Weather, Drivers, and Wins

A Machine Learning Perspective on Formula 1

Gourav Sharma - 301475592

Krish Bedi - 301563666

Ruben Dua – 301540990

Introduction

Formula 1 (F1) racing is a pinnacle of high-performance motorsport, where success is influenced by a combination of driver skill, team strategy, vehicle technology, and external factors such as weather. Predicting race outcomes in such a dynamic environment requires integrating diverse datasets and leveraging advanced machine learning techniques.

This project focuses on analyzing F1 race and weather data from 2018 to 2022 to predict race position categories (Top 3, Top 10, or Other) for the 2023 season. By collecting and processing data from online sources such as the Ergast API and FastF1, we incorporated key factors, including driver and constructor performance metrics, weather conditions, and pit stop efficiency.

The workflow involves scripting for data collection, preprocessing, and feature engineering, followed by the application of machine learning models. A Voting Classifier, combining the strengths of Decision Trees, Random Forests, Gradient Boosting, K-Nearest Neighbors, and Support Vector Classifiers, was implemented to enhance predictive accuracy.

This report provides an overview of the methods, tools, and outcomes, detailing the integration of various datasets, the preprocessing steps, and the results of the machine learning analysis. By simulating real-world conditions and integrating diverse performance metrics, this project aims to provide actionable insights into factors influencing F1 race outcomes and advance the understanding of predictive analytics in motorsport.

Data Collection and Sources

APIs Used:

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

Data Gathered

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

The 2024 F1 season, which is now coming to an end, features 24 races across four continents, exposing drivers and teams to a wide range of weather conditions. From scorching heat that accelerates tire wear to heavy rainfall reducing visibility and grip, weather significantly influences car performance, driver strategy, and race outcomes. Strategic preparation for these challenges is critical for success, making weather data an integral component of this analysis.

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. The graph highlights the thrilling points battle between Lewis Hamilton and Max Verstappen, along with other drivers' points tallies, during the controversial 2021 season's 22 races. It showcases their consistent dominance and closely contested performances, providing a clear view of performance trends and the intensity of the championship fight.

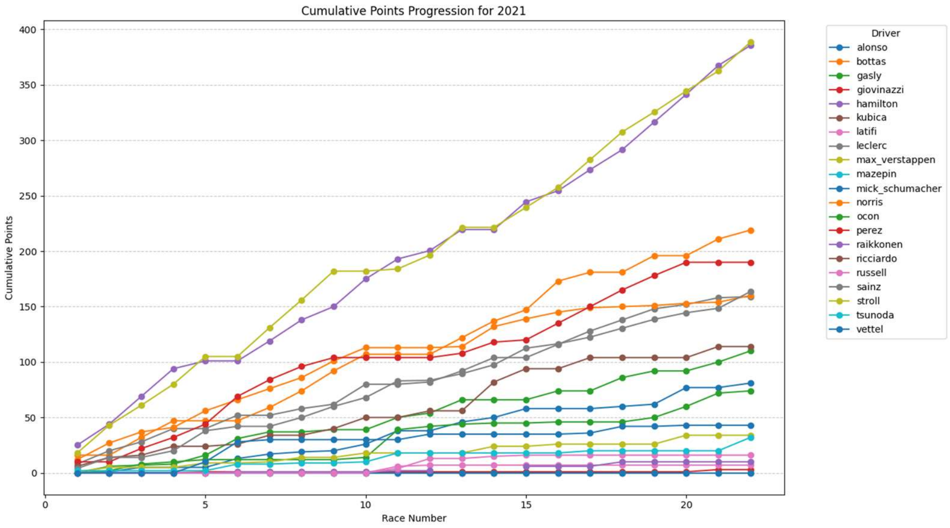


Fig 1: Cumulative Points Progression (Cumulative Points Vs Race Number)

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. We use the graph to demonstrate the importance of qualifying sessions and their correlation with race-day outcomes, showing how starting positions influence performance.

