Formula 1 Championship Analysis

Team Members - Shraddha Gupta, Chaithra Nair, Aishwarya Mocherla

Statement of goal

The goal of this project is to analyze how various factors influence the **final race positions** of drivers in Formula 1 championships and leverage these insights to build predictive models. **Starting grid positions**, alongside variables such as **constructor performance**, **fastest lap times**, **constructor points**, and other **race-specific metrics**, are examined to uncover their influence on race outcomes.

This analysis is significant for multiple stakeholders:

- **For Teams & Drivers**: It offers actionable insights to optimize qualifying performance and race strategies, helping teams make informed decisions to improve race day outcomes.
- **For Fans**: The findings enhance understanding of how various factors impact race dynamics, making the sport more engaging and accessible to enthusiasts.
- **For Researchers**: The study contributes to motorsport analytics by providing a data-driven exploration of the complex interplay of factors influencing Formula 1 races.

By examining the relationships between these variables and final positions, the project identifies significant patterns and correlations that can enhance prediction accuracy. This comprehensive approach provides valuable insights into the dynamics of Formula 1 races and deepens the understanding of the factors driving success.

Data Description

The dataset used for this project originates from **Kaggle** and can be cross-referenced with official records from the **Formula 1 (F1) website** to ensure accuracy and reliability. It spans data from **1950 to 2023**, encompassing the entire history of the Formula 1 World Championship. However, for this analysis, data from **2018 to 2023** was selected to focus on recent race dynamics, current regulations, and their relevance to modern Formula 1. This period offers a more consistent representation of the factors influencing race outcomes in the contemporary competitive landscape.

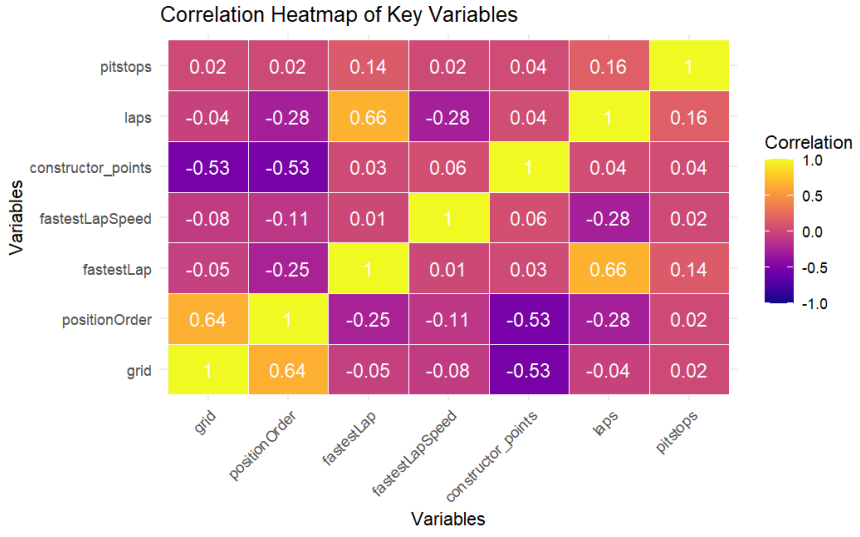
The dataset includes detailed information on **drivers**, **constructors**, and **races**, capturing essential features such as **starting grid positions**, **final race positions**, **fastest lap times**, **constructor points**, and other race-specific metrics. These insights are vital for the Formula 1 community, including teams, fans, and researchers, as they provide a structured framework for statistical analysis and predictive modeling. While specific to F1, the dataset holds broader relevance for exploring real-world performance metrics applicable to competitive sports and data-driven decision-making.

