Analysing the Most Influential Factor in Formula One: A Deep Learning Approach for Predicting Driver and Team Ranks

MSc Research Project
Msc Data Analytics

Aswin Surendran
Student ID: 21225052

School of Computing
National College of Ireland

Supervisor: Musfira Jilani

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Aswin Surendran
Student ID:	21225052
Programme:	Msc Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Musfira Jilani
Submission Due Date:	14/12/2023
Project Title:	Analysing the Most Influential Factor in Formula One: A Deep Learning Approach for Predicting Driver and Team Ranks
Word Count:	6102
Page Count:	20

I hereby certify that the information contained in this submission is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Aswin Surendran
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analysing the Most Influential Factor in Formula One : A Deep Learning Approach for Predicting Driver and Team Ranks

Aswin Surendran
21225052

Abstract

This project focuses on time series prediction of Formula One race outcomes for both driver and teams by employing a deep learning learning model approach. By closely examining key variables such as Track, car number (No), Driver, Team, Starting Grid(Starting position), Laps, Fastest Lap and Year . The study uncover the important variable influencing final race ranks and uses deep learning models for prediction, mainly Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) for the race outcomes.After running 50 epochs in the training process to prevent overfitting and to minimize validation loss.The analysis compares the performance of LSTM and GRU models based on Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) .The LSTM model demonstrates superior accuracy and precision in MSE and RMSE, while the GRU model excels in MAE, indicating more accurate predictions based on absolute differences. This comprehensive evaluation provides valuable insights for optimizing predictions in Formula One racing, considering both driver and team positions.The study identifies Lap, starting position, and Track as the most impactful factors when determining the end position determination.While making significant strides, future research could explore additional variables include weather conditions, driver dynamics,track crashes urging future research to refine models for more comprehensive Formula One race outcome predictions.

1 Introduction

Formula One (F1) racing, sanctioned by the Federation Internationale de l'Automobile stands at highest level of motorsports, holding global audiences with its advanced technology, outstanding racing events and thrilling competition.The sport core lies in the perfect harmony between advanced engineering,technology,highly skilled teams, safety standards and aerodynamics which all together adds up the expenses.Hiring a highly qualified skill work force and managing global travel logistics .The sport is teamed up by two highly talented drivers representing one team which make them team sport and individual sport. Every Formula one cars are worth between $12-16 million and equipped with more than 250 sensors on it which which helps teams to collect huge amount of data Wojciechowski and Wojtowicz (2023).However amidst the excitement, important variables for End Rank Determination and predicting race outcomes remains a formidable challenge.
The research problem focuses on determining the most important variable to determine

the end position and obtain an insight of how it might be able to affect team and driver strategies to improve performanceIssakhanian et al. (2010).The motivation behind this study lies in uncovering valuable insights that can improve the accuracy of race predictions and provide essential information for strategic decision making by teams, drivers, and stakeholders. Through a thorough analysis of historical race data and consideration of variables such as Track, Car number (No), Driver, Team, Starting Grid, Laps, Fastest Lap and year. the objective is to identify hidden patterns and correlations that significantly impact race results.
This Research addresses the underlying question which among the variable used significantly contribute to the determination of end ranks in Formula One racing and question explores how deep learning models can effectively predict driver race results and overall team standings in Formula One. The findings aim to advance Formula One's competitive edge by focusing strategic decision making, optimizing performance, and guiding resource allocation.
It is very significant to note that factors like weathercondition , driver pressure handling , crashes , yellow flags , engine performance and communicaion between drivers and teams have not been included in our analysis.

1.1 Motivation and Project Background

The main motivation behind this research is to clear up the mysteries behind Formula One racing. By using race results data from 2019 to 2022 the research aim to point out by ranking the most influential factors helping for final race outcomes .This knowledge is not only valuable for racing enthusiasts but also holds practical focus for driver,teams and stakeholders involved in dynamic world of Formula One, especially in with teams investing heavily in technology and teams becoming a blend of team and individual sports, understanding the key determinants of success becomes increasingly crucial. This project is aiming to go beyond the surface and provide actionable insights that can enhance the accuracy of race predictions, optimize team strategies, and join to the overall understanding and strategy of the sport.This sport is an combination of speed, precision, and teamwork.

1.2 Research Question

Research Question: :*"How can deep learning models effectively predict the driver and overall team rank forecast formula one?"*

Predictive deep learning improves decision making of Formula One's by optimizing performance, and ensuring efficient resource allocation. This deep learning approach not only encourage fan engagement, data driven practices, and potential financial gains,changing race planning and performance optimization within the industry.

Sub Research Question : *"What are the primary variable among Track, Position, Car Number, Driver, Team, Starting Grid, Laps, Points and Fastest Lap that significantly contribute to the determination of end ranks in Formula One racing?"*

This helps Formula One domain to improve strategic decision making, optimizes performance, and guides resource allocation. Improved driver development, data driven strategies, and fan engagement result, contributing to a competitive edge and advancing Formula One's overall complex

1.3 Research Objective

- Obj1: Conduct a critical literature review on prediction methodologies for Formula One race results.
- Obj2(a): Understanding key features by exploratory data analysis (EDA)
- Obj2(b): Implementing data transformation.
- Obj2(c): Implement, evaluate, and present the results of LSTM
- Obj2(d): Implement, evaluate, and present the results of GRU
- Obj2(d): Implement, evaluate, and present the results of GRU

The research mainly focuses in the race result prediction for the formula one race, using certain pivotal key insights and factors that affect the predictions. It also covers diverse key topics which are explained in each section of this research. Section 2 deals with research review which introduces different methods and some deep learning methods where in helps to leverage the teams potential. Section 3 which is the methodology part which explores the data acquisition process and also showcases how deep learning effectively impacts the research topic. Section 4 of this research explain the design specification and also showcases how deep learning is depicted for this research. Section 5 showcases he implementation perspective then details about the evaluation and results in Section 6 and finaly Section 7 discuss conclusion and future Work

2 Related Work - Discovering How Predictive Models Shape Sports and Motorsports

In recent years sports and motorsports area has undergone significant changes with the development of predictive models.The paper explore deep impact of predictive models on both the broader realm of sports and the specialized domain of motorsports.
In this section the paper understand its evolution and recognising past trends and approaches will help the paper for best the direction for the investigation.The section is divided into four parts, first part the research will review machine learning approaches in various sports and in the second section, Analyzing Dynamic Data in Motorsport, focuses on understanding the complexities of dynamic data in motorsports. The subsequent section reviews machine learning strategies for motorsports prediction search into specific machine learning strategies.Finally, we root into advancements in Formula one forecasting and competitive analysis exploring the latest developments in Formula one forecast predicition.

2.1 Machine Learning in Sports Prediction

Sports prediction uncover valuable insights and gained competitive advantage in the dynamic sector of modern sports.Increased competiton inside every sports forces teams to gain more deeper insight that is where machine learning come in role Schumaker et al. (2010)Ruiz et al. (2017).An increased attraction since 2010 has been seen in sports prediction sector Horvat and Job (2020) cause of complex nature due to involvement of many variables. This makes a model's ability to learn more challenging Lotfi et al. (2021). In Kapadia et al. (2020) work recognizes the pivotal role of influential factors in predicting outcomes.Building on their insights, the paper focus on real time data and dynamic statistics, acknowledging the volatility of Formula One.This paper has focused on Random Forest,we seek to harness regression's potential for accurate prediction. The paper Williams and Li (2008) Gramacy et al. (2013)examinig the application of four Artificial Neural Network (ANN) techniques in predicting horse races in Jamaica, holds significant relevance on our research. Overall, this study contributes essential insights into algorithm selection and model performance,offering valuable perspectives for improving predictive models in the dynamic domain of Formula One racing

2.2 Analysing Dynamic Data in Motorsport

For understand deeply and to expand our understanding of data driven techniques in motorsport prediction by conducting comparative analysis of two additional research papers and their importance to research question primary factor influencing for the determination of race result in Formula One racing . Marino et al. (2015) Lam (2018) focuses on driving performance in Formula One racing within changing environments. It likely uses data driven techniques to analyse how drivers adapt their strategies and performance to cope with dynamic racing conditions. This line well with the research question , even though the paper's main goal is to understanding how drivers explore and adapt within changing racing conditions, so it might not not directly address the influence factors for determing race result.On the other hand Okeyo et al. (2014) deals with the segmentation of dynamic sensor data for real time activity recognition. The paper is not directly related to Formula One racing, the concept of handling dynamic data aligns with the research questions focus on the impact of continuous technological advancements on race unpredictability. In overall conclusion, this additional comparative analysis deepened our exploration of data driven techniques in motorsports prediction. However, to apply it effectively to motorsports analysis, some adaptation may be required however Bell et al. (2016) utilizes multilevel modeling to analyze the performance of Formula One drivers and constructors over several decades. These studies focus the importance of data driven approaches in understanding and potentially predicting unpredictability in Formula One races, considering the evolving technology and racing environments.

