scale project. Our future work includes transforming the problem to a classification task, aiming to predict race positions using categorical loss functions, such as cross entropy loss, and expanding the practical applications of our models.

10 Appendix

```
Qualifying: China 2025
Meeting Key: 1255 | Session Key: 9994
Final positions for each driver:
1. Oscar PIASTRI (#81) 183.56089473684207
2. George RUSSELL (#63) 165.28523809523807
3. Lando NORRIS (#4) 211.4254375
4. Max VERSTAPPEN (#1) 235.87550000000002
5. Lewis HAMILTON (#44) 176.81115789473682
6. Charles LECLERC (#16) 179.08478947368417
7. Isack HADJAR (#6) 189.97838888888887
8. Kimi ANTONELLI (#12) 161.47228571428573
9. Yuki TSUNODA (#22) 194.98627777777781
10. Alexander ALBON (#23) 180.06310526315787
11. Esteban OCON (#31) 165.79942857142854
12. Nico HULKENBERG (#27) 166.19628571428572
13. Fernando ALONSO (#14) 172.70299999999997
14. Lance STROLL (#18) 173.05214285714285
15. Carlos SAINZ (#55) 172.0292857142857
16. Pierre GASLY (#10) 137.575375
17. Oliver BEARMAN (#87) 150.77771428571432
18. Jack DOOHAN (#7) 157.29099999999997
19. Gabriel BORTOLETO (#5) 138.070875
20. Liam LAWSON (#30) 154.18985714285716
track_temperature: 36.1
pressure: 1014.9
rainfall: 0
wind_speed: 2.6
wind_direction: 252
humidity: 14.0
air_temperature: 26.0
```

Figure 10: A sample of the raw data structure extracted from OpenF1, showing a Qualifying session with relevant data.

```
Practice 1,2025,Australia,14,Jack DOOHAN,7,196.93789473684208,37.7,1016.8,0,1.3,109,67.0,22.9
Practice 1,2025,Australia,15,Pierre GASLY,10,172.86436363636363,37.7,1016.8,0,1.3,109,67.0,22.9
Practice 1,2025,Australia,16,Kimi ANTONELLI,12,155.2949583333333,37.7,1016.8,0,1.3,109,67.0,22.9
Practice 1,2025,Australia,17,Liam LAWSON,30,175.40266666666668,37.7,1016.8,0,1.3,109,67.0,22.9
Practice 1,2025,Australia,18,Gabriel BORTOLETO,5,181.97266666666667,37.7,1016.8,0,1.3,109,67.0,22.9
```

Figure 11: A sample of the cleaned dataset showing driver performance metrics and session-specific weather conditions, ready for model input.

References

[1] Formula 1. Las vegas grand prix 2024 - f1 race, 2024. Accessed: 2025-05-13. URL: https://www.formula1.com/en/racing/2024/las-vegas/circuit.

[2] Formula 1. Qatar grand prix 2024 - f1 race, 2024. Accessed: 2025-05-13. URL: https://www.formula1.com/en/racing/2024/qatar/circuit.

[3] Formula 1. United states grand prix 2024 - f1 race, 2024. Accessed: 2025-05-13. URL: https://www.formula1.com/en/racing/2024/united-states/circuit.

[4] Formula 1. Autodromo nazionale monza, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/italy-autodromo-nazionale-monza.FiJN1jnQlRLeHqOxIt13m.

[5] Formula 1. Autódromo hermanos rodríguez, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/mexico-autodromo-hermanos-rodriguez-mexico-city.1K2WPfBcI8kTXjcTHcbsBM.

[6] Formula 1. Autódromo josé carlos pace, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/brazil-autodromo-jose-carlos-pace-sao-paulo.5z2RfrmiTTfEP6Wnxv1yIW.

[7] Formula 1. Bahrain international circuit, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/bahrain-bahrain-international-circuit-sakhir.2CaIdaOTCgQ3Yfnb37NmSS.

[8] Formula 1. Baku city circuit, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/azerbaijan-baku-city-circuit-baku.5KEzWNQG1x2nLSpfg5xBZh.

[9] Formula 1. Circuit de barcelona-catalunya, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/spain-circuit-de-barcelona-catalunya-barcelona.6F5mWJGRuYkQ1XPX48pl8.

[10] Formula 1. Circuit de gilles villeneuve, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/canada-circuit-de-gilles-villeneuve-montreal.3ootdKogPeMEGqcNpleNrN.

[11] Formula 1. Circuit de monaco, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/monaco-circuit-de-monaco-monte-carlo.2ZWRtIcSI6ZzVGX1uGRpkJ.

[12] Formula 1. Circuit de spa-francorchamps, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/belgium-circuit-de-spa-francorchamps.3LltuYaAXVRU8iezEsjzGw.

[13] Formula 1. Circuit zandvoort, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/netherlands-zandvoort-circuit-zandvoort.6XdtPTIMZzx5wLKP9mm7Ev.

[14] Formula 1. Emilia romagna grand prix 2025 - f1 race, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/racing/2025/emiliaromagna/circuit.

[15] Formula 1. Hungaroring, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/hungary-hungaroring-budapest.3SzOBTKiIf7ov9bnezZ5dU.

[16] Formula 1. Marina bay street circuit, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/singapore-marina-bay-street-circuit.7LXNQUCHTyR5yMQPlIk7Lv.

[17] Formula 1. Melbourne grand prix circuit, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/australia-melbourne-grand-prix-circuit.4LZLnnP4nXDFVVD01EdvvG.

[18] Formula 1. Miami grand prix 2025 - f1 race, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/racing/2025/miami/circuit.

[19] Formula 1. Red bull ring, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/austria-red-bull-ring-spielberg.1s7EC2bGta7o1thZkVPbbq.

[20] Formula 1. Saudi arabian grand prix 2025 - f1 race, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/racing/2025/saudi-arabia/circuit.

[21] Formula 1. Shanghai international circuit, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/china-shanghai-interational-circuit-shanghai.7ESK3ZQ5DHqihx3DpPZKWL.

[22] Formula 1. Silverstone circuit, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/great-britain-silverstone-circuit.2DtFVI1FjkYgLVdGhbAIv0.

[23] Formula 1. Suzuka international racing course, 2025. Accessed: 2025-05-13. URL: https://www.formula1.com/en/information/japan-suzuka-international-racing-course.2XjOiKgIHRRBVVpp5N3S5t.

[24] PyTorch Contributors. Pytorch documentation. URL: https://docs.pytorch.org/docs/stable/index.html.

[25] Snowflake Inc. Streamlit • a faster way to build and share data apps. URL: https://streamlit.io/.