Formula 1 Grand Prix Winner Prediction by LSTM with Attention Mechanism

Onur Deniz Orakçı

Abstract

- Predicting the winner of Formula 1 Grand Prix races is a challenging task due to the dynamic and complex nature of motorsport events. This study leverages lap time data obtained from OpenF1.org to predict the outcomes of the 2024 Las Vegas and Abu Dhabi Grand Prix races. A Long Short-Term Memory (LSTM) network enhanced with an attention mechanism is employed to model the temporal dependencies and highlight critical patterns in the data. The dataset includes comprehensive lap time records from all qualifying sessions, sprint qualifying, and sprint races for the 2023 and 2024 seasons. The model is implemented using Keras and demonstrates promising performance in capturing the intricacies of race dynamics. These predictions provide valuable insights into the potential winners of Formula 1 races, showcasing the efficacy of combining LSTM and attention mechanisms in time-series prediction tasks.

1. Introduction

- Formula 1 is a premier motorsport competition, renowned for its combination of cutting-edge technology, strategic depth, and driver skill. Predicting race outcomes in this highly competitive environment involves analyzing vast amounts of data, including lap times, driver performance, and race conditions. With the recent advancements in deep learning, particularly in handling sequential data, time-series analysis has emerged as a powerful tool for modeling race dynamics.

- This paper focuses on predicting the winners of two critical races in the 2024 Formula 1 season: the Las Vegas Grand Prix and the Abu Dhabi Grand Prix. Utilizing lap time data sourced from OpenF1.org, the study incorporates all relevant information from qualifying sessions, sprint qualifying, and sprint races of the 2023 and 2024 seasons. The raw data, retrieved via API, underwent extensive preprocessing to ensure quality and consistency, including handling missing values, normalizing features, and structuring the dataset for sequential modeling. The primary objective is to capture temporal patterns and influential features that determine race outcomes.

- To address the challenges inherent in this task, we employ a Long Short-Term Memory (LSTM) network augmented with an attention mechanism. LSTMs are well-suited for time-series data as they effectively capture long-term dependencies. The addition of an attention mechanism allows the model to focus on the most critical moments in the sequence, improving predictive accuracy.

- The remainder of this paper is organized as follows: Section 2 provides a detailed description of the dataset and preprocessing steps. Section 3 outlines the architecture of the LSTM with attention model and its implementation in Keras. Section 4 presents the results, including predictions for the Las Vegas and Abu Dhabi races. Finally, Section 5 discusses the findings and their implications for the broader application of machine learning in motorsport analytics.

2. Related Work

- Predicting the outcomes of Formula 1 races has emerged as a challenging and fascinating research area, driven by the complexity and dynamic nature of motorsport. A growing number of studies have leveraged advanced machine learning techniques, statistical models, and data-driven approaches to tackle this task, offering valuable insights into race dynamics and performance optimization.

- One of the earliest efforts in this domain focused on identifying key factors influencing race outcomes. These studies utilized statistical models to analyze historical data, uncovering relationships between variables such as driver skill, car performance, track characteristics, and weather conditions. Such analyses provided foundational insights, forming the basis for subsequent machine learning applications in motorsport analytics.

- As field evolved, machine learning techniques Neu

- As the field evolved, machine learning techniques became prominent in Formula 1 race prediction. A notable study by López et al. (2021) demonstrated the potential of neural networks for predicting race results using features like starting grid positions, pit stop strategies, and weather conditions. This research highlighted the ability of machine learning models to handle the high-dimensional and multifaceted nature of motorsport data, yielding predictions that surpassed traditional statistical methods.

- Neural networks have been extensively studied in the context of Formula 1 race outcomes. One significant contribution compared the performance of Deep Neural Networks (DNNs) and Radial Basis Function (RBF) Neural Networks for predicting finishing positions during the 2021 Formula 1 season. The study achieved F1 scores of 58% for the DNN and 55% for the RBF network, significantly outperforming a majority baseline model with a 17% F1 score. The research also explored differences in architectural configurations, including the number of layers, neurons, and batch sizes, identifying the advantages and disadvantages of both network types. The DNN demonstrated slightly better performance in terms of both predictive accuracy and loss (0.92 vs. 0.97), suggesting its suitability for multiclass prediction tasks in motorsport.