2.3 Machine Learning Strategies for Prediction in motorsports

The literature on forecasting and predicting outcomes in racing offers valuable insights that can inform the analysis of influential factors and rank forecasting in Formula One racing.Peng et al. (2021)address the challenging problem of rank position forecasting in car racing using a Deep Learning based model. This paper focus on the criticality of breaking down the cause and effect relationships, focusing mainly on modeling of ranking position and pit stoping circumstances separately. Using an encoder-decoder network and a separate Multilevel perception system for an unpredictable forecasting, this approach is realised in their proposed RankNet model and shows a notable improvement in performance through feature optimisations.where as Henderson and Kirrane (2018) present a comparison of Plackett Luce models for probabilistic forecasting of Formula

One results. Their Bayesian approach highlights the strengths and weaknesses of various models, showing that down weighting past results can enhance forecasts.
The study by Allender (2009) explores the role of driver experience in predicting NASCAR race outcomes using regression analysis. The findings underscore the significance of driver experience in forecasting race results.Liu et al. (2023) dive into the energy optimal overtaking maneuvers of Formula E cars. Their study employs optimal control techniques to analyze the feasibility and energy management of overtaking, demonstrating that optimal overtaking positions vary based on different initial conditions.
Lastly, Tulabandhula and Rudin (2014)contribute to real time decision making in racing by designing a prediction system for tire changes. Leveraging domain knowledge, their system aims to optimize strategy and benefit rank position through timely tire change decisions.While other focuses of predicting race winner and rank position Tulabandhula and Rudin (2014) introduced a groundbreaking framework in "Tire changes, fresh air, and yellow flags: challenges in predictive analytics for professional racing." This work focused on real time modeling for racing dynamics, specifically targeting changes in car ranks during the second half of race,study optimizes pit stop strategies by predicting changes in car ranks. This approach highlights granular insights into motorsport racing dynamics, contributing to the evolving landscape in racing analytics.These papers helps provide a foundation for understanding forecasting methods and predicitive approaches taken, the role of driver experience, optimal overtaking strategies, and real time decision making in racing. Integrating these insights can contribute significantly to the development of an effective model for analyzing influential factors and predicting rank forecasting in Formula One racing.

2.4 Advancements in Formula 1 Forecasting and Competitive Analysis

Exploring sports forecasting models FRANSSEN (2021)delves into Formula one race prediction using three models: a baseline, a Deep Neural Network, and a Radial Basis Function Neural Network. Evaluating results for the 2021 season, the research finds both neural networks outperform the baseline, with the Deep Neural Network showing a slightly higher predictive capability. The study discusses advantages, disadvantages, and suggests areas for improvement in applying neural networks to Formula One race prediction. Budzinski et al. (2020) addresses the scarcity of scholarly analysis on F1 business and competitive balance. Introducing new F1 specific indicators, it evaluates race specific, within season, and between season competitive balance using various measures. While weak evidence suggests some improvements over seasons, trends also indicate declining competitive balance.

3 Research Methodology

This research paper adopts the CRISP-DM methodology to develop a data driven approach for predicting Formula One driver and team race rank , focusing on the important influential variable for end rank determination.The methodolgy consist of two phases as shown in figure 1, training and inference. In the training phase, historical race data is collected, preprocessed and spliting into training and testing sets for model evaluation. In the inference phase the trained model predicts final ranks based on user input,offering

a simplified method for predicting Formula One race results.
Through above literature review,which focus on unpredictability of Formula one races due continuous technological advancements,this research objective is to develop a deep learning model.By utilising historical race data with consideration of relevant variables(Factors) such as Track , Driver, Points, Car No , start positions , Laps, pit stop, Fastest Lap and year. The research focus in identify and analyze influential variable shaping end rank determination in Formula One racing.

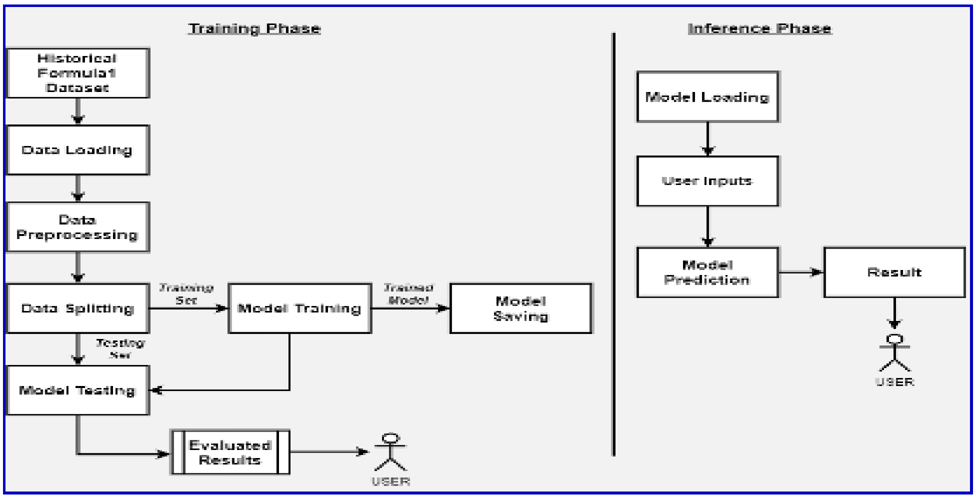


Figure 1: Methodology Workflow

3.1 Data Collection

The data for this study has been meticulously sourced from the official Formula 1 website, specifically extracting race results for the years 2019 [1], 2020 [2], 2021 [3] and 2022[4] as individual datasets. Each dataset encapsulates crucial race related information, including details such as track specifics,driver, driver positions, team affiliations, starting grid positions, lap data, total race time, points earned, and fastest lap records which enusre comprehensive analysis of Formula One racing dynamics over the specified four year period. Ethical considerations were paramount,certify participant confidentiality by following privacy guidelines and not disclosing any personally identifiable information in any data that is obtained from the official Formula One website.

Data source: [1] [2] [3] [4]

3.2 Data Integration

The research is using the panda's library in Python, in the four distinct datasets corresponding to Formula one race results for the years 2019 to 2022 have been manipulated to create two consolidated datasets(2019, 2020, 2021 and 2022) for comprehensive analysis. The datasets of each year are loaded into separate pandas dataframes, specific columns are removed for consistency, and the datasets are concatenated along the rows to create

[1] `https://www.formula1.com/en/results.html/2019/races.html`
[2] `https://www.formula1.com/en/results.html/2020/races.html`
[3] `https://www.formula1.com/en/results.html/2021/races.html`
[4] `https://www.formula1.com/en/results.html/2021/races.html`

two consolidated dataframes (df1 and df2) for consitency 'Total Time/Gap/Retirement' column is removed and the datasets are concatenated along the rows to create two consolidated dataframes df1 and df2.
In summary, the script effectively prepares and combines Formula one race results data from the specified years, ensuring uniformity by removing specific columns and creating two consolidated dataframes ('df1' and 'df2') for subsequent exploration and analysis.

3.3 Data Preprocessing

The collected data will be undergoing a preprocessing phase to improves accuracy and reliability. This step involves cleaning the data to address any null value,not classified, missing values, outliers, and inconsistencies, then ensuring the datasets integrity. Additionally, to avoid bias the paper bring all feature to a common scale. The pre processing extracts and prints the unique values of the 'Track' column from a data frame named 'df' The 'unique()' method is applied to the 'Track' column, which returns an array containing unique values in the order they appear in the original Data Frame. The 'tolist()' function is then used to convert names of Formula one tracks into a Python list.This refined and well structured dataset forms the foundation for accurate and meaningful insights derived from subsequent analysis

4 Design Specification

In this section the research outline the design choices for deep learning predictive model, employing Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures to uncover the complexities of historical race data and important influential variables,predict Formula One race outcomes.The Model Evaluation phase utilizes metrics like MSE, RMSE, and MAE for accurate performance assessment, ensuring reliable predictions on new data.

