**Analyzing Formula 1 Performance: Application and Insights from Decision Modeling**

Thanks to [Clicked](https://www.clicked.com/), I was part of an amazing opportunity and project to analyze Formula 1 performance data using decision modeling.

A race car lined up on a track

Description automatically generated

Formula 1 race

By leveraging linear regression, the aim was to predict win rates for drivers and understand the key factors influencing these outcomes.

This is a concise summary of my methodology, codes, and findings, which I am excited to share.

**Objectives**

* To understand the factors that influence driver win rates in Formula 1.
* To use linear regression to model the relationship between performance metrics and win rates.
* To visualize the findings and share them on GitHub and Streamlit.

**Hypothesis:**

- The efficiency and effectiveness of pit stops are critical factors that significantly impact the driver’s performance during a race.

- The diversity of the team (skills, experience, backgrounds) positively influences team performance and race outcomes.

Tools used: Python and VS code for interactive coding sessions, numPy, and Pandas for data manipulation and analysis, and scikit-learn for building machine learning models. I also used Matplotlib/Seaborn for data visualization.

**Data Overview**

Out of 14 datasets, I used datasets that included results, drivers, constructors’ results, pit stops, and constructor standings. The key features I focused on were:

* Total Races
* Total Wins
* Age
* Constructor Performance Points

**Exploratory Data Analysis (EDA)**

Through EDA, several insights were discovered:

* British drivers have the highest representation, with Hamilton leading in total points and wins.
* Ferrari is the top constructor, significantly contributing to driver success.
* Top drivers like Hamilton and Verstappen dominate recent seasons.

A graph of drivers with numbers

Description automatically generated with medium confidence

A graph with different colored lines

Description automatically generated

A graph with numbers and lines

Description automatically generated

**Cross section of EDA**

A graph of drivers by total

Description automatically generated

A graph of a number of drivers

Description automatically generated

A graph of a number of points

Description automatically generated with medium confidence

**EDA continued**

**Correlation Analysis**

**Understanding Correlation:**

Correlation is a statistical measure that describes the extent to which two variables are related. It ranges from -1 to 1.

1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship and 0 indicates no relationship.

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

**Correlation Matrix for Results and Pairplots**

**Interpreting the Correlation Matrix:**

In simple terms, a correlation matrix shows the correlation coefficients between variables in a dataset.

Positive Correlation (+): If the correlation coefficient is close to +1, it means that as one variable increases, the other variable also increases.

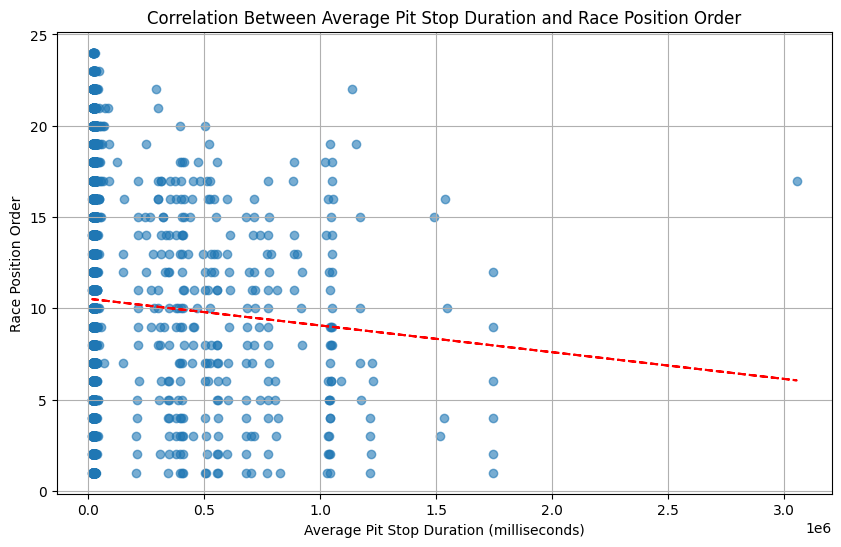
Negative Correlation (-): If the correlation coefficient is close to -1, it means that as one variable increases, the other variable decreases.

No Correlation (0): If the correlation coefficient is close to 0, it means there is no linear relationship between the variables.