Key Variables and Their Measurement:

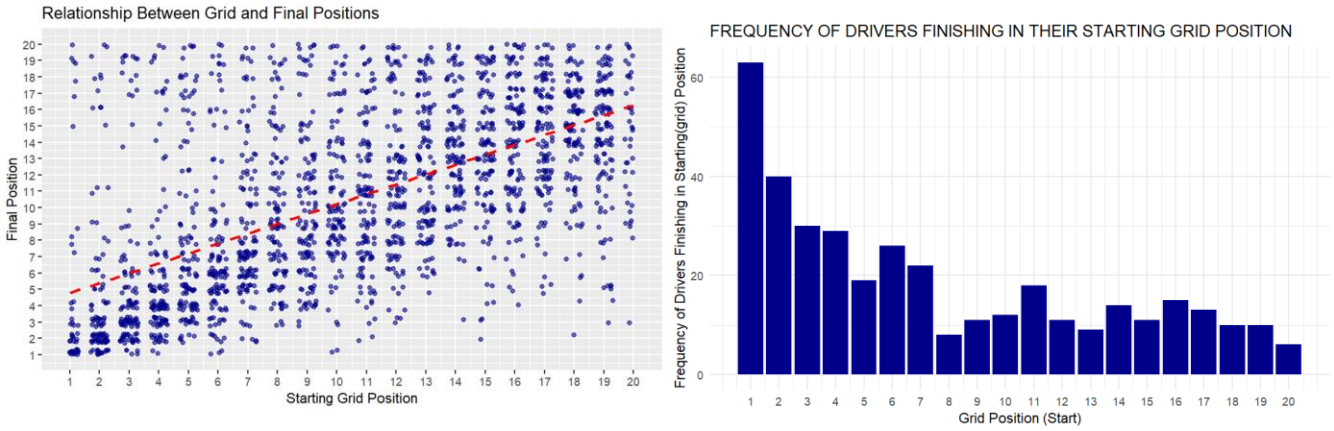
- **DriverID**: A unique identifier for each driver, recorded as an integer.
- **RaceID**: A unique identifier for each race, recorded as an integer.
- **Year**: The season or year of the race, recorded as an integer.
- **ConstructorID**: A unique identifier for each constructor (team), recorded as an integer.
- **Grid**: The starting position of the driver on the grid, recorded as an integer.
- **PositionOrder**: The final position of the driver in the race, recorded as an integer.
- **Points**: The points earned by the driver in the race, recorded as an integer.
- **FastestLap**: The fastest lap time achieved by the driver in the race, recorded as a numeric value.
- **Constructor Points**: The total points earned by the constructor in the race, recorded as an integer.
- **FastestLapTime**: The time duration of the fastest lap achieved, recorded as a string in HH:MM:SS format, converted to numeric for analysis.

Graphs

1. **Correlation Analysis of Key Variables in Formula 1:** The correlation heatmap provides a comprehensive overview of the relationships between critical variables influencing Formula 1 race outcomes, such as grid positions, final race positions, constructor points, laps completed, fastest lap times, fastest lap speeds, and pit stops. A strong positive correlation (0.64) between grid and final positions highlights the significant role of starting positions in determining race outcomes, while a moderate negative correlation (-0.53) between constructor points and final positions reflects the competitive edge provided by stronger constructors. Additionally, the positive correlation (0.66) between laps completed and fastest lap times indicates that drivers with consistent performance over the race achieve better lap times. Pit stops, however, show minimal correlations, suggesting a more complex or context-dependent influence on results.