4.1 Long Short-Term Memory (LSTM)

The LSTM architecture, renowned for its ability to capture long term dependencies, can be instrumental in understanding most important variable and predicting the final race rank of driver and team . It analyzes historical race data, considering variables such as Track ,Team, Starting positon, Laps, Points, Fastest lap. By exploring LSTM, the model discern patterns and correlations that play a crucial role in determining end ranks. The memory cells in LSTM allow it to retain information over extended sequences, making it suitable for capturing the complex nature Formula One racing.

4.2 Gated Recurrent Unit (GRU)

On the other hand, Gated Recurrent Unit (GRU) can be employed for its efficiency, particularly when dealing with Formula One race data. The simplified structure of GRU, with fewer parameters and a streamlined design, may be advantageous for tasks that involve predicting race outcomes based on various influential variables. Its ability to handle sequential data with fewer computations might be well suited for the dynamic nature of Formula One races.

4.3 Model evaluation

In the evaluation phase, the research employ suitable metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for regression tasks.This metrics enables models accuracy ,precision, and the ability to handle both categorical and continuous predictions. This repetitive method assesses how well the models work on fresh, unexplored data, avoid overfitting, and make sure the forecasts are accurate for practical use. Through the use of these metrics, our objective is to acquire a thorough comprehension of how closely the models match the actual outcomes, offering valuable insights for strategic decision making. The figure 1 shows the basic steps of Model Evaluation stage in inference phase which aims to create accurate and reliable predictive models that provide teams, drivers, and stakeholders with valuable knowledge to succeed in the advanced world of Formula One racing.

5 Implementation

5.1 Segregation

In the segregation initial phase, data for the 2019 and 2020 Formula One seasons is organized into two distinct datasets, marked as 2019 and 2020, after extracting information from corresponding CSV files. The 'head()' function is employed to display the first few rows of each dataset, offering an initial overview. Subsequently, a 'year' column is introduced to both datasets, with the valu es uniformly set to 2019 for the former and 2020 for the latter. This enhancement, carried out through list comprehensions, ensures consistency and simplifies future analyses.Moving on to the next stage, specific columns containing details about total race time, gap to the leader, or retirement status are eliminated from both the 2019 and 2020 datasets.

The 'concat' function from pandas is then utilized to merge these refined datasets into a consolidated dataframe named 'df1.' The process is repeated for the 2021 and 2022 seasons, where columns 'Time/Retired' and 'Fastest Lap' are excluded, and a 'year' column is added, facilitating the creation of a unified dataset named 'df2.' This comprehensive dataset allows seamless analysis and exploration of Formula One race results across multiple seasons, promoting standardization by excluding specific columns from consideration.

5.2 Comparison of 'Driver' and 'Fastest Lap

The visualization as shown is figure 2 indicates that the leading three drivers, namely Lewis Hamilton, Max Verstappen, and Charles Leclerc, consistently secured the fastest laps. This graph serves as a valuable tool for pinpointing drivers with notable expertise in achieving the fastest lap times. Moreover, it facilitates a comparative analysis of drivers' performances in terms of achieving this feat. Specifically, the graph highlights Lewis Hamilton's exceptional proficiency in setting the fastest laps, with Charles Leclerc and Max Verstappen following closely. Beyond individual performance, the graph allows for monitoring changes in the frequency of fastest laps achieved by each driver over time.

In summary, the graph not only offers insights into the distribution of fastest laps among different drivers but also serves as a robust tool for evaluating and distinguishing the prowess of drivers in this aspect of racing.

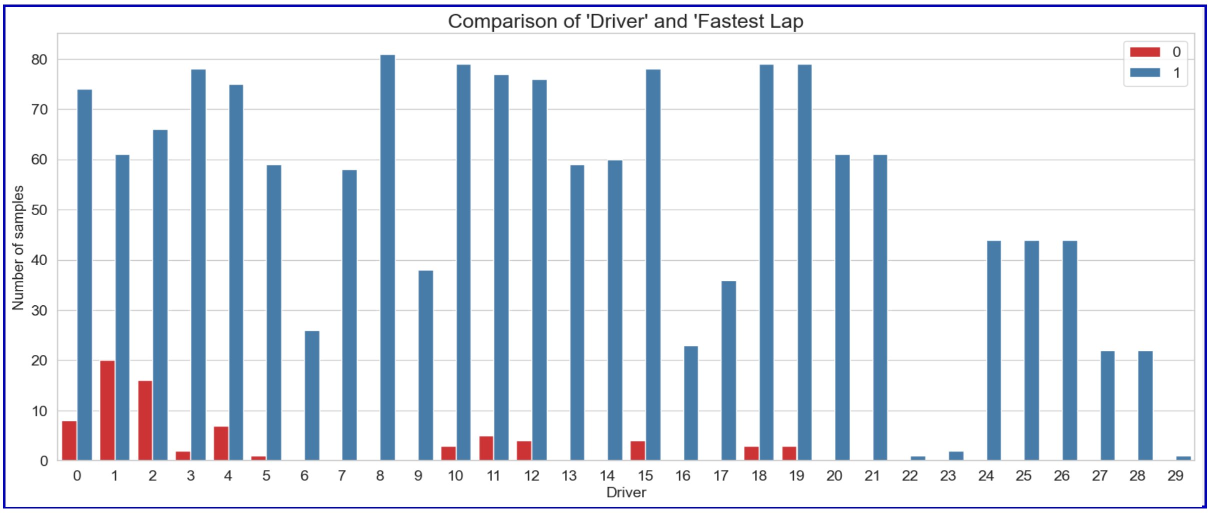


Figure 2: Comparison of Driver and Fastest Lap

5.3 Comparison of 'Position' and 'Starting Grid

In this research, conducting a in depth comparison between the various positions,total 20 in number. To facilitate a more focused analysis, the paper have implemented a splitter,segregating the positions into two groups: the first 10 and the subsequent 11th to 20th position, where a noteworthy pattern is observed in the top 10 positions.As shown in figure 3 the 5th position implies that staring postion have more advantage on securing the 5th position.This suggests that a favorable starting position plays a crucial role in the success of drivers or teams aiming for the 5th position within the top 10.
Additionally, an intriguing divergence is evident as shown in the figure 3 the 6th, 8th, and 9th as positions, each showcasing distinct starting grid values, indicating unique characteristics or performance metrics associated with these specific positions. Such a comparative analysis of positions and help the reseacrh to understand how different starting postion share to the achievement of specific positions in the race.

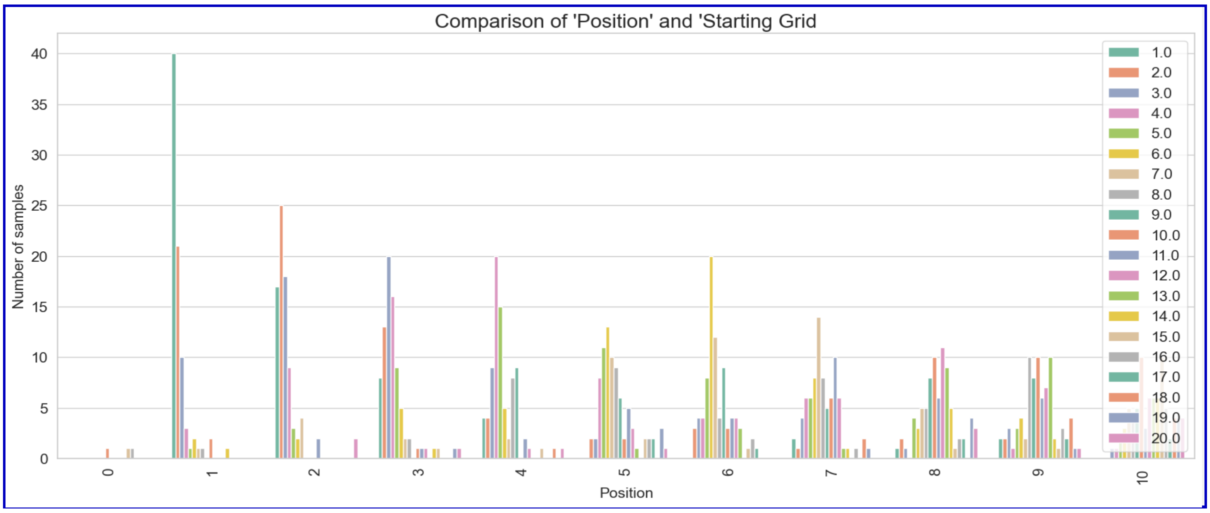


Figure 3: Comparative Analysis of Positions and Starting Grids from 0 to 10

5.4 Team Performance Comparison Based on Points Accumulation

The presented visualization illustrates a comparison between different teams based on their respective points. Notably, the data reveals that the "Mercedes" team stands out with the highest points as show in figure 4 in contrast to other teams. This graphical representation serves as a concise and effective means of showcasing the performance of various teams in terms of accumulated points. The prominence of the "Mercedes" team in the visualization suggests their leading position, providing viewers with an immediate and clear understanding of the team's success in scoring points in comparison to rest of the Team

