Formula 1 Data Analysis

Introduction

Formula 1 (F1) is the pinnacle of motorsport, combining advanced technology, elite driving, and strategic teamwork. The Constructor Championship, awarded to the team with the most points in a season, represents excellence in engineering and performance. Predicting these outcomes is valuable for teams optimizing strategies, fans seeking deeper engagement, and analysts exploring race dynamics. This project aims to develop a machine learning model to predict race position categories ("Other," "Top 10," "Top 3") for constructors by analyzing historical data from 2018 to 2022 and validating on 2023 data. Key factors such as driver reliability, team efficiency, and environmental conditions are incorporated to ensure accurate predictions.

Data Gathering

APIs Used:

- **Ergast API:** Provides extensive historical F1 data, including race results, qualification results, driver information, and standings.
- **FastF1 API:** Offers detailed weather conditions.

Season Schedule

The season schedule dataset includes key details such as the round number, race name, circuit information, and geographical data, including latitude and longitude. This dataset outlines the chronological order of races, capturing the specific dates and times for each event every year.

Weather Data

Features: Air Temperature, Humidity, Pressure, Rainfall, Track Temperature, Wind Direction, Wind Speed.
Usage: Weather data plays a significant role in Formula 1 races by influencing car performance, driver strategy, and overall race outcomes:

- Temperature impacts tire performance, with extreme heat accelerating wear and cold reducing grip.
- Rainfall challenges drivers with reduced visibility and grip, necessitating strategic tire changes.
- Wind alters car stability, especially on high-speed straights or during overtakes.
- Humidity strains drivers physically, affecting concentration and stamina.
- Altitude affects engine cooling and downforce due to thinner air.

Integration in Model:

Encoded weather data into categories for better interpretability and used features such as wind direction and speed to assess car stability and predict race outcomes. This allows the model to simulate real-world scenarios where weather conditions directly impact race results.

Race Data

The race data serves as the primary dataset, capturing the outcomes of each race across multiple seasons. It includes essential details such as race identifiers, year, driver and constructor information, and race-specific results like positions, points scored. Additionally, it provides key performance metrics, including fastest lap times and average speeds. This dataset forms the backbone of the analysis, offering insights into standings, race-day dynamics, and overall performance trends for both drivers and constructors.

Qualification Data

The qualification data provides detailed insights into the drivers' performance during the qualifying sessions. It includes information such as round number, driver identifiers, and qualifying times for Q1, Q2, and Q3 sessions. This dataset is pivotal in determining the starting grid positions, which play a significant role in influencing race-day strategies and outcomes. By analyzing qualifying performance, the dataset helps assess how initial positioning correlates with overall race results.

Driver Data

The driver data contains essential personal and identification details for each Formula 1 driver participating in the seasons from 2018 to 2023. Key features include unique driver IDs, names, season numbers, driver codes, dates of birth, and nationalities. This dataset is crucial for linking individual drivers to their performance metrics, enabling a comprehensive analysis of how driver-specific factors contribute to race and season outcomes.

Pit Stop Data

The pit stop data captures critical metrics related to team strategies and pit crew efficiency. It includes the number of pit stops made during a race, the average time taken per stop, and the total pit stop duration for each event. This dataset is vital for analyzing how well-executed pit stops influence race outcomes, with faster and more efficient stops often contributing to better

race positions. By examining these metrics, the data provides insights into the strategic decisions made by teams during races.

Data Integration and Cleaning

Data cleaning involved combining multiple datasets, including driver, qualification, race, weather, and schedule data, into a unified structure. This integration was achieved by merging on shared keys such as `Year`, `Round`, Constructor ID and and `Driver ID`. Unnecessary columns, including `Wins`, `Constructor Name`, and `EventName`, were removed to streamline the dataset and ensure relevance to the analysis. These steps ensured a clean and consistent dataset, ready for feature engineering and modeling.

Error Handling: Adjusted error probabilities for drivers (Driver Confidence) and constructors (Constructor Confidence) to reflect reliability.

Feature Engineering

Feature engineering was conducted to enhance the dataset's usability and improve model performance. These transformations prepared the dataset for effective and efficient modeling.. Key steps included:

- **Categorization**: Weather data such as AirTemp, Humidity, Rainfall, WindSpeed, and WindDirection were categorized into discrete classes (e.g., Cold, Warm, Hot) to improve interpretability and facilitate analysis.
- **Label Encoding**: Categorical features were converted into numerical representations. Unique identifiers like Driver ID and Constructor ID were transformed into encoded columns (e.g., Driver ID_Encoded) for seamless integration into machine learning models.
- **Scaling**: Continuous variables such as Total Points and Total Wins were standardized using the StandardScaler to ensure uniform scaling and eliminate bias from varying feature ranges.