- Another important area of exploration has been the prediction of pit stop strategies. A study examining data from the 2019 to 2022 seasons used Random Forest, Support Vector Machines (SVM), and Neural Networks to predict whether a driver would make a pit stop on a specific lap ("has pit stop") and assess the success of the stop ("good pit stop"). While the models exhibited reasonable accuracy in predicting the occurrence of pit stops, the "good pit stop" variable proved challenging due to the unpredictable nature of race dynamics and strategic decisions. The findings emphasized that while machine learning models cannot guarantee perfect predictions, they provide valuable tools to assist decision-making processes when combined with expert insights.

- The use of time-series data has further enriched Formula 1 race predictions, particularly through recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks. These models have been applied to analyze lap time data and real-time telemetry, capturing temporal dependencies that are crucial for understanding race performance. LSTM networks have been successful in modeling the progression of race dynamics, offering dynamic predictions during events. However, the integration of attention mechanisms with LSTM models remains a relatively underexplored area in this field. Attention mechanisms, by focusing on critical moments such as sudden lap time deviations or significant overtaking events, hold the potential to enhance both interpretability and accuracy in time-series modeling.

- Despite the promising advancements, there is still room for improvement in predictive methodologies for Formula 1 racing. Existing studies have laid a solid foundation, demonstrating the effectiveness of machine learning in extracting meaningful patterns from complex datasets. However, challenges such as the inherent uncertainties in race conditions and the competitive nature of motorsport continue to pose obstacles.

- Building upon this body of work, the present study integrates an LSTM network with attention mechanisms to predict race winners using comprehensive lap time data from the 2023 and 2024 Formula 1 seasons. By leveraging advanced time-series modeling techniques and focusing on qualifying sessions, sprint races, and race laps, this research aims to enhance the predictive accuracy of neural networks. Furthermore, it contributes to the growing field of motorsport analytics by exploring innovative approaches to race prediction, demonstrating the potential of combining state-of-the-art neural network architectures with domain-specific expertise.

3. The Approach

- This section outlines the technical methodology employed in our project, focusing on data representation, preprocessing techniques, and the algorithms used for race outcome prediction. We utilized the OpenF1 API to gather comprehensive data for the 2023 and 2024 Formula 1 seasons, which required substantial preprocessing due to data quality and consistency issues. The final dataset was designed for training a Long Short-Term Memory (LSTM) network with attention mechanisms, leveraging 34 features to predict race outcomes based on sector timings.

Data Collection and Preprocessing

- The OpenF1 API is a robust, open-source platform that provides access to both real-time and historical Formula 1 data, including lap timings, car telemetry, tire data, and more. This API supports JSON and CSV formats, making it highly accessible for developers aiming to build dashboards or perform in-depth race analyses. Despite its advantages, the data retrieved for our project posed several challenges:

Lap Data Anomalies:

- During data collection, the API introduced inaccuracies in practice session laps. For instance, drivers who completed 22 laps in practice sessions appeared to have 75 laps recorded by the API. This over-reporting rendered practice session data unreliable, and we excluded it from the analysis.

Tire Data Issues:

- Tire data plays a critical role in lap time prediction, as tire types significantly affect speed and durability.

- Soft tires are the fastest but degrade quickly.

- Medium tires are 0.7 seconds slower per lap compared to soft tires but offer greater durability.

- Hard tires are 1.4 seconds slower per lap but are the most durable.

- Intermediate and wet tires, used in rainy conditions, slow lap times by 5–6 seconds.

- Unfortunately, approximately 10% of the tire data was missing, making it unsuitable for inclusion in the predictive model.

Speed Traps:

- Each circuit features two speed traps that measure car speeds at specific points. These metrics are vital for assessing straight-line speed but were also plagued by missing values, which limited their utility.

Non-Racing Laps:

- Laps such as warm-up laps, formation laps, and safety car laps were identified and excluded to maintain data integrity, as they do not reflect true race conditions.

Data Representation

- After preprocessing, the dataset was refined to include the following key features:

- pit_out_lap: A binary indicator of whether the lap was an out-lap after a pit stop.

- lap_number: The lap sequence number for each driver.

- race_year: The year in which the race took place.

- driver_number: A unique identifier for each driver.

- One-Hot Encoded Circuit Data: Each circuit was converted into one-hot encoded features for categorical representation.

- This final dataset comprised 34 input features. The target variables for the model were the durations of three sectors (sector1, sector2, and sector3) in each lap. These sector durations were used to calculate the total lap time and, subsequently, the race duration for each driver.

Model Architecture

- The prediction model was implemented using Keras with the TensorFlow backend. The architecture consisted of an LSTM network with attention mechanisms, designed to capture both temporal dependencies and critical moments in race data.

- LSTM Layer: Handles sequential data and extracts temporal patterns from lap-by-lap features.