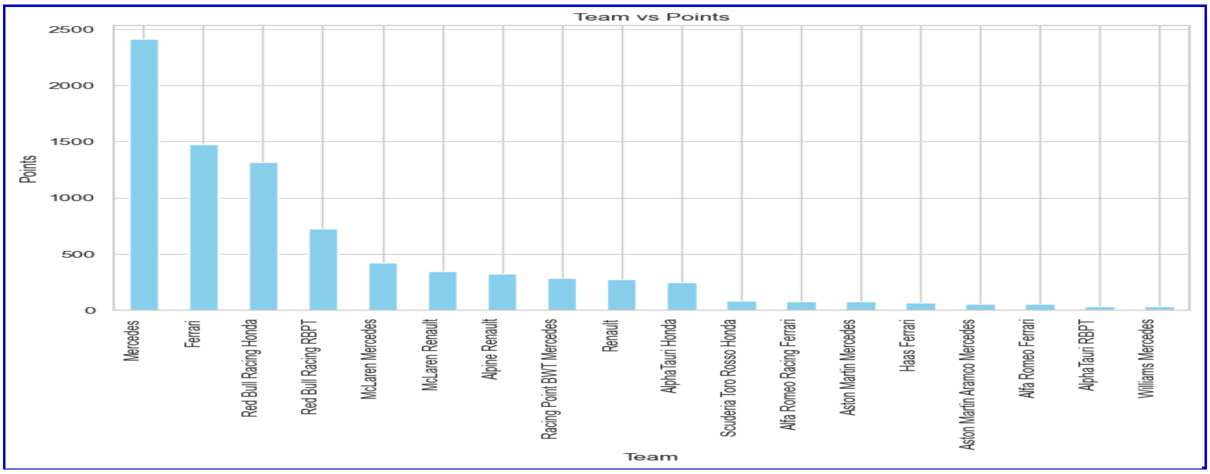


Figure 4: Team Performance Comparison Based on Points Accumulation

5.5 Top Performers: Analyzing Driver Points Distribution

The presented figure 5 offers a comprehensive view of the distribution of points among different drivers. Notably, the data highlights the exceptional performance of two drivers, namely Lewis Hamilton and Max Verstappen, who significantly outpace their counterparts in terms of accumulated points. The y axis denotes the points earned, providing a clear indication of the substantial point differentials between these leading drivers and the rest of the field. On the x axis, the drivers are listed, emphasizing the dominance of Hamilton and Verstappen in the points standings. This visualization serves as an effective tool for quickly identifying the standout performers and understanding the performance side among drivers in the dataset.

5.6 Ranking The Influential Variable in Determining the Position

The figure 6 scrutinizes the pivotal variable influencing race position determination, employing the y axis to depict ranking scheme within the 0.0 to 0.35 range. Across the x axis, variables such as Track, No, Driver, Team Quantity, Starting Grid, Laps, and Fastest Lap are considered.

The analysis unveils that starting position and lap holds the most influenced, securing the top rank with a coefficient of 0.34, 0.35. So the insight suggest that starting

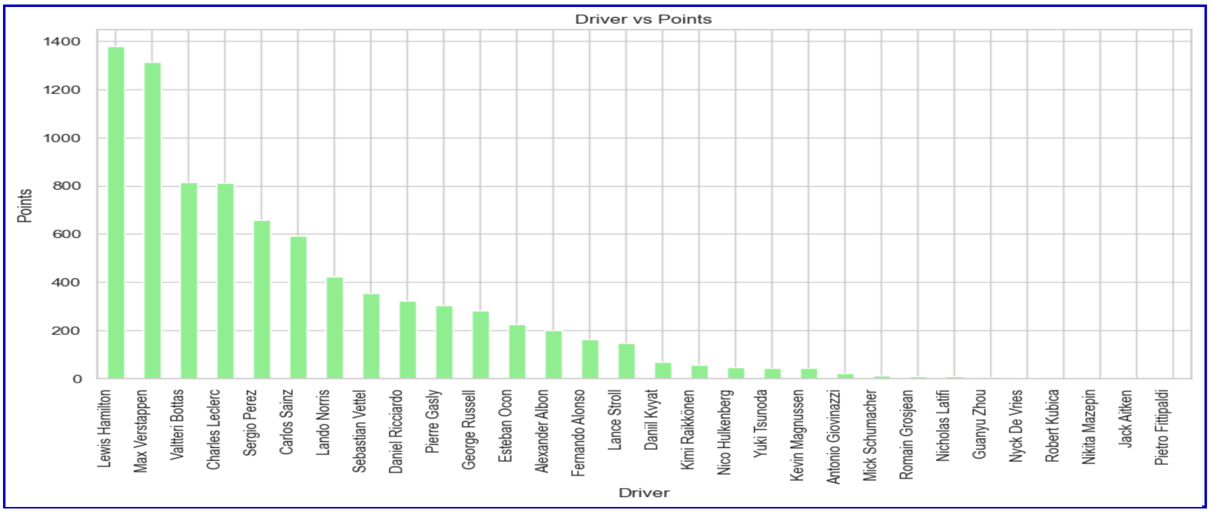


Figure 5: Points Comparison Across Different Driver

position(Starting Grid) and Lap is the most important in determining a driver's final position.On the other hand the least influential variable has been identified as Fastest Lap, obtaining Rank 7 with a coefficient of 0.01.
Therefore, a driver starting in the 1st position is likely to have a better chance of winning or achieving a top position in the race, according to the analysis.

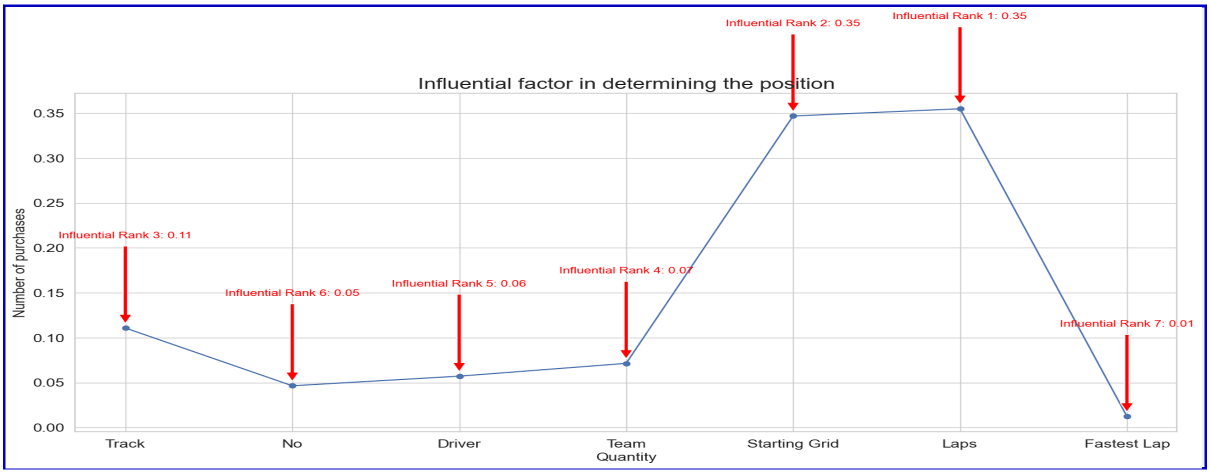


Figure 6: Influential factor in determining the position

5.7 Ranking The Influential Factors In Points Distribution

The figure 7 meticulously examines the influential factors in determining points, utilizing the y axis to represent the ranking within the 0.0 to 0.4 range. Across the x axis, selected variables including Track, No, Driver, Team Quantity, Starting Grid, Laps, and Fastest Lap are considered.
The key finding is that 'Starting Grid' emerges as the most influential variable, securing the top influential rank with a notable coefficient of 0.46. This implies a strong correlation between the starting grid position and the accrued points.In other words, a favorable starting position significantly contributes to a driver's overall points allocation.