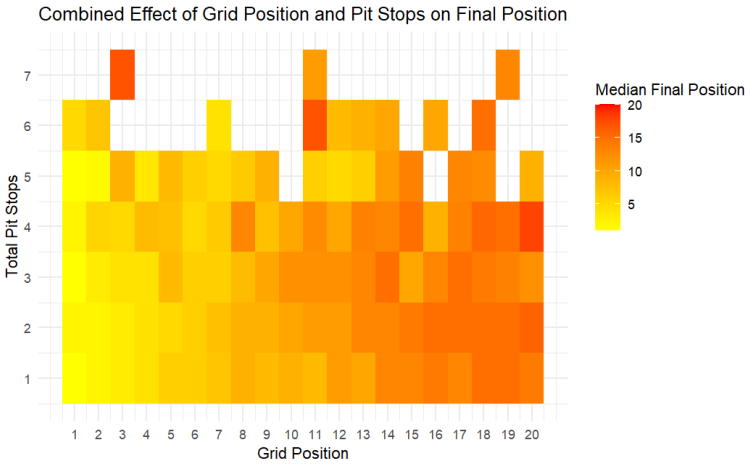


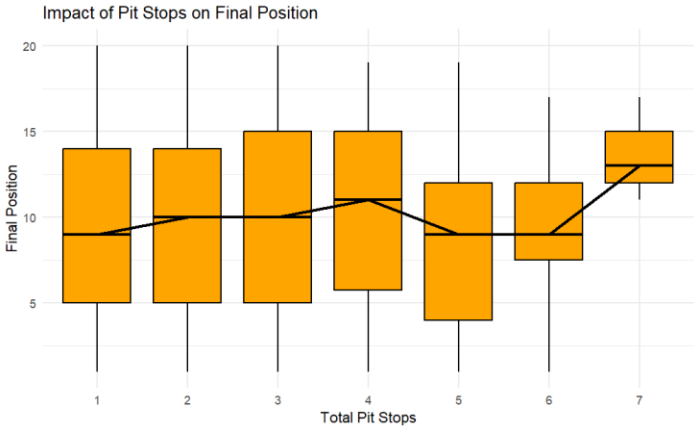
2. **Relationship Between Grid and Final Positions:** The scatter plot illustrates the relationship between starting grid positions and final race positions in Formula 1. A noticeable trend indicates that drivers starting closer to the front of the grid generally achieve better final positions, as evidenced by the positive correlation and the red regression line. However, the dispersion of points highlights variability, suggesting that while grid position plays a significant role, other factors such as driver skill, team performance, and race dynamics also influence race outcomes.



The bar chart highlights the frequency of drivers finishing in the same position as their starting grid placement. Notably, drivers starting from the front of the grid, particularly positions 1 and 2, are more likely to maintain their starting position throughout the race. This trend diminishes as the grid positions increase, suggesting that starting near the front provides a significant advantage. However, lower grid positions exhibit more variability, emphasizing the influence of race conditions, overtaking strategies, and team performance on final outcomes.

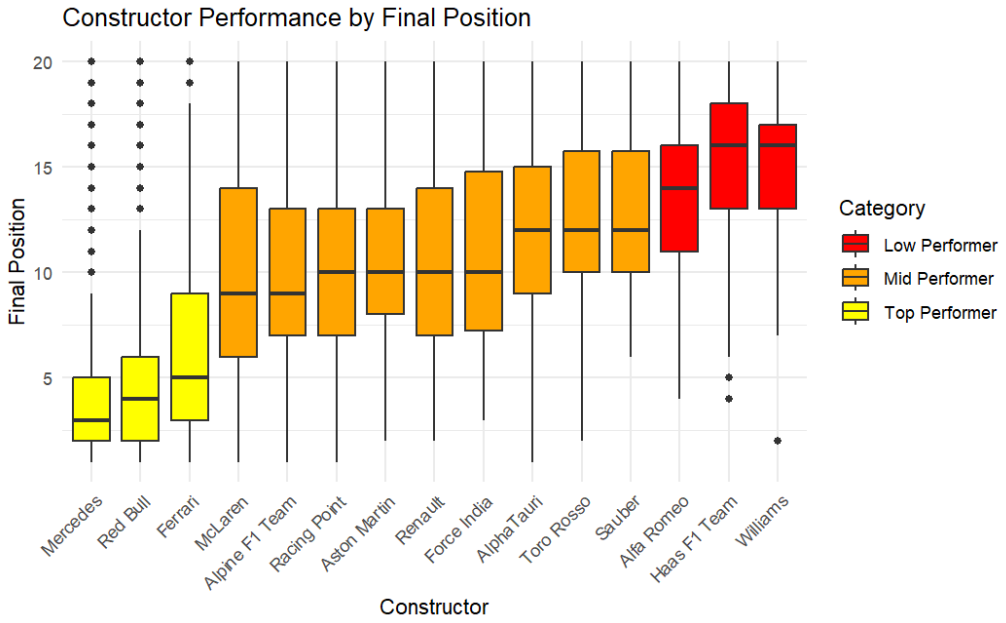
3. **Combined Effect of Grid Position and Pit Stops on Final Position:** Using a heatmap to depict this complex interaction allowed us to visually dissect how starting position combined with pit stop frequency impacts race outcomes. The gradation of colors illustrates a clear pattern where drivers starting further back benefit more significantly from a higher number of pit stops, likely due to aggressive overtaking strategies and optimal pit stop timing. This insight is particularly valuable for teams and drivers in formulating race strategies when not in pole positions.