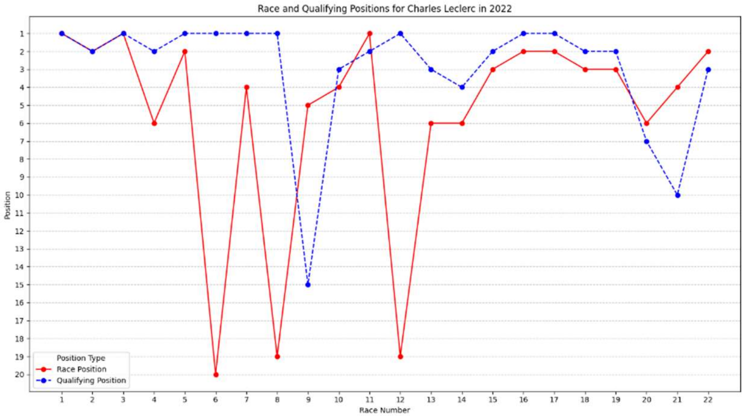


Fig 2: Qualification Position Data for Charles Leclerc in 2022

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.

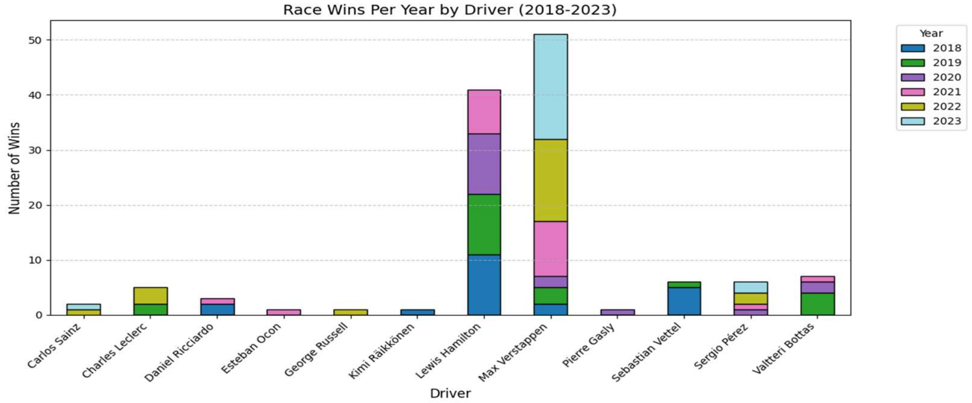


Fig 3: Race Wins Per Year By Driver (2018 - 2023)

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.

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 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.

Feature Engineering

Feature engineering was conducted to enhance the dataset's usability and improve model performance.

- **Experience Factor**:
  - Total points and wins for all drivers were calculated after every race. These cumulative metrics serve as an experience factor, highlighting their all-time performance and the competitiveness of drivers
- **Calculated Error Probabilities for Drivers and Teams**
  - Using incidents recorded during races, error probabilities were calculated and adjusted for both drivers and teams. These metrics were refined into two key features:

    - Driver Confidence: Represents the reliability and consistency of drivers by inversely reflecting their likelihood of errors based on historical incidents.
    - Constructor Confidence: Indicates team reliability using their race-day performances over the years.

Data Pre-Processing

To prepare the dataset for effective and efficient modeling, several transformations were applied:

- **Categorization**: Weather-related features such as AirTemp, Humidity, Rainfall, WindSpeed, and WindDirection were categorized into discrete classes (e.g., Cold, Warm, Hot). This improved interpretability and simplified analysis by grouping values into meaningful categories.
- **Label Encoding**: Categorical features were converted into numerical representations to integrate seamlessly with machine learning models. For example, unique identifiers like Driver ID and Constructor ID were transformed into encoded columns (e.g., Driver ID_Encoded).
- **Scaling**: Continuous variables like Total Points and Total Wins were standardized using the StandardScaler. This ensured uniform scaling across features and eliminated bias from varying feature ranges, enhancing the model's performance.

These preprocessing steps enhanced the dataset's quality, enabling robust feature engineering and accurate predictive modeling.

Machine Learning Models

For this project, we used a combination of machine learning models to improve prediction accuracy. Here's how the modelling 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**: Splits data into decision-based branches, offering high interpretability and effective handling of mixed data types. Works well as part of ensembles but prone to overfitting in complex datasets.
    - **RandomForestClassifier**: Combines multiple decision trees to reduce overfitting and boost accuracy. Excellent for feature importance analysis, robust on large datasets, but less interpretable due to its ensemble nature.
    - **GradientBoostingClassifier**: Sequentially builds trees to correct errors, ideal for capturing non-linear relationships. Offers strong accuracy on structured data but requires careful tuning to avoid overfitting and high training time.
    - **KNeighborsClassifier**: Classifies data based on proximity in feature space. Simple and effective for small datasets, though computationally intensive on larger datasets and sensitive to scaling and irrelevant features.
    - **Support Vector Classifier (SVC)**: Identifies the optimal hyperplane for class separation, excelling in high-dimensional, non-linear datasets. Robust in well-separated data but computationally expensive and less effective with noise.