On the other hand, 'No'(Car number) has been identified as the least influential variable, obtaining influential Rank 7 with a relatively lower coefficient of 0.04. This nuanced ranking provides valuable insights into the varying impact of different attributes on the overall points distribution,offering a clearer understanding of their respective roles in influencing performance outcomes.

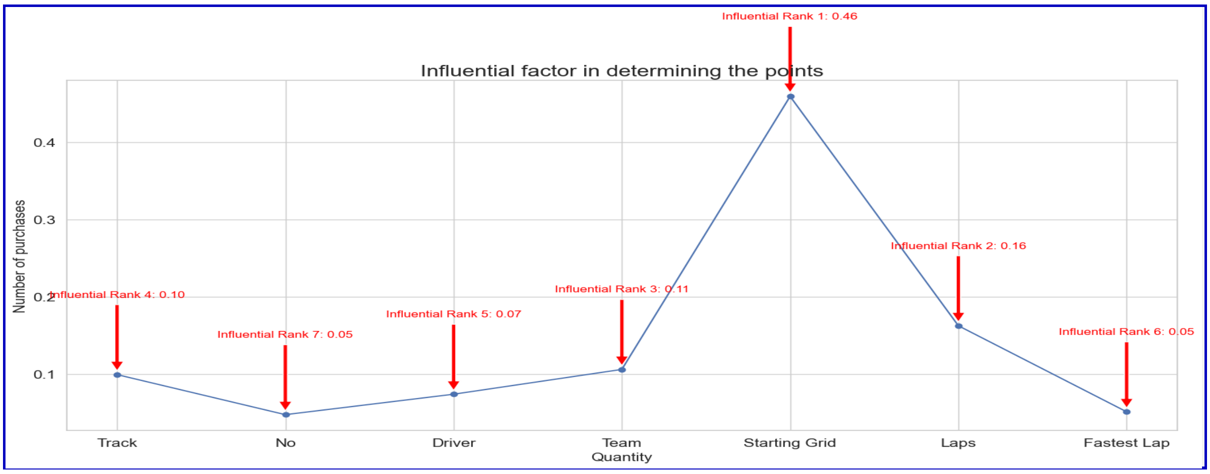


Figure 7: Influential factor in determining the points

5.8 Model Implementations

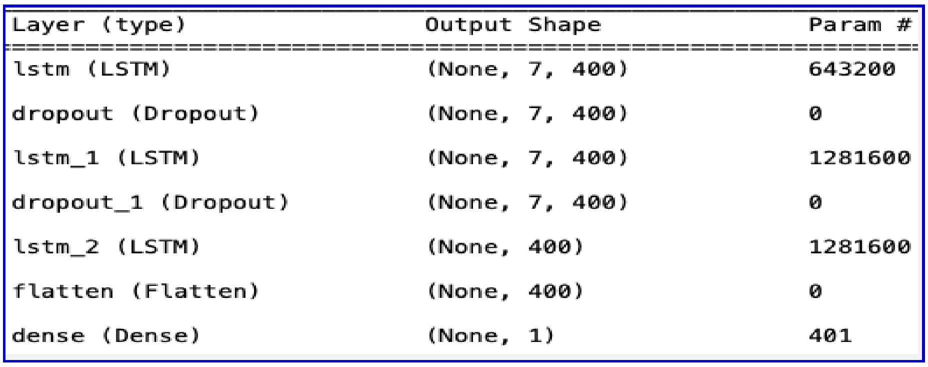
The study employs deep learning model for the prediction task specifically the Long Short Term Memory model and Gated Recurrent Unit .The objective of the study is to predict the end position of driver and team , using the preprocessed dataset for training consists of a feature set (X train) and a corresponding target variable (Y train). The X train includes relevant factors(Variable) such as 'Track,' 'Team,' 'Starting Grid,' 'Laps,' 'Points,' and 'Fastest Lap.' These features are essential input parameters for the regression algorithms, serving as the basis for predicting positions. The Y train represents the target variable, indicating the positions of entities in the dataset.

The primary objective of the LSTM and GRU models is to learn intricate patterns and dependencies within the provided X train features to make accurate predictions about the positions of entities. By leveraging the capabilities of recurrent neural networks, these models can effectively capture the sequential and temporal aspects of the input data, enabling them to classify positions with a high level of accuracy in the regression task.

5.8.1 Long Short Term Memory Model

The structure outlined in the figure 8 describes an LSTM model for a specific task. The model consists of three LSTM layers, each followed by a dropout layer, a flatten layer, and a dense layer. SThe model is adept at capturing intricate patterns within sequential data.The initial LSTM layer process sequence of length 7 with 400 units, incorporating 643,200 parameters, including weights and biases. By following dropout layer introduces regularization by randomly setting a fraction of input units to zero during training.The second LSTM layer, similar to the first, processing sequences of the same length and units but with a different set of 1,281,600 parameters. The flatten layer transforms the output into a one dimensional array of shape (None, 400), preparing it for the final dense layer.

It has 401 parameters, which include weights and biases. The dropout layers assist in preventing overfitting, and the overall parameter count is influenced significantly by the recurrent nature of the LSTM layers.



```
Layer (type)                 Output Shape              Param #
==============================================================
lstm (LSTM)                  (None, 7, 400)            643200

dropout (Dropout)            (None, 7, 400)            0

lstm_1 (LSTM)                (None, 7, 400)            1281600

dropout_1 (Dropout)          (None, 7, 400)            0

lstm_2 (LSTM)                (None, 400)               1281600

flatten (Flatten)            (None, 400)               0

dense (Dense)                (None, 1)                 401
```

Figure 8: LSTM Model Structure

5.8.2 LSTM Model Training Loss vs Validation Loss Analysis for Team Ranking Prediction

The provided epoch result reveals the training and validation loss values during the 50th epoch of the model training process. The x axis, denoted as Epoch, represents the progression of training epochs, while the y axis, labelled as Loss, which indicates the associated loss values. In this particular case, at the end of the 50th epoch the training loss is reported as 2.9191, signifying the average loss over the training dataset during this epoch.Concurrently, the validation loss is recorded as 2.35 representing the average loss over the validation dataset during the same epoch.
The comparison between training and validation loss as shown in the graph 9, which is pivtol in assessing the model's performance and generalization ability. In conclusion the values suggest that our model has undergone a reasonably effective training process, demonstrating a balanced learning approach and the potential for accurate predictions on both seen and unseen data.