The correlation matrices provided valuable insights:

* **High Correlation**: There was a strong correlation between race results and race identifiers.
* **Moderate Correlation**: Driver IDs and constructor IDs showed a moderate correlation, indicating some level of interdependence.
* **Low Correlation**: Grid positions had a low correlation with other variables, suggesting minimal direct impact on race outcomes.

**Pit Stop Duration vs. Race Position**



**Correlation between pit stop and race position**

Analyzing the correlation between pit stop duration and race position revealed:

* **Negative Correlation**: There is a slight negative correlation, as suggested by the downward slope of the red trend line. Shorter average pit stop durations are generally associated with better race positions.
* **Data Distribution**: Many data points are clustered around shorter pit stop durations (close to 0), indicating that most pit stops are relatively quick. There is a wide spread of race position orders for short pit stops, showing variability in race outcomes despite quick stops.
* **Outliers**: Some data points with very long pit stop durations (over 1.5 million milliseconds) are spread across various race positions, suggesting that extremely long pit stops can lead to poorer race positions, but there are exceptions.

**Implications for Performance**:

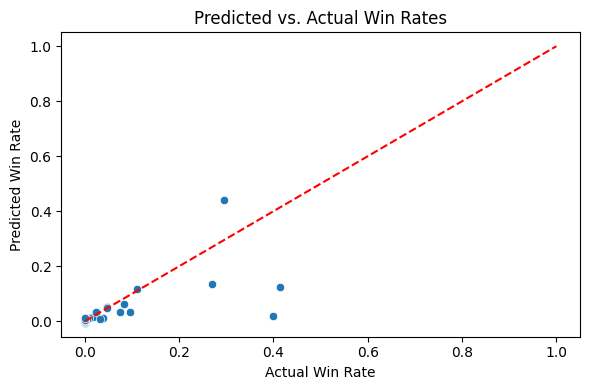
* Teams with consistently shorter pit stops may achieve better race positions.
* While pit stop duration is a factor, it is not the sole determinant of race position, as indicated by the variability in race positions for similar pit stop durations.

The chart demonstrates that faster pit stops tend to correlate with better race positions, though the relationship is not extremely strong. Many factors contribute to race performance, and pit stop efficiency is one of them. The negative trend line highlights the importance of minimizing pit stop duration to improve race outcomes.

**Linear Regression Model**

The linear regression model used the following features to predict win rates:

* Total Races
* Total Wins
* Age
* Constructor Performance Points



**Model Performance**:

* **Mean Squared Error (MSE)**: 0.0016, indicating a low average squared difference between the observed and predicted values. This means the model’s predictions are quite close to the actual values, indicating a good fit.
* **R-Squared (R²)**: 0.45, this means the model explains 45% of the variance in win rates. This is a moderate level of explanatory power, suggesting that while the model captures some important relationships, there is still considerable variance unexplained.

**Key Findings**:

* **Total Wins and Constructor Points**: These have a significant impact on win rates. This ties back to the primary hypothesis that the development and composition of a constructor team (as indicated by constructor points) significantly impact a team’s chances of winning races.
* **Pit Stops**: The analysis confirmed the secondary hypothesis that efficient pit stops are crucial, as shorter pit stop durations are associated with better race positions.
* **Model Insights**: While the model captures important relationships, there’s room for improvement with additional features and advanced techniques.

**Decision Tree Analysis**

In addition to the linear regression model, I explored decision tree analysis. This method helps identify the most significant variables and their interactions that influence driver performance. The decision tree provided a clear visualization of how different factors contribute to race outcomes, supporting the secondary hypothesis that team diversity and pit stop efficiency play critical roles.

A diagram of a number of samples

Description automatically generated with medium confidence

**Decision Tree**

**Conclusion**

This decision modeling exercise has provided valuable insights into the factors influencing Formula 1 driver performance. The linear regression model, despite its moderate explanatory power, highlights the importance of total wins and constructor points. The decision tree analysis further helps in understanding variable significance and interactions. Future work could enhance the models by incorporating more variables and sophisticated modeling approaches.

For a detailed look at the data and visualizations, you can check out my [GitHub repository](https://github.com/YummyAmy/Decision-Modeling) and the interactive [Streamlit app](https://decision-modeling.streamlit.app/) created for Formula 1 model.

Feel free to connect with me on [LinkedI](http://www.linkedin.com/in/ameti-obong-u-395a25111)n for more insights and updates on more data analytics. Let’s keep driving!!!