Models

For this project, we used a combination of machine learning models to improve prediction accuracy. Here's how the modeling process worked:

- **Model Selection**: Multiple classifiers were chosen, including DecisionTreeClassifier, RandomForestClassifier, GradientBoostingClassifier, KNeighborsClassifier, and Support Vector Classifier. Each model has strengths in handling different types of data patterns.

- **DecisionTreeClassifier**

  A Decision Tree Classifier splits data into branches based on decision rules to predict outcomes. It’s intuitive and interpretable, making it ideal for understanding relationships in the data. Decision trees handle both numerical and categorical data well but are prone to overfitting, which can reduce accuracy in complex datasets. It works best as part of an ensemble for robust predictions.

- **RandomForestClassifier**

  The Random Forest Classifier builds multiple decision trees and averages their predictions, reducing overfitting and improving accuracy. It's robust to noisy data and handles large datasets effectively. Random Forest is excellent for feature importance analysis and performs well across various tasks, but its ensemble nature can make it less interpretable than single trees.

- **GradientBoostingClassifier**

  Gradient Boosting builds trees sequentially, each correcting errors from the previous one. It's powerful for handling non-linear relationships and provides excellent accuracy on structured data. This model is highly customizable with hyperparameters like learning rate and estimators. However, it can be slow to train and may overfit if not carefully tuned.

- **KNeighborsClassifier**

  The K-Nearest Neighbors Classifier predicts outcomes by looking at the closest data points in feature space. It’s simple and effective for small datasets with clear patterns. KNN is non-parametric and doesn’t assume a specific data distribution. However, it can be slow for large datasets and is sensitive to irrelevant features and data scaling.

- **Support Vector Classifier (SVC)**

  The Support Vector Classifier aims to find the best boundary (hyperplane) to separate data into classes. It's effective for high-dimensional and non-linear datasets using kernel functions. SVC provides robust performance when the data is well-separated but can struggle with noisy data and is computationally expensive for large datasets. It’s a strong performer in complex tasks.

- **VotingClassifier**: The models' predictions were combined using a voting ensemble.
  - **Hard Voting**: Majority voting determined the final prediction based on the most common result across models.
  - **Soft Voting**: Considered the probability of each prediction, resulting in a more nuanced decision.

- **Hyperparameter Tuning**: Key settings, like tree depth, number of estimators, and learning rate, were adjusted to optimize model performance.
- **Evaluation**: The ensemble model was trained on data from 2018 to 2022 and tested on 2023 results. This approach provided better accuracy and reliability compared to using individual models alone.

Results

Predictions: Predictions for the 2023 data were saved in ML_Outputs/classifier_predictions.csv

	Precision	**Recall**	**F1 Score**	**Support**
Top 3	0.76	0.73	0.74	174
Top 10	0.59	0.63	0.61	143
Other	0.64	0.61	0.62	64
accuracy			0.67	381
Macro Average	0.66	0.66	0.66	381
Weighted Average	0.68	0.67	0.67	381

Model Accuracy : 0.67

.

Feature Importance: iska graph

The model identified key features that significantly influenced predictions:

- **Driver Confidence**: Reflects a driver's reliability and consistency, calculated from historical performance and error probabilities.
- **Constructor Confidence**: Measures team reliability, factoring in pit stop efficiency and race-day performance.
- **Quali Position**: Starting grid positions derived from qualifying sessions, which strongly impact race-day outcomes.
- **Total Points**: Represents overall performance across races, highlighting drivers' and teams' season-long success.

Conclusions

Key Findings

- **Driver and Constructor Reliability**: Teams and drivers with higher reliability consistently perform better, making these strong predictors of race outcomes.
- **Weather Conditions**: Adverse weather significantly impacts race performance; teams with adaptive strategies tend to excel.
- **Pit Stop Efficiency**: Small improvements in pit stop times can positively influence race positions, emphasizing the importance of operational precision.

Practical Implications

- **For Teams**: These insights can support better strategic decisions for qualifying, pit stops, and race-day execution, particularly under challenging conditions.
- **For Broadcasters and Analysts**: Predictive analytics can make race coverage more engaging and informative for audiences by providing real-time insights.

Limitations

- **Data Imbalance**: Limited data for Top 3 positions may reduce prediction accuracy for this category.
- **Simplification of Variables**: Converting continuous variables into categories may lose detailed nuances.
- **Dynamic Factors**: The model does not account for real-time variables like driver health, on-track incidents, or mid-race adjustments.