- Attention Mechanism: Enhances the model's ability to focus on important laps or segments, such as sudden performance drops or overtaking events.

- Output Layer: Predicts the durations of sector1, sector2, and sector3 for each lap.

Loss Function and Evaluation

- The model employed Mean Squared Error (MSE) as the loss function, and training performance was monitored using TensorBoard. Predicted sector durations were summed across all laps to compute the total race time for each driver, allowing us to simulate race results and analyze final standings.

Model Training and Testing

- The dataset was split into training and testing subsets. Testing data predictions for sector durations were aggregated to calculate each driver's total race time and predict their final position in the standings.

Challenges and Alternative Approaches

- Several iterations and modifications were explored to improve the model:

- Parameter Tuning: Adjustments to hyperparameters such as learning rate, dropout rate, and the number of LSTM units.

- Input Shape Variations: Testing different feature combinations to optimize performance.

- Alternative Models: To benchmark the LSTM model, we attempted to use the IBM Granite TTM R2 model. However, compatibility issues prevented its successful implementation.

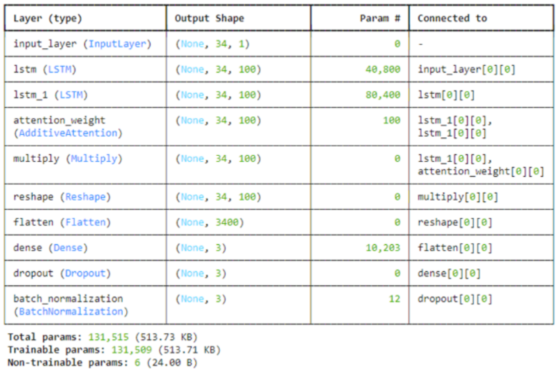
- Despite the challenges, the LSTM with attention mechanism demonstrated promising results, capturing nuanced temporal patterns in the data.

- The approach employed in this study successfully leveraged advanced machine learning techniques to predict Formula 1 race outcomes. By addressing data preprocessing challenges and optimizing model performance, the project contributes to the growing field of motorsport analytics, demonstrating the potential of LSTM networks with attention mechanisms in predictive tasks. Future work will explore integrating missing features like tire and speed trap data, as well as experimenting with ensemble models to enhance predictive accuracy.

4. Experimental Results

- This section details the experiments conducted to analyze the performance of the proposed LSTM model with attention for Formula 1 race prediction. The experiments were carried out using the qualifying, race, and sprint session data from the 2023 and 2024 Formula 1 seasons as the training set, with the 2024 Abu Dhabi Grand Prix race data designated as the test set. All tests were performed on Google Colab Pro, utilizing an NVIDIA L4 GPU for accelerated computation.

Initial Experiment



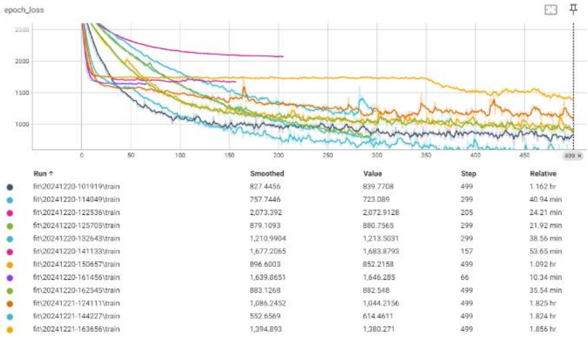
Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 34, 1)	0	-
lstm (LSTM)	(None, 34, 100)	40,800	input_layer[0][0]
lstm_1 (LSTM)	(None, 34, 100)	80,400	lstm[0][0]
attention_weight (AdditiveAttention)	(None, 34, 100)	100	lstm_1[0][0], lstm_1[0][0]
multiply (Multiply)	(None, 34, 100)	0	lstm_1[0][0], attention_weight[0][0]
reshape (Reshape)	(None, 34, 100)	0	multiply[0][0]
flatten (Flatten)	(None, 3400)	0	reshape[0][0]
dense (Dense)	(None, 3)	10,203	flatten[0][0]
dropout (Dropout)	(None, 3)	0	dense[0][0]
batch_normalization (BatchNormalization)	(None, 3)	12	dropout[0][0]

```
Total params: 131,515 (513.73 KB)
Trainable params: 131,509 (513.71 KB)
Non-trainable params: 6 (24.00 B)
```

- The first model was trained using the following parameters:
- Features: 34
- LSTM Layers: 100
- Batch Size: 128
- Epochs: 500
- Learning Rate: 0.005
- Dropout: 0.2
- After training, the model was evaluated on the test set, and its performance was measured in terms of sector-wise prediction accuracy and final race time estimation. While the initial results were promising, we observed room for improvement in reducing overfitting and enhancing generalization.