- **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

The model achieved an overall accuracy of 67%, with strong performance in predicting Other finishes (Precision: 0.76, Recall: 0.73, F1-Score: 0.74), indicating confidence in identifying performances outside the Top 3 and Top 10. Predictions for Top 10 (F1-Score: 0.61) and Top 3 (F1-Score: 0.62) showed moderate accuracy, reflecting challenges in distinguishing high-performing drivers from mid-tier competitors.

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

Key Insights:

- **Strengths:** High precision for Other predictions (0.76) demonstrates confidence in identifying these outcomes, with balanced Macro Average (0.66) indicating fair accuracy across all classes.
- **Weaknesses:** Moderate performance for Top 10 and Top 3 predictions reflects challenges in feature representation, with low support for Top 3 (64 samples) contributing to variability.
- **Implications:** Reliable for Other predictions; improvements in features and addressing class imbalance are needed for better Top 3 and Top 10 accuracy.

This confusion matrix compares the predicted race position categories (Top 3, Top 10, Other) against the actual outcomes. The matrix provides insights into the model's performance by showing how well it predicted each category.

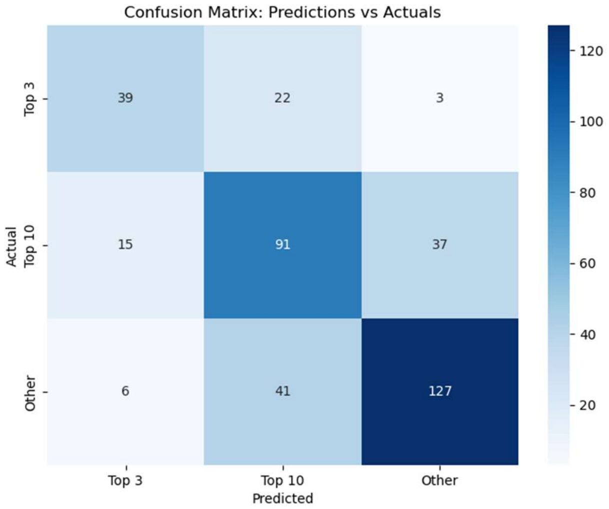
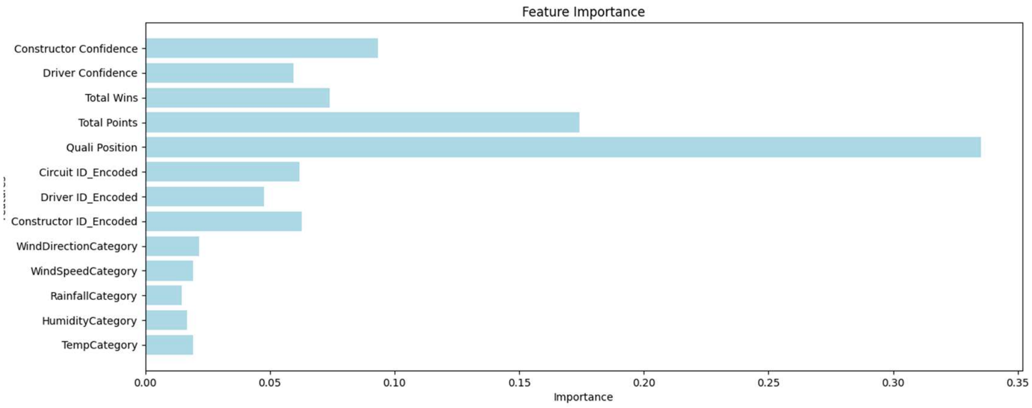


Fig 4: Prediction vs Actual

Key Observations:

- **Diagonal Values (Correct Predictions)**:
  - **Top 3**: The model correctly predicted 39 out of 64 actual Top 3 positions.
  - **Top 10**: The model accurately identified 91 out of 143 Top 10 positions.
  - **Other**: The model correctly predicted 127 out of 174 races in the Other category.
  - These diagonal values represent the true positives for each category.
- **Off-Diagonal Values (Misclassifications)**:
  - Some Top 3 races were misclassified as Top 10 (22 instances) or Other (3 instances).
  - Several Top 10 races were misclassified as Top 3 (15 instances) or Other (37 instances).
  - A small number of Other races were misclassified as Top 3 (6 instances) or Top 10 (41 instances).
- **Strengths**:
  - The model performed well in identifying the Other category, with the highest correct predictions (127), indicating it can reliably classify races where drivers finish outside the Top 10.
- **Weaknesses**:
  - There is significant confusion between Top 10 and Other categories, as seen from the 41 instances where Other was predicted instead of Top 10.