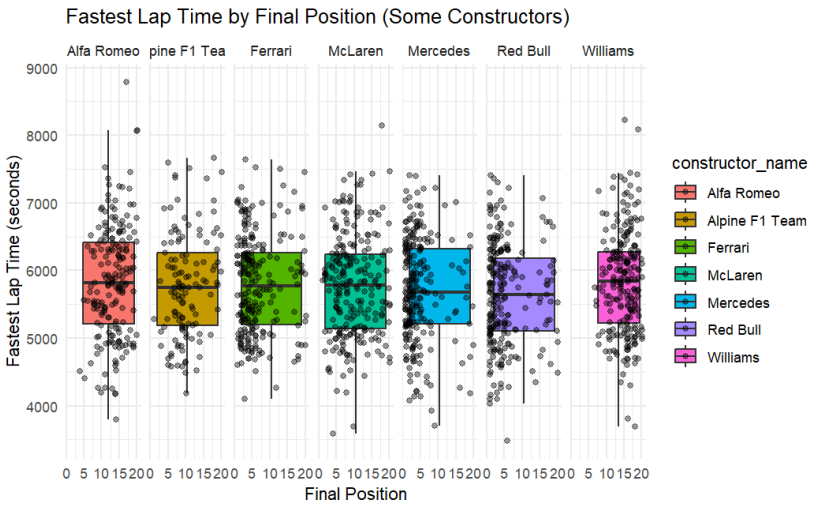


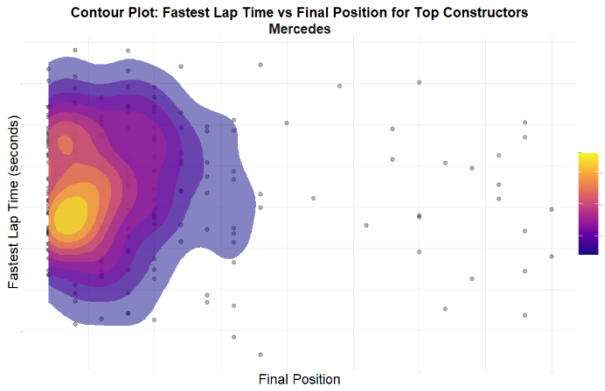
The boxplot reveals that the number of pit stops generally shows limited variability in final race positions, as median positions remain inconsistent across different pit stop counts. Given the weak trends and inconsistency, incorporating pit stops into the model may not significantly improve predictive accuracy.

4. Constructor Performance by Final Position: This box plot categorizes constructors into low, mid, and top performers based on their final positions across races. This visualization helps to differentiate performance levels among constructors, offering insights into consistency and competitiveness. The spread and median of the boxes reveal that top performers like Mercedes, Red Bull, and Ferrari not only finish in higher positions but also exhibit less variability in performance, underscoring their dominance in the sport.



5. **Impact of Fastest Lap Time on Final Position Across Constructors**: Teams like **Mercedes, Red Bull, and Ferrari** demonstrate dense regions at lower lap times and top positions, highlighting their consistency and competitive edge. Mid-field teams such as **McLaren** and **Alpine F1 Team** display moderate lap times with greater variability, reflecting the impact of race strategies and external factors. In contrast, lower-tier teams like **Williams** and **Alfa Romeo** show slower lap times and more scattered results, underscoring their performance gaps. Also, for instance the contour plot for Mercedes reveals a clear relationship between **Fastest Lap Time** and **Final Position** across top constructors, with faster lap times generally linked to better results. These patterns support the inclusion of **Fastest Lap Time** in the models as it effectively captures team-specific race performance trends.





