Yellow Jacob

Green Raphaël

Purple : Luna

Red : Idriss

Hello Everyone, we are going to present our group project about Formula 1. We fist going to explain how formula 1 racing work and then explain our analysis. **Welcome to our presentation on 'RaceCraft Prodigy: F1 Winner Predictor**

Formula 1 involves high-speed races with teams designing and building their own advanced cars. Race weekends include practice, qualifying, and the race itself, with points awarded for finishing positions and additional factors.

Each team has two drivers, and strict regulations govern on-track behavior and technical specifications. Predicting race results is crucial for teams, utilizing data science to analyze car performance, driver skills, and strategies. The competitive nature of F1, combined with data-driven decision-making, enhances fan engagement and contributes to technological advancements in motorsport. There are multiple Grand Prix across the globe.

We begin by importing and exploring datasets, ensuring data integrity and identifying key variables. Descriptive statistics reveal correlations between manufacturer performance, grid/finish rankings, and driver nationality's impact on race outcomes.

The modeling phase involves data preparation, classification, and regression. Classification algorithms predict race outcomes categorically, while regression techniques forecast specific metrics like finishing positions.

The final step compares the strengths and weaknesses of both models. This systematic process aims to uncover patterns, enhance predictive accuracy, and provide valuable insights into Formula 1 race dynamics.

**2. Data Scraping and Cleaning (Luna)**

For our analysis, we extracted data from the Ergast API, spanning from 1990 to 2021. We've leveraged diverse datasets, including race details, race results, driver and constructor standings, qualifying results, and weather conditions. These datasets provide a comprehensive view of the multifaceted nature of Formula 1 racing.

**3. Exploratory Data Analysis: study of the link between our variables and victory**

"In our exploratory phase, we investigated correlations between various factors and race outcomes. We examined the influence of start grid positions, manufacturer performance, driver nationality, and weather conditions on race results. These insights were crucial in understanding the variables that significantly impact Formula 1 race dynamics."

**3.1 Correlation between Start Grid and Finish Ranking**

Formula 1 race data initially showed no clear correlation between start grid positions and podium standings across all years, likely due to dataset complexity. A focused exploration in 2020 revealed a subtle linear trend, leading to the creation of the corr\_grid\_podium function. Illustrated with 2019 data, the function systematically calculates and displays start grid–podium correlations for each circuit. The resulting bar plot provides insights into a notable relationship, with correlations exceeding 80% in certain circuits. What is interesting is that the Monaco circuit arrived very high in the correlation ranking and this circuit is very narrow and a lot of people say that this circuit is boring since it is too hard to pass in front a driver. Our number verify this claim.

**3.2 Correlation between manufacturer and grid and finish ranking**

Here is a summarized table showcasing the average podium and grid positions for each Formula 1 manufacturer, providing insights into their performance relative to starting positions and final podium outcomes.

Here again, without much surprise, some manufacturers stand out from the crowd, such as Ferrari, McLaren and Mercedes. We can see that people find Ferrari as the best team ever

**3.3 Correlation between driver and final victory**

We then turn to the link between nationality and victory. We're going to find out whether this variable can explain or predict final victory. We'll simply plot the percentage of victories by nationality. Unsurprisingly, the drivers with the most Grand Prix titles are German (Vettel, Schumacher, Rosberg), English (Hamilton), Brazilian (Senna) and Finnish (Räikkönen). There is a disparity between nationalities, which means that it is an important factor.

Moreover, there's a slight negative correlation between age and ranking (-0.055): in other words, the older you are, the lower your ranking and the closer you are to victory.

**3.4 Weather analysis**

We also conducted a weather analysis that would be available in the final report

**4.1 Data Preparation (Jacob)**

Data preparation is a critical phase in our project, laying the foundation for accurate and insightful modeling. It's where we transformed raw, diverse datasets into a clean, coherent format ready for analysis. Let's explore how we achieved this.