Feature Importance



The model identified several key features that significantly influenced predictions:

- **Constructor Confidence**: Measures team reliability, including pit stop efficiency.
- **Driver Confidence**: Reflects a driver's consistency based on past performance.
- **Quali Position**: Highlights the importance of starting grid positions in determining race outcomes.
- **Total Wins and Points**: Indicate season-long success and competitiveness.
- **Weather Factors**: Categorized wind, temperature, and rainfall capture environmental impacts on race performance. While these features were incorporated, their importance was lower than anticipated, suggesting that race outcomes are more influenced by driver and constructor factors than weather conditions.

These features played a critical role in enhancing the model's accuracy and interpretability.

Conclusion

This project successfully utilized machine learning to analyze Formula 1 race and weather data from 2018 to 2022, providing actionable predictions for the 2023 season. By integrating diverse datasets from sources like the Ergast API and FastF1, key performance metrics such as driver performance, constructor reliability, and weather conditions were incorporated into the analysis. The implementation of a Voting Classifier, leveraging the strengths of multiple machine learning models, demonstrated robust predictive capabilities.

Although driver and constructor performance emerged as the most influential factors, weather conditions, despite their real-world impact, were found to have a lesser effect in the model. This highlights the adaptability of F1 teams in mitigating environmental challenges. The project underscores the importance of data-driven insights in understanding race dynamics and offers a strong foundation for future advancements in predictive analytics within motorsport.

Practical Implications

- **For Teams**: These insights can support better strategic decisions for qualifying 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.

Problems Faced

- **Extraction of Weather Data** - We tried to use the GHCN extractor for fetching weather features for training the ML model, but the GHCN data was far too inconsistent in the availability of data during the race weekends which we tried to find using the closest station to the circuits using the Haversine formula.
- **Integration Challenges**:
  Combining multiple datasets from various sources, including Ergast API and FastF1, required significant effort to resolve inconsistencies in formats, timestamps, and data granularity. Handling missing or conflicting entries further added to the preprocessing workload.

References

- **Ergast API**: http://ergast.com/mrd/terms/
- **FastF1 API**: https://docs.fastf1.dev/api.html#fastf1.api.weather_data
- **Redbull**: https://www.redbull.com/th-th/how-weather-impacts-f1-racing
- **GHCN**: Global Historical Climatology Network daily (GHCNd) | National Centers for Environmental Information (NCEI)

Experience Summaries

Krish Bedi (301563666)

- Conducted in-depth data analysis of Formula 1 races to identify key factors influencing race outcomes, leveraging advanced statistical and machine learning techniques.
- Collected and integrated data from online APIs, ensuring consistency and quality across driver, weather, and team performance datasets.
- Enhanced machine learning model accuracy from 0.58 to 0.67 by hyper-tuning features and optimizing parameters for ensemble classifiers.
- Preprocessed and encoded datasets to ensure cleanliness and readiness for machine learning pipelines, enabling accurate and reliable predictions.

Ruben Dua (301540990)

- Analyzed multi-year Formula 1 datasets to explore the relationship between weather, driver confidence, constructor reliability, qualifying performance and race results.
- Collected data online through APIs, to ensure accurate and comprehensive datasets for analysis.
- Applied machine learning techniques, including ensemble methods, to predict race positions using factors like weather, qualifying positions, qualifying times and other metrics.
- Processed and integrated data on total wins and points for each driver as a metric for experience in machine learning model.

Gourav Sharma (301475592)

- Designed and implemented advanced machine learning models, including ensemble methods such as Random Forests and Gradient Boosting, to predict race outcomes based on weather, driver, and constructor performance metrics.
- Collected and integrated large-scale data online from APIs like Ergast and FastF1, ensuring robust quality through preprocessing techniques such as handling missing values and data normalization.
- Engineered key features, including driver and constructor confidence metrics, historical performance and statistical patterns to improve model accuracy and interpretability.
- Visualized data through comprehensive graphs and charts, showcasing feature importance and performance trends, providing actionable insights derived from predictive models.