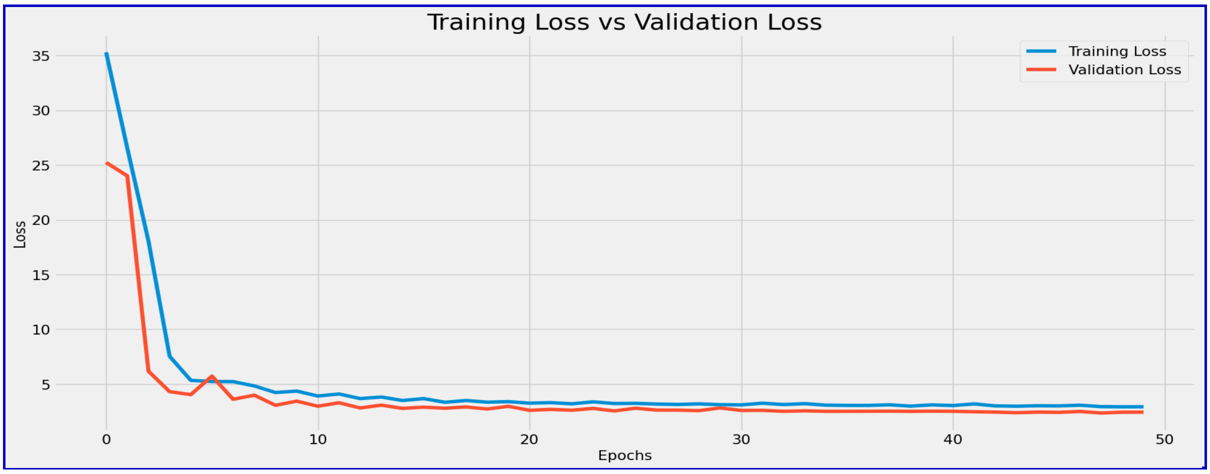


Figure 9: LSTM Model Training Loss vs Validation Loss for Team

5.8.3 LSTM Model Training Loss vs Validation Loss Analysis for Driver Ranking Prediction

The training and validation loss values for the 50th epoch of the models training process are displayed in the epoch result that is provided. The training epoch progression is represented by the x axis epoch, and the corresponding loss values are shown by the y axis loss. At epoch 50, the training loss is reported as 3.31, reflecting the average loss over the training dataset during this specific epoch. Simultaneously, the validation loss is documented as 2.35, indicating the average loss over the validation dataset during the same epoch. Analyzing these values provides a crucial understanding of the model's performance and its ability to generalize to unseen data. The close proximity of the training and validation loss values as shown in graph 10 suggests that the model is not overfitting to the training data, demonstrating a balanced learning process. Continuous monitoring of these loss values across epochs is essential for evaluating the training dynamics and ensuring the model's capability to generalize Analysingll to new, unseen data.
In this case the provided loss values at epoch 50 suggest a reasonably effective training process for the LSTM model, although further analysis and monitoring across epochs may provide a more comprehensive understanding of its overall performance

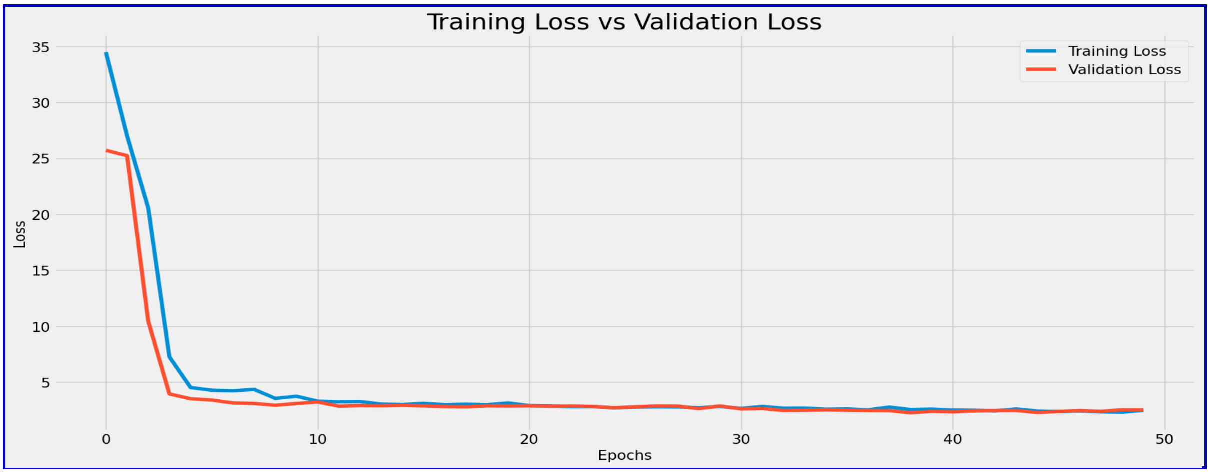
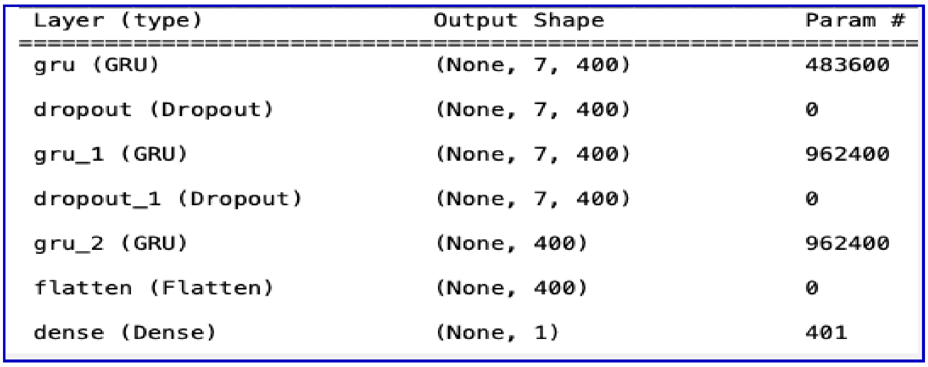


Figure 10: LSTM Model Training Loss vs Validation Loss for Individuals

5.8.4 Gated Recurrent Unit Model

The structure outlined in figure 11 describe an GRU model configuration with multiple layers, each contributing to the model's ability to capture sequential patterns and make predictions. The first GRU layer, labelled 'gru,' has an output shape of (None, 7, 400), indicating that it processes sequences of length 7 with 400 units in each sequence. This layer contributes 483,600 parameters, which include weights and biases. Following the GRU layer, a dropout layer labelled 'dropout_2' introduces regularization by randomly setting a fraction of input units to zero during training, mitigating overfitting. The second GRU layer, denoted as 'gru_1,' mirrors the first in terms of processing sequences of length 7 with 400 units, contributing an additional 962,400 parameters. A corresponding dropout layer labelled 'dropout_3' follows, serving the same regularization purpose as before. The third GRU layer, labelled 'gru_2,' processes sequences with a length of 400 units, resulting in an output shape of (None, 400) and 962,400 parameters. Subsequently, a flatten layer labelled 'flatten_1' transforms the output into a one dimensional array of shape (None, 400), preparing it for the final layer. The last layer, a dense layer labeled 'dense_1,' produces the model's output with a single unit indicating the predicted result. This layer has 401 parameters, which include weights and biases. The GRU model architecture, like the LSTM model, is designed to capture sequential dependencies and temporal dynamics in the data, making it suitable for tasks involving sequences or time series data



```
 Layer (type)                Output Shape              Param #
=================================================================
 gru (GRU)                   (None, 7, 400)            483600

 dropout (Dropout)           (None, 7, 400)            0

 gru_1 (GRU)                 (None, 7, 400)            962400

 dropout_1 (Dropout)         (None, 7, 400)            0

 gru_2 (GRU)                 (None, 400)               962400

 flatten (Flatten)           (None, 400)               0

 dense (Dense)               (None, 1)                 401
```

Figure 11: GRU Model Structure

5.8.5 GRU Model Training Loss vs Validation Loss Analysis for Team Ranking Prediction

The epoch results provide a insight into the training and validation loss values upon reaching the 50th epoch in the GRU model training process. The x axis represents the progression of training epochs, and the y axis indicates the corresponding loss values. At epoch 50, the training loss is reported as 2.89, representing the average loss over the training dataset during this epoch. Simultaneously, The validation loss is concurrently documented as 2.5016, representing the average loss over the validation dataset in the same epoch. The comparison between training and validation loss as show in the graph 12 is crucial for evaluating the model's generalization performance.

The values suggest that the model is learning well from the training dataset and is also capable of generalizing to unseen data, as evidenced by the relatively close proximity of the training and validation loss values. Additional monitoring and analysis across epochs may yield more information about the GRU model's overall performance and generalizability.