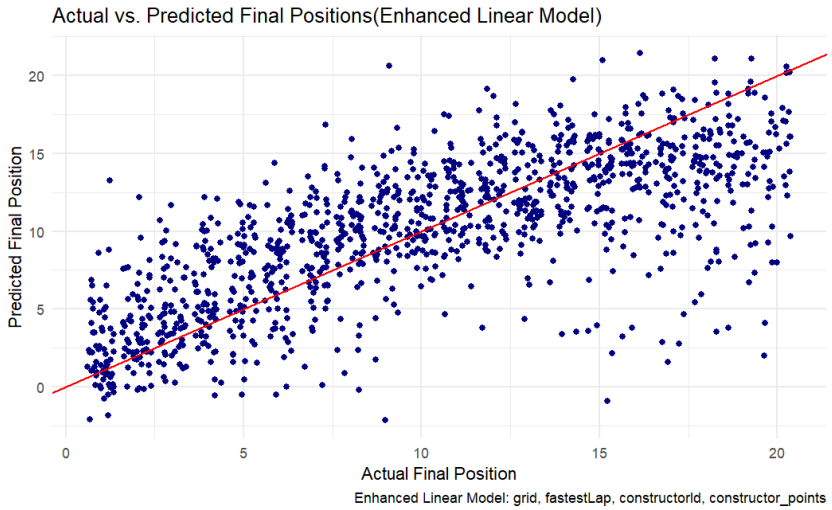
Each of these visualizations was meticulously crafted to ensure that they provide the most comprehensive insights into how various elements such as starting position, pit stop strategy, and constructor performance affect outcomes in Formula 1 racing, grounding our analysis firmly within the context of real-world race dynamics and strategic decisions.

Models:

The data was split into training (January to July) and testing (August onwards) sets to reflect the natural progression of a Formula 1 season. This approach ensures the model learns patterns from the first half of the season and tests its performance on the second half, aligning with real-world race timelines.

1. LINEAR MODEL (lm(formula = positionOrder ~ grid, data = training_data)

A simple linear regression model was developed to predict final race positions using starting grid positions. The model revealed a positive slope of 0.618, indicating that each increase in grid position worsens the final result by approximately 0.62 positions. With an R-squared value of 0.37, grid position alone explains only 37% of the variability, emphasizing the need for additional predictors. The model achieved RMSE values of 4.69 (test), with MAE around 3.6. While the scatter plot shows general alignment along the diagonal, significant variability for mid-field and lower positions highlights the limitations of relying solely on grid position for predictions.



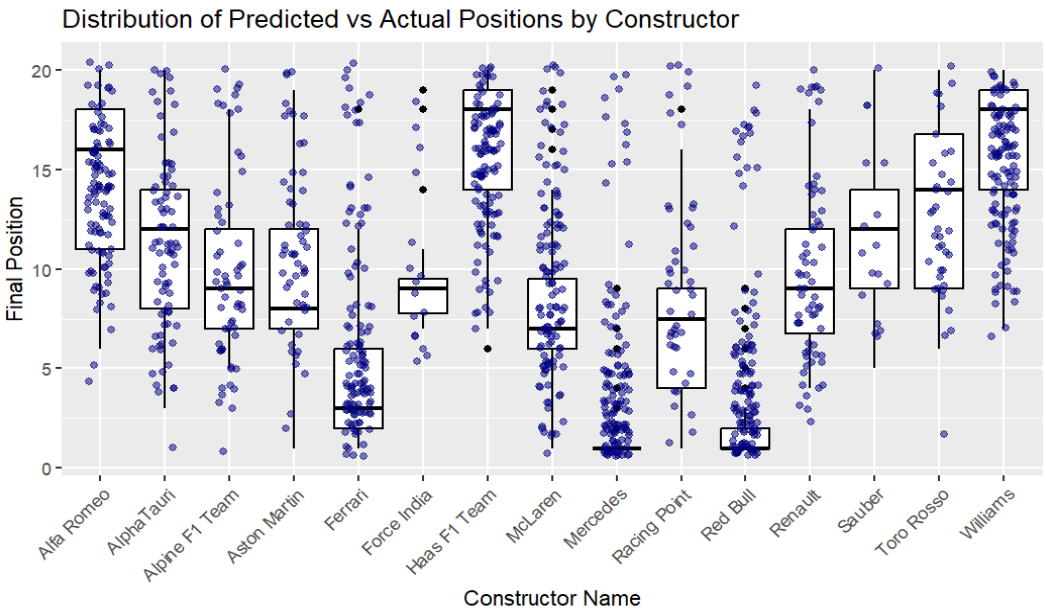
2. LINEAR MODEL WITH ADDITIONAL FEATURES: (lm(formula = positionOrder ~ grid + fastestLap + as.factor(constructorId) + constructor_points, data = training_data)