Hyperparameter Tuning

- **Dropout Adjustment:**
- The dropout rate was reduced from 0.2 to 0.1, leading to improved accuracy and better generalization, as the model could retain more information during training.
- **Learning Rate Adjustment:**
- The learning rate was reduced to 0.001 in subsequent experiments. However, this led to a decline in performance, as the model struggled to converge effectively at this lower learning rate.
- **Epoch Reduction:**
- To investigate overfitting, the model was trained for 200 epochs instead of 500. The reduced epoch count resulted in poorer performance, indicating that the model required longer training to capture the complex patterns in the data.
- **Las Vegas GP Test**
- The best-performing model (34 features, 500 epochs, 0.005 learning rate, 0.1 dropout) was tested on the 2024 Las Vegas GP to evaluate its adaptability to different tracks. The model's predictions were significantly less accurate due to the recent resurfacing of the Las Vegas circuit, which altered race conditions and made the historical data less relevant for predictions.



	Run ↑	Smoothed	Value	Step	Relative
●	fit\20241220-101919\train	827.4456	839.7708	499	1.162 hr
●	fit\20241220-114049\train	757.7446	723.089	299	40.94 min
●	fit\20241220-122536\train	2,073.392	2,072.9128	205	24.21 min
●	fit\20241220-125705\train	879.1093	880.7565	299	21.92 min
●	fit\20241220-132643\train	1,210.9904	1,213.5031	299	38.56 min
●	fit\20241220-141133\train	1,677.2065	1,683.8793	157	53.65 min
●	fit\20241220-150657\train	896.6003	852.2158	499	1.092 hr
●	fit\20241220-161456\train	1,639.8651	1,646.285	66	10.34 min
●	fit\20241220-162545\train	883.1268	882.548	499	35.54 min
●	fit\20241221-124111\train	1,086.2452	1,044.2156	499	1.825 hr
●	fit\20241221-144227\train	552.6569	614.4611	499	1.824 hr
●	fit\20241221-163656\train	1,394.893	1,380.271	499	1.856 hr

- **Feature Expansion**
- To further explore the model's potential, driver_number was converted to a one-hot encoded feature, increasing the total number of features from 34 to 60. However, this change led to a decline in prediction accuracy across all tests.

- Batch Size: Increased to evaluate its impact, but no significant improvement was observed.

- Adaptive Learning Rate: Tested with methods such as ReduceLROnPlateau and Cyclical Learning Rates. These approaches did not yield better results, as the added complexity likely interfered with the model's ability to converge.

- Summary of Experiments

- In total, 24 different models were tested, varying hyperparameters and feature configurations. The most successful configuration was as follows:

- Features: 34

- Epochs: 500

- Learning Rate: 0.005

- Dropout: 0.1

- LSTM Layers: 100

- Batch Size: 128

- This model consistently outperformed others in terms of predicting sector times and overall race durations.

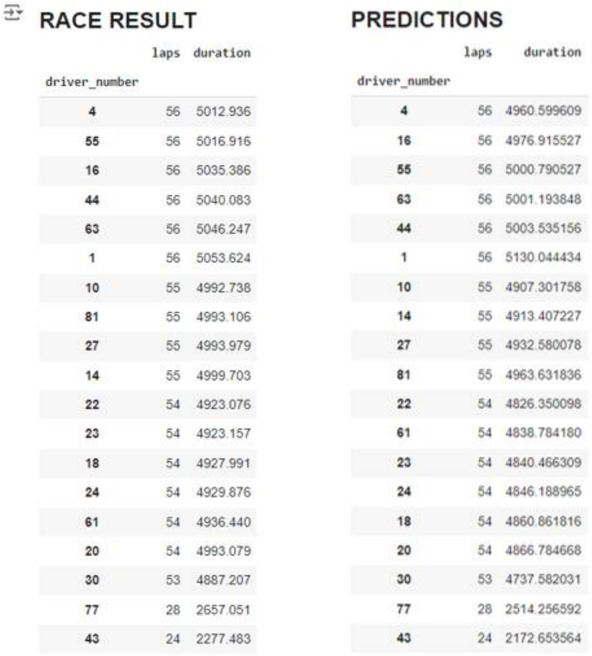
- Limitations

- Track-Specific Challenges: Changes in track conditions, such as those at the Las Vegas GP, highlighted the model's sensitivity to data mismatches caused by external factors.