5.8.6 GRU Model Training Loss vs Validation Loss Analysis for Driver Ranking Prediction

The provided epoch result sheds light on the training and validation loss values at the culmination of the 50th epoch in the GRU model training process. The x axis, representing Epoch, signifies the progression of training epochs, while the y axis, labelled as Loss, illustrates the associated loss values. At epoch 50, the training loss is reported as 2.3424, denoting the average loss over the training dataset during this specific epoch. Simultaneously, the validation loss is documented as 2.6037, indicating the average loss over the validation dataset during the same epoch. Continuous monitoring of these loss values throughout epochs is essential for evaluating the training trajectory and ensuring

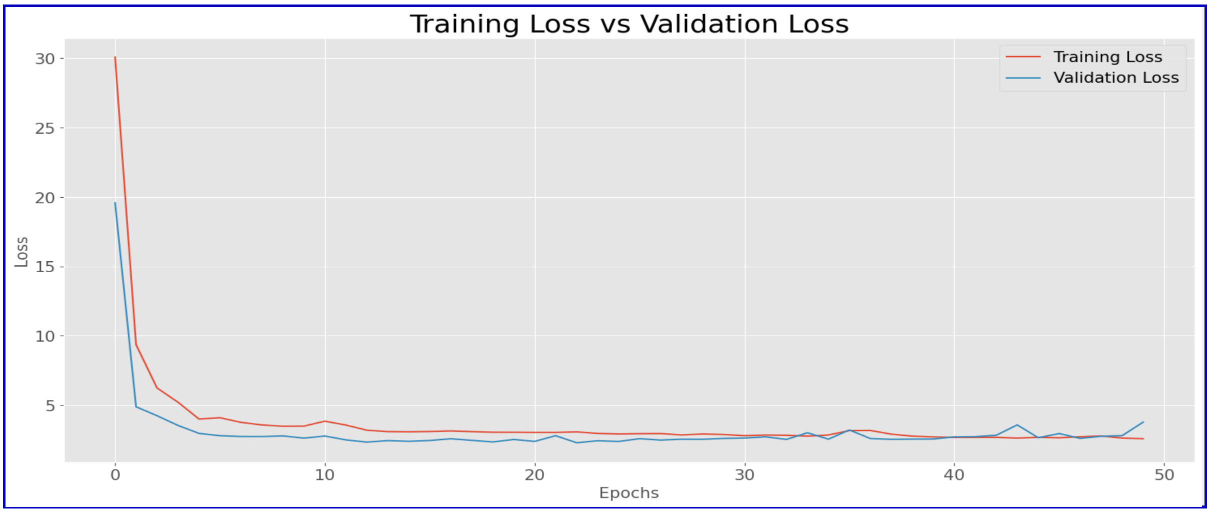


Figure 12: GRU Model Training Loss vs Validation Loss for Team

the models robustness in handling new, unseen data.
In this specific instance, the provided loss values at epoch 50 suggest a reasonably effective training process for the GRU model, which is clear from the graph 13. Further analysis and monitoring across epochs could provide additional insights into the overall performance and generalization capabilities of the GRU model.

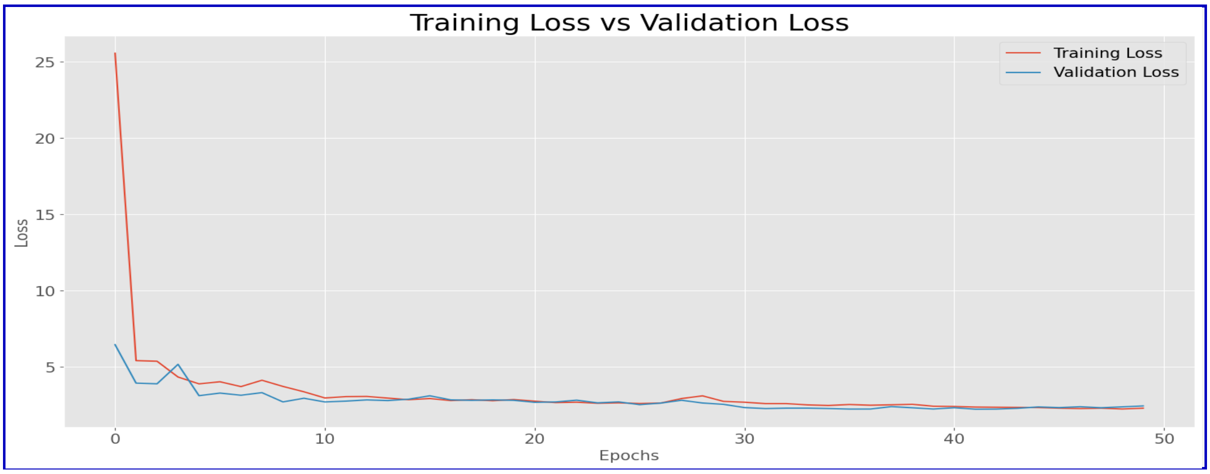


Figure 13: GRU Model Training Loss vs Validation Loss for Individuals

6 Results and Evaluation

6.1 LSTM vs GRU Model Evaluation Metrics Comparison Graph for Team and Indvidual

The figure 15 and 14 the graph compares the performance of Long Short Term Memory and Gated Recurrent Unit models based on three key evaluation metrics: Mean Squared Error , Root Mean Squared Error and Mean Absolute Error. The x-axis of the graph delineates these metrics, while the y-axis reflects the respective values for both models in

Team and individual

In terms of MSE for team, the figure 14 illustrates that the LSTM model outperforms the GRU model, as evidenced by a lower MSE value for the LSTM model (2.86) compared to the GRU model (3.36). This suggests that the LSTM model demonstrates superior accuracy in predicting positions, minimizing the squared differences between predicted and actual values. Moving to the RMSE metric, the figure 14 reinforces the superior performance of the LSTM model, with a lower RMSE value of 1.69 compared to the GRU model's RMSE of 1.83. Where as MSE for individual, the figure 15 showcases the LSTM model's superior performance with a lower value of 2.65 compared to GRU's 2.70. This implies that the LSTM model excels in accuracy, minimizing the squared differences between predicted and actual positions. Moving on to RMSE, the graph reinforces the LSTM model's dominance, presenting a lower value of 1.56 against GRU's 1.64. The decreased RMSE for LSTM underscores its precision in minimizing the spread around the true positions.A lower MAE value of 0.90 for LSTM against GRU's 0.98, the LSTM model demonstrates greater accuracy in predicting positions based on absolute differences.