The enhanced linear regression model incorporates additional predictors—grid position, fastest lap, constructor ID, and constructor points—to improve predictive performance over the simple linear model. The Residual Standard Error (RSE) decreased to 3.97 (from 4.58 in the simple model), and the Adjusted R-squared increased to 0.528, indicating better explanatory power. Key variables, such as fastestLap and constructor_points, were found significant, highlighting their influence on race outcomes. Model performance on the test data yielded an RMSE of 4.31 and an MAE of 3.16, both improved from 4.69 and 3.68, respectively, in the simple model.

A key limitation of the linear regression model is that it treats the target variable, positionOrder, as continuous, even though it is inherently ordinal (representing race positions). This can lead to unrealistic predictions, such as negative values for final positions. To address this, an ordinal regression model was explored, which better aligns with the ordinal nature of the target variable.

3. ORDINAL REGRESSION MODEL: formula: positionOrder ~ grid + fastestLap + as.factor(constructorId) + constructor_points

The ordinal regression model improved prediction accuracy by treating positionOrder (final race position) as an ordinal variable. Incorporating grid, fastestLap, constructorId, and constructor_points, the model achieved a **Mean Absolute Error (MAE) of 2.9** and **Root Mean Square Error (RMSE) of 3.88**, showing significant improvement in performance.

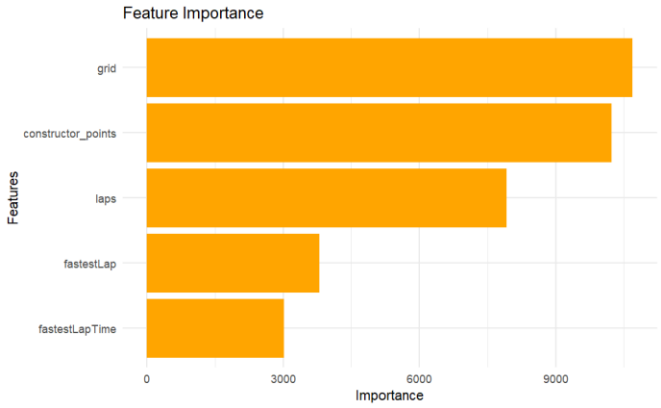


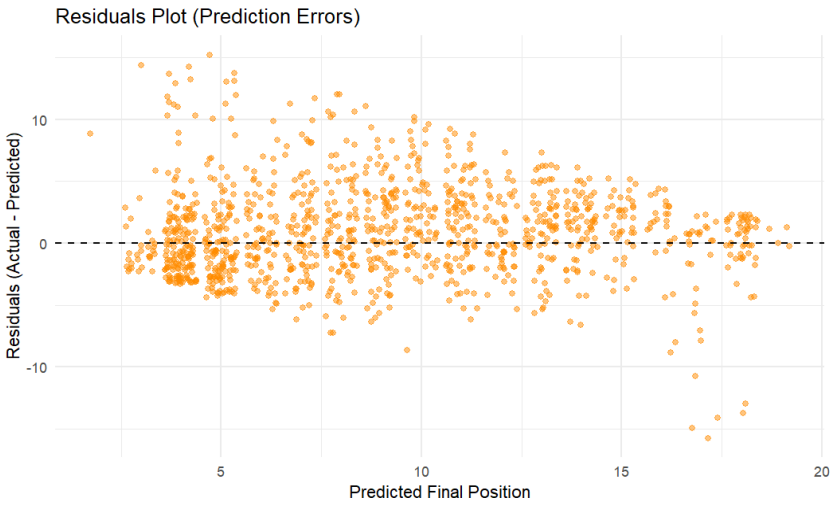
The plot reveals that the ordinal regression model performs well for top-performing constructors like **Ferrari**, **Mercedes**, and **Red Bull**, where predicted positions are tightly clustered and closely align with actual results, reflecting consistent race performance. In contrast, mid- and lower-tier constructors such as **Alfa Romeo**, **Williams**, and **Haas F1 Team** exhibit greater variability and noticeable gaps between predictions and actual outcomes, highlighting the model's struggle to account for unpredictable race dynamics. This suggests that while the model effectively captures patterns for dominant teams, its predictions for inconsistent performers are less reliable due to external factors like strategy changes or race incidents.