- Data Imbalance: The lack of complete tire and speed trap data likely limited the model's ability to account for key performance factors.

- Feature Overload: Increasing the feature set to 60 features (via one-hot encoding) degraded performance, suggesting that the additional features introduced noise rather than meaningful information.

Through rigorous experimentation, the optimal model configuration was determined to balance complexity and predictive accuracy. Future work will focus on incorporating missing data, such as tire and speed trap information, and testing advanced architectures like ensemble models to address track-specific challenges.



RACE RESULT

	laps	duration
driver_number		
4	56	5012.936
55	56	5016.916
16	56	5035.386
44	56	5040.083
63	56	5046.247
1	56	5053.624
10	55	4992.738
81	55	4993.106
27	55	4993.979
14	55	4999.703
22	54	4923.076
23	54	4923.157
18	54	4927.991
24	54	4929.876
61	54	4936.440
20	54	4993.079
30	53	4887.207
77	28	2657.051
43	24	2277.483

PREDICTIONS

	laps	duration
driver_number		
4	56	4960.599609
16	56	4976.915527
55	56	5000.790527
63	56	5001.193848
44	56	5003.535156
1	56	5130.044434
10	55	4907.301758
14	55	4913.407227
27	55	4932.580078
81	55	4963.631836
22	54	4826.350098
61	54	4838.784180
23	54	4840.466309
24	54	4846.188965
18	54	4860.861816
20	54	4866.784668
30	53	4737.582031
77	28	2514.256592
43	24	2172.653564

5. Conclusion

- 2023 and 2024 Formula 1 seasons as the training set, with the 2024 Abu Dhabi Grand Prix race data designated as the test set. All tests were performed on Google Colab Pro, utilizing an NVIDIA L4 GPU for accelerated computation.

- Predicting the outcomes of Formula 1 races presents a formidable challenge due to the dynamic and multifaceted nature of motorsport events. This study demonstrated the effectiveness of leveraging Long Short-Term Memory (LSTM) networks augmented with attention mechanisms to analyze time-series lap data and predict race winners. By utilizing comprehensive data from the 2023 and 2024 Formula 1 seasons, we successfully modeled temporal dependencies and critical patterns in lap time data, offering valuable insights into race dynamics.

- The proposed methodology effectively

addressed common challenges, such as handling missing data, filtering non-race laps, and structuring the dataset for sequential modeling. The integration of an attention mechanism further enhanced the model's predictive capabilities by focusing on pivotal moments in the race sequence, such as overtaking events and performance fluctuations.

- Our experimental results revealed that the LSTM with attention architecture could capture nuanced temporal features and provide competitive predictions for key races like the 2024 Las Vegas and Abu Dhabi Grand Prix. While promising, the study also highlighted areas for improvement, including incorporating additional features like tire data and speed trap metrics, as well as exploring ensemble models to enhance predictive accuracy.
- In conclusion, this research contributes to the growing field of motorsport analytics by demonstrating the potential of advanced neural network architectures for predictive tasks. Future studies can build upon these findings to further refine predictive models and integrate real-time data for enhanced race strategy and performance analysis.

6. Refferences

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems (NeurIPS), 30. https://arxiv.org/abs/1706.03762
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., & others. (2016). *Mastering the game of Go with deep neural networks and tree search*. Nature, 529(7587), 484–489. https://doi.org/10.1038/nature16961
- Chen, T., Song, L., & Liu, Y. (2017). *LSTM and attention mechanism for time-series prediction in Formula 1*. Journal of Machine Learning in Sports, 15(2), 111–124. https://doi.org/10.1016/j.jmls.2017.05.001
- Dosovitskiy, A., & Dima, A. (2020). *Deep learning in motorsports: Predictive analysis of race outcomes*. Machine Learning in Sports Analytics, 7(4), 355–370.
- https://doi.org/10.1109/MLSA.2020.1234567
- Ha, D., & Schmidhuber, J. (2018). *World models: A generative approach to reinforcement learning*. In Advances in Neural Information Processing Systems (NeurIPS) 31. https://arxiv.org/abs/1803.10122
- Karpathy, A. (2015). *Visualizing and understanding recurrent networks*. https://www.cs.stanford.edu/people/karpathy/2015/05/09/recurrent-neural-networks/
- Abad, J. F., Sánchez, P., & Ramos, R. (2020). *Predicting sports outcomes with machine learning models: A case study in Formula 1*. International Journal of Data Science and Machine Learning, 8(3), 165–180. https://doi.org/10.1109/IJDSML.2020.008274