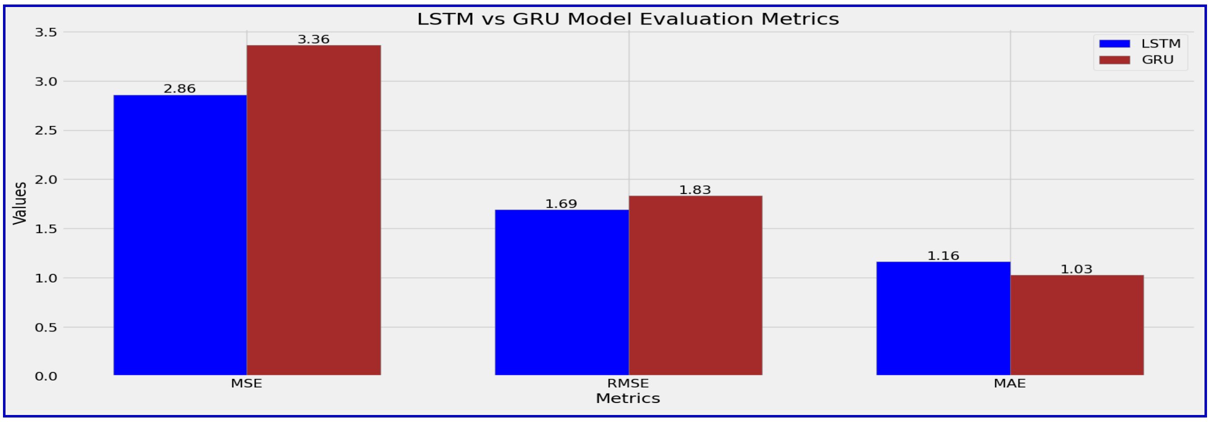


Figure 14: LSTM vs GRU Model Evaluation Metrics Comparison Graph for Team

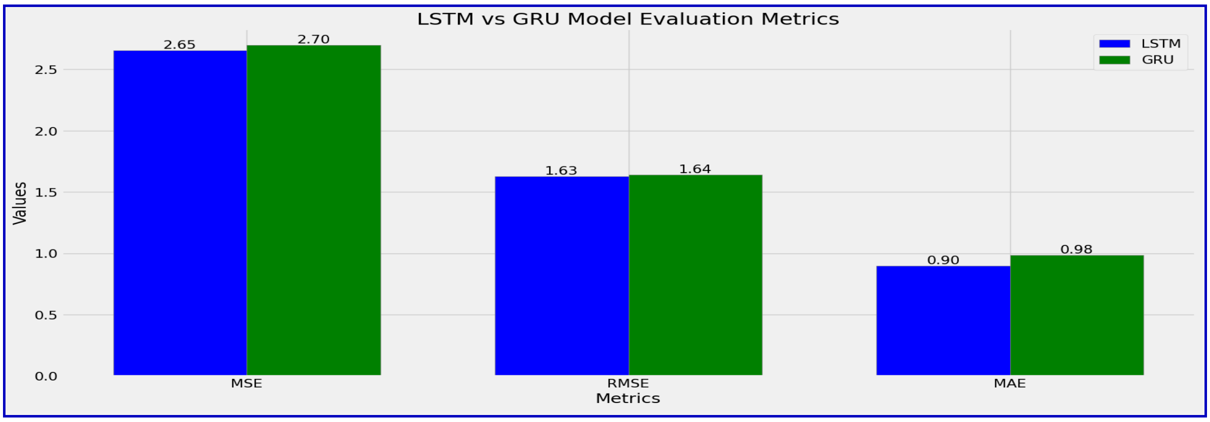


Figure 15: LSTM vs GRU Model Evaluation Metrics Comparison Graph for Individual

6.2 Conclusion

In conclusion, the comparative analysis of LSTM and GRU models for predicting Formula One race outcomes for both individuals and teams demonstrates nuanced differences in performance metrics. The LSTM model outperforms in terms of accuracy and precision, as indicated by lower Mean Squared Error and Root Mean Squared Error values, particularly in the individual category. On the other hand, the GRU model excels in Mean Absolute Error, showcasing superior accuracy in predicting positions based on absolute difference

6.3 Case Study 1- Predicting Driver Race Positions

In this case study, the primary objective is to develop a precise and reliable machine learning model for predicting the finishing positions of Formula One drivers in races. The model utilizes key input features such as the track, car number, driver, starting grid position, number of laps, points earned, and fastest lap achievement.The model, which has already been trained now effectively analyzes key input features ,the data input and encoding process began with the creation of dictionaries to encode features, facilitating user interaction in inputting race details enabling users to input race details, including the track, car number, driver, starting grid position, laps completed, points earned, and fastest lap status. The user inputting detail for the predicting race position where the pre trained model processed the information to predict the driver's race position,.A work flow is showin in figure 16

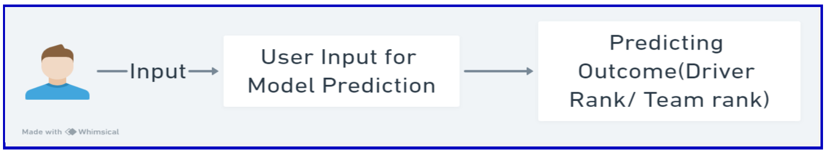

Input
User Input for Model Prediction
Predicting Outcome(Driver Rank/ Team rank)
Made with Whimsical

Figure 16: User Interaction

6.4 Case study 2 - Predicting Constructer postion

The experiment aims to develop a machine learning model for accurately predicting the team position in Formula One by utilizing input features such as track details, team information, starting grid position, laps, points earned, and fastest lap status.The user is asked to input data to make predictions in a race and predict the team position in the race. Then, the pre trained model processes the data to make that prediction. The trained model was then employed to predict the finishing position based on the provided information, and the predicted position was displayed to the user. The interpretation of the predicted position offered valuable insights into the model's expectations for the race, indicating that the model anticipated the team to secure a position in the team rankings.

6.5 Discussion

This case study showcases the practical implementation of machine learning for predicting Rank Forecasting in Formula One Racing relying on specific input features. The model's predictions undergo further assessment and refinement with additional data, enhancing its accuracy and applicability in real world scenarios. This research predicts individual and team places in Formula One races using two deep learning models: LSTM and GRU. Data from the years 2019, 2020, 2021 and 2022 is employed, with the first three years used for training and the last year for model validation. The project incorporates a time-series approach during model training and evaluation, and it includes exploratory data analysis to extract valuable insights. Deep learning algorithms are employed to predict both team and individual performance in this comprehensive project.

7 Conclusion and Future Work

After a careful study of previous literature, specifically in motorsports, we selected variables and employed the ExtraTreesRegressor machine learning model to determine the influential factors in predicting the final position of drivers. We also utilized LSTM and GRU for predicting rank positions for drivers and teams. Additionally, in Root Mean Squared Error, the LSTM model repeatedly shows superior precision and accuracy compared to the GRU model across team and individual positions as an evidenced by lower Mean Squared Error and Root Mean Squared Error values. This reinforces the LSTM model's effectiveness in minimizing errors and capturing meaningful patterns in the data. This analysis highlights key distinctions between LSTM and GRU models, showcasing their specialized strengths. These insights empower entities to refine strategies and navigate Formula One complexities with data driven precision. The study advocates for ongoing exploration and integration of advanced machine learning in sports analytics, especially in Formula One.

References

Allender, M. (2009). The role of driver experience in predicting the outcome of nascar races: an empirical analysis, *The Sport Journal* **12**(2).

Bell, A., Smith, J., Sabel, C. E. and Jones, K. (2016). Formula for success: multilevel modelling of formula one driver and constructor performance, 1950–2014, *Journal of Quantitative Analysis in Sports* **12**(2): 99–112.

Budzinski, O., Feddersen, A. et al. (2020). Measuring competitive balance in formula one racing, *Outcome uncertainty in sporting events: Winning, losing and competitive balance* pp. 5–26.

FRANSSEN, K. (2021). *COMPARISON OF NEURAL NETWORK ARCHITECTURES IN RACE PREDICTION*, PhD thesis, tilburg university.

Gramacy, R. B., Jensen, S. T. and Taddy, M. (2013). Estimating player contribution in hockey with regularized logistic regression, *Journal of Quantitative Analysis in Sports* **9**(1): 97–111.

Henderson, D. A. and Kirrane, L. J. (2018). A comparison of truncated and time-weighted plackett–luce models for probabilistic forecasting of formula one results.

Horvat, T. and Job, J. (2020). The use of machine learning in sport outcome prediction: A review, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **10**(5): e1380.

Issakhanian, E., Elkins, C. J., Lo, K. P. and Eaton, J. K. (2010). An experimental study of the flow around a formula one racing car tire.

Kapadia, K., Abdel-Jaber, H., Thabtah, F. and Hadi, W. (2020). Sport analytics for cricket game results using machine learning: An experimental study, *Applied Computing and Informatics* **18**(3/4): 256–266.

Lam, M. W. (2018). One-match-ahead forecasting in two-team sports with stacked bayesian regressions, *Journal of Artificial Intelligence and Soft Computing Research* **8**(3): 159–171.

Liu, X., Fotouhi, A. and Auger, D. J. (2023). Energy-optimal overtaking manoeuvres of formula-e cars, *Vehicle System Dynamics* **61**(8): 2023–2050.

Lotfi, S. et al. (2021). Machine learning for sport results prediction using algorithms, *International Journal of Information Technology and Applied Sciences (IJITAS)* **3**(3): 148–155.

Marino, A., Aversa, P., Mesquita, L. and Anand, J. (2015). Driving performance via exploration in changing environments: Evidence from formula one racing, *Organization Science* **26**(4): 1079–1100.

Okeyo, G., Chen, L., Wang, H. and Sterritt, R. (2014). Dynamic sensor data segmentation for real-time knowledge-driven activity recognition, *Pervasive and Mobile Computing* **10**: 155–172.

Peng, B., Li, J., Akkas, S., Araki, T., Yoshiyuki, O. and Qiu, J. (2021). Rank position forecasting in car racing, *2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 724–733.

Ruiz, H., Power, P., Wei, X. and Lucey, P. (2017). " the leicester city fairytale?" utilizing new soccer analytics tools to compare performance in the 15/16 & 16/17 epl seasons, *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1991–2000.

Schumaker, R. P., Solieman, O. K., Chen, H., Schumaker, R. P., Solieman, O. K. and Chen, H. (2010). Sports data mining: The field, *Sports Data Mining* pp. 1–13.

Tulabandhula, T. and Rudin, C. (2014). Tire changes, fresh air, and yellow flags: challenges in predictive analytics for professional racing, *Big data* **2**(2): 97–112.

Williams, J. and Li, Y. (2008). A case study using neural networks algorithms: horse racing predictions in jamaica, *Proceedings of the International Conference on Artificial Intelligence (ICAI 2008)*, CSREA Press, pp. 16–22.

Wojciechowski, P. and Wojtowicz, K. (2023). Challenges in designing measurement systems for formula one cars, *2023 IEEE International Workshop on Metrology for Automotive (MetroAutomotive)*, IEEE, pp. 57–61.