With the lowest **AIC value of 5942.28** compared to 7018.18 and 6689.32 for the simple and enhanced linear models, the ordinal model avoided unrealistic predictions and provided reliable insights. It stands out as the most accurate and robust approach for predicting Formula 1 race outcomes.

4. RANDOM FOREST MODEL: (positionOrder ~ grid + fastestLap + constructor_points + laps + fastestLapTime)

The Random Forest model was introduced to improve the prediction of final race positions by capturing complex, non-linear relationships in the data. The key predictors included grid position, fastest lap, constructor points, laps completed, and fastest lap time. This ensemble learning approach aggregated multiple decision trees to provide robust predictions, effectively addressing the limitations of linear and ordinal models.

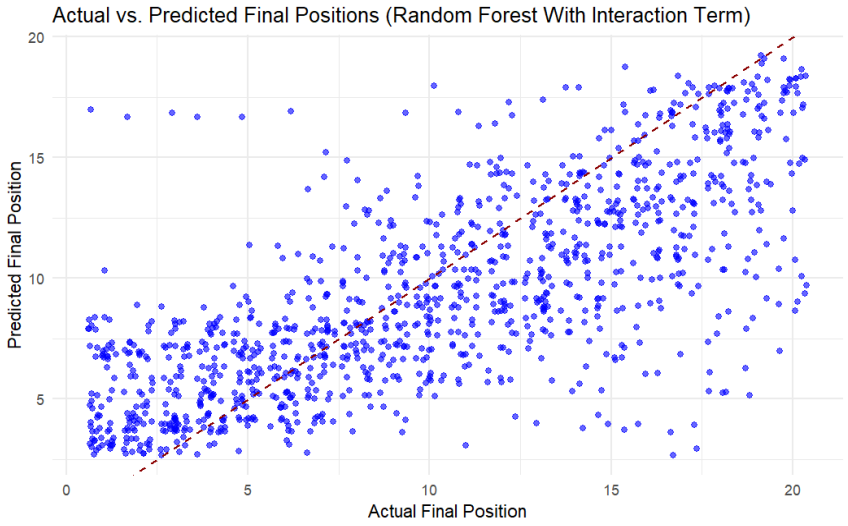




The model achieved a **Mean Absolute Error (MAE)** of **2.86** and a **Root Mean Square Error (RMSE)** of **3.85**, outperforming both the linear and ordinal regression models. Additionally, the feature importance plot highlights **grid position** and **constructor points** as the most influential variables, confirming their strong impact on race outcomes. The Random Forest model demonstrated its ability to handle the complexity of Formula 1 data and delivered reliable predictions, making it a valuable tool for race position forecasting.

5. RANDOM FOREST MODEL WITH INTERACTIVE ELEMENT: (formula: positionOrder ~ grid + constructor_points + laps + fastestLapTime + constructor * constructor points + fastestLap

To enhance the predictive power of the Random Forest model, an interaction term between constructor ID and constructor points was introduced. This interaction aimed to capture the relationship between constructor-specific performance and their earned points, which might influence final race positions.