**Merging Datasets:**

Our first challenge was to integrate multiple datasets into a single, unified dataframe. We aligned data based on common attributes like race season, round, and driver IDs. We focused on maintaining the integrity of the data, ensuring that the merge process did not introduce errors or biases.

**Handling Missing Values:**

In dealing with missing data: We employed strategies like data imputation or, in some cases, exclusion of records with excessive missing values. Our approach was tailored to each variable's significance and the nature of its missingness. This step was essential to avoid skewed analyses.

**Normalization and Standardization:**

To ensure fair contribution from all variables, we normalized our data using techniques like StandardScaler and MinMaxScaler. This prevented variables with larger scales from disproportionately influencing the model's performance.

Normalization also helped in dealing with outliers, making our models more robust and less prone to overfitting.

**Categorical Variables:**

Our datasets contained numerous categorical fields like driver names, constructor teams, and weather conditions. For nominal data, we applied one-hot encoding, creating binary columns for each category. FOR EXAMPLE variables like weather conditions, where no inherent order exists. For ordinal data, with a natural order or ranking, we used label encoding for variables like grid positions, where the numerical order carries significance in the analysis.

**Creation of New Variables:**

We didn't just rely on existing variables. Recognizing the need for deeper insights, we created new variables that could offer more predictive power. For instance, calculating drivers' age provided a perspective on experience, while 'qualifying\_time' shed light on starting advantages. These new variables allowed us to capture subtler aspects of racing dynamics, enhancing the depth and accuracy of our analysis.

In essence, the data preparation phase was a meticulous process of transforming raw data into a structured and enriched dataset, ready for the sophisticated analyses in our modeling phase. This foundational work was key to the success of our project, ensuring that our models were built on a solid and comprehensive database.

**4. Modeling (Idriss)**

Training set : 1990-2020 : technology shift at this period not reliable to train the model before that

Test set : 2021

"For the modeling phase, we prepared our data meticulously, merging datasets, handling missing values, and creating new variables for a deeper analysis. We selected SVM for classification and linear regression for continuous prediction. Our models were rigorously trained and validated using cross-validation techniques to ensure robustness."

"Our predictive tasks were twofold: classifying race winners and predicting continuous rankings. To tackle these, we selected two types of models."

"For classification, we used a Support Vector Machine (SVM). SVM is renowned for its effectiveness in handling complex classification tasks, and its ability to manage high-dimensional data made it an ideal choice for our project."

"In terms of regression, we employed linear regression models. Linear regression is excellent for interpreting relationships between variables, making it suitable for understanding how different factors like driver skill or car performance translate into race rankings."

"Training these models was an iterative process. We experimented with different parameter settings, especially for the SVM, to identify the combination that yields the most accurate predictions. This was computationally intensive but crucial for fine-tuning our model's performance."

"In linear regression, we focused on understanding and quantifying the relationships within our data. This not only helped in predicting outcomes but also in gaining deeper insights into the factors that drive success in Formula 1."

For regression, we chose to use linear regression. This is a very standard model, easy to interpret and relatively quick to run.

First, we'll run the regression on all the explanatory variables in order to store their scores in our comparison dictionary.

Secondly, we'll regress on different combinations of variables in order to understand which types of variables best predict the race winner. We could use LASSO to this

We employed techniques like batch normalization and dropout for neural networks to improve convergence and reduce overfitting. We employed cross-validation techniques to assess the generalizability and robustness of our models.

**5. Model Evaluation**

For classification tasks, we used metrics such as accuracy, precision, recall, and F1-score. For regression tasks, we employed mean squared error (MSE) and R-squared. These metrics provided a comprehensive view of our models' performance, balancing the trade-off between different types of errors. We analyzed the performance of each model, focusing on their ability to accurately predict race outcomes. The insights gained from this evaluation informed our understanding of the key variables influencing Formula 1 race dynamics.

**6. Results**

On what we have seen yet in the SVM model gives better results than the many linear regression we have run.