The updated model achieved a Mean Absolute Error (MAE) of 2.94 and a Root Mean Square Error (RMSE) of 3.90 on the test data. While the slight increase in error compared to the previous Random Forest model suggests a minimal impact of the interaction term, it added depth to the analysis of constructor performance. Feature importance highlights the grid position, constructor points, and laps as the most influential predictors, followed by the interaction term, fastest lap, and fastest lap time. The inclusion of the interaction term provides a more nuanced understanding of how constructor-specific factors affect race outcomes.

Results:

MODEL	MAE	RMSE
Simple Linear Model (~grid)	3.68	4.69
Enhanced Linear Model (~grid + fastestLap + constructorId + constructor points)	3.16	4.31
Ordinal Regression Model (~grid + fastestLap + constructorId + constructor points)	3.26	4.54
Random Forest Model (~grid + fastestLap + constructor points + laps + fastestLapTime)	2.86	3.85
RF with Interaction Term (~grid + fastestLap + constructor points + laps + fastestLapTime + constructor points * constructor_id)	2.94	3.92

Conclusion

The analysis successfully explored the key factors influencing final race positions in Formula 1 championships using a combination of statistical and machine learning models. Starting from a simple linear regression model, which highlighted the significant role of starting grid positions, the analysis progressed to enhanced models by incorporating variables such as fastest lap, constructor performance, and interaction terms.

The **Ordinal Regression Model** effectively addressed the ordinal nature of the target variable, improving upon the limitations of the linear models. The **Random Forest Models** showcased their strength in capturing complex, non-linear relationships in the data, providing robust and accurate predictions. The interaction-enhanced Random Forest model added further insights into constructor-specific effects, demonstrating its value for analyzing Formula 1 race dynamics.

Key insights include the dominance of **grid positions** and **constructor points** as the most influential predictors of race outcomes. However, variables like laps completed and fastest lap times further highlighted the intricate interplay of race dynamics, team strategy, and driver performance. Top-performing constructors like **Mercedes**, **Ferrari**, and **Red Bull** exhibited consistent results, while mid- and lower-tier teams displayed greater variability, indicating the need for strategic interventions.

Overall, this comprehensive analysis provides actionable insights for teams to optimize race strategies and qualifying performance. The findings enhance fan engagement by demystifying race outcomes and contribute to motorsport analytics by offering a data-driven approach to understanding Formula 1 dynamics. Future work could incorporate additional contextual factors, such as weather conditions and race incidents, to further refine predictive accuracy.

Limitations

Our analysis of Formula 1 race outcomes and strategies has several limitations that are essential to understanding its scope and applicability. While the models capture key variables such as grid position, fastest lap, constructor performance, and pit stops, they do not include all factors influencing race outcomes. Critical elements like tire strategies, driver skills, weather conditions, safety car interventions, and on-track incidents significantly affect race dynamics but fall outside the scope of our dataset.

Additionally, the dataset covers only the 2018–2023 seasons, excluding earlier data that might reveal long-term trends and shifts in team performance or technological advancements in the sport. The analysis also does not account for mid-season changes such as driver substitutions, car upgrades, or evolving team strategies, which can dramatically alter the competitive landscape. Furthermore, the data lacks details on in-race events like lap-by-lap position changes and tire degradation, which are vital for understanding race dynamics and driver performance.

Addressing these gaps would enable a more comprehensive and predictive analysis, making the findings more robust and reflective of the multifaceted nature of Formula 1 racing.

Future Work

Incorporation of Additional Variables: Including critical factors such as tire strategies, weather conditions, and safety car deployments, which significantly influence race outcomes but are currently missing from the dataset.

Refining Feature Engineering: To refine our predictive models, an expansion of feature engineering by integrating more interactive terms could be highly beneficial. By creating new variables that represent interactions between existing features, such as the combination of grid positions with specific race tracks or weather conditions, we could capture more complex influences on race outcomes, thereby enhancing the model's accuracy and robustness.

Advanced Modeling Techniques: Explore advanced machine learning models, including deep learning and simulation-based approaches, to improve predictive accuracy and capture complex interactions among variables.

Interactive Dashboards: Develop interactive visualizations to present findings in a user-friendly format, enabling teams, researchers, and fans to explore the data and insights effectively.