Lap Time Analysis: A Measure of Rookie Viability in Formula 1

[redacted]

Western Governors University

Table of Contents

A. Project Highlights

Research Question or Organizational Need:

The capstone project addressed the question: "How does the progression of lap times for rookie drivers in Formula 1 during their first season correlate with their potential for long-term success in the sport?" This question aimed to establish a data-driven approach to assess rookie drivers' future success in Formula 1 based on their lap time progression in the first season.

Scope of the Project:

The project focused on analyzing historical and current performance data of rookie Formula 1 drivers. The scope included data collection, cleaning, feature engineering, model development, and validation. The project did not cover real-time data ingestion or in-season updates, and it was not deployed in a live racing environment.

Overview of the Solution:

The solution involved developing a predictive model using a Random Forest Regressor to assess the viability and success of rookie Formula 1 drivers based on their first-season performance. The model incorporated engineered features such as segmented normalized lap times and consistency scores. Python and its libraries (pandas, scikit-learn, and Matplotlib) were utilized for data manipulation, machine learning, and data visualization, respectively. The CRISP-DM methodology guided the project's execution.

Tools and Methodologies:

- **Data Analytics Tools and Techniques:** The project employed supervised machine learning models, including decision trees, random forests, and gradient boosting.
- **Data Source:** The dataset, sourced from Kaggle's "Formula 1 World Championship (1950 - 2023)" (Rao, 2023), comprised various F1 specific CSV files, including lap times, drivers, races, and results.
- **Programming Language:** Python was used for the project, and the code was housed in a Jupyter Notebook file.
- **Methodology:** The CRISP-DM process was adopted, encompassing stages from business understanding and data preparation to modeling and evaluation.

B. Project Execution

The project followed the initial plan outlined in the Task 2 document closely.

- **Project Plan:** Our main goal was crafting a model to predict the future success of rookie Formula 1 drivers. We did this by focusing on their lap times – specifically, how these times, when processed through our feature engineering (like segmented normalized lap times and consistency scores), could predict the points a rookie would earn, as points are ultimately the best guage of success for a driver.

- **Project Planning Methodology:** As discussed, we used the CRISP-DM methodology to guide the project, which was a perfect fit. It started with getting an understanding of the business need – predicting rookie drivers' success through the points they'd accumulate, which was known to hinge on lap time performance. This led to a comprehensive data understanding phase, where the significance of the points metric was deeply analyzed.

Subsequent phases of data preparation and modeling were rigorously executed, with the points metric being a focal point for feature engineering and model training. The evaluation phase validated the model's effectiveness in aligning with our business objectives, using metrics like accuracy, R^2 score, Mean Absolute Error, and Mean Squared Error, which were pivotal in assessing the model's ability to predict points accurately .

- **Project Timeline and Milestones:** The timeline set out in Task 2 was achieved. We began with data collection, focusing on lap times, then moved to feature engineering, which was pivotal in transforming the raw data into meaningful predictors. After implementing and validating the model, we finally had a model adept at using lap time features to predict rookie drivers' success in terms of points.

C. Data Collection Process

- **Data Selection and Collection:** Our choice of datasets from Kaggle was driven by the need to obtain comprehensive and relevant Formula 1 data. We selected the CSV files that were most pertinent to our research question, specifically: 'results.csv,' 'constructors.csv,' 'drivers.csv,' 'lap_times.csv,' and 'races.csv.' This selection was made to capture a broad range of data points – from lap times and race results to driver and constructor details – all crucial for analyzing rookie driver performance. The decision to structure these files within the 'F1_Dataset' folder and read them using pandas in Python was made for efficiency and to streamline our data processing workflow .

- **Handling Data Quality and Completeness:** The datasets required cleaning and preprocessing to ensure accuracy and reliability of our analysis. We specifically tackled missing values (marked as '\N') to avoid analysis errors. The 'positionText' attribute in the 'results' dataset was a unique challenge, as it contained mixed data types (integers and strings). We mapped non-integer values (like "R" for retired and "D" for disqualified) to numerical codes, enabling us to maintain consistency and usability in our machine learning model. The handling of the 'grid' attribute, particularly the '0' value indicating a start from the pit lane, was also critical to accurately reflect race start positions in our analysis. These steps were essential to ensure the quality and completeness of our data, making it suitable for robust analysis and model training .

- **Data Governance, Privacy, and Security:** Our adherence to data governance standards was extremely important. We ensured the integrity of the publicly available lap time records and statistics was preserved throughout the project. In terms of privacy, it was important that our data did not include any personally identifiable information; hence, we used aggregated metrics at the driver ID level. The minimal security concerns stemmed from the public nature of the dataset and its lack of sensitive information. Complying with ethical, legal, and regulatory standards was a fundamental aspect of our project, which is why we exclusively used open-source data and code libraries. This approach ensured that our project was conducted in an ethical and legally compliant manner, respecting the norms of data usage and analysis in public domain research .

In summary, every step of our data collection process was carefully considered and executed with the goals of our project in mind. From the selection of specific datasets to the detailed cleaning and preprocessing tasks, each decision was made to align with our objective of accurately predicting the potential success of rookie drivers based on lap time progression analysis.

C.1 Advantages and Limitations of Data Set

Advantages:

1. Comprehensive Metrics: The datasets from Kaggle provided a rich array of metrics crucial for our analysis. This included detailed lap times, driver and constructor information, and race results. Such comprehensive data was vital for an understanding of rookie drivers' performance.

2. Relevance to Research Question: The datasets were highly relevant to our research question, particularly the 'lap_times.csv' and 'results.csv.' These files allowed us to focus on the progression of lap times and relate them to the overall performance of rookie drivers, as measured by the points they scored during the season.

3. Suitability for Predictive Modeling: The datasets were well-suited for our chosen predictive model - a Random Forest Regressor. This model, as proposed in Task 2, was designed to analyze rookie driver performance based on various engineered features like consistency score, segmented normalized lap time, and categorization of laps into fast,

medium, or slow. The richness and variety of the data allowed us to engineer these features effectively .

Limitations:

1. Data Cleaning and Preprocessing Required: Although the datasets were rich in information, they required significant cleaning and preprocessing. This included handling missing values and transforming non-numeric data into a format suitable for our model. Such preprocessing was necessary to ensure data quality and reliability but did add additional steps to our analysis process.

2. Handling Mixed Data Types: The 'results' dataset, in particular, presented a challenge with the 'positionText' attribute, which included both integers and strings. This required careful mapping to numerical codes, adding complexity to our data preparation phase.

3. Focus on Engineered Features: While the datasets allowed us to develop engineered features essential for our analysis, this focus meant that we relied heavily on these transformations to derive meaningful insights. This reliance on engineered features meant that our model's effectiveness was directly tied to the accuracy and relevance of these features.

The datasets provided a strong foundation for our analysis, offering comprehensive and relevant data that aligned with our project's objectives. However, the need for extensive data cleaning and reliance on engineered features added layers of complexity to our data preparation and analysis

process. Despite these challenges, the datasets were instrumental in enabling us to develop a predictive model that could accurately assess rookie driver performance in Formula 1.

D. Data Extraction and Preparation

In this project, the data extraction and preparation process played a crucial role in ensuring the reliability and accuracy of our analysis. We employed a systematic approach to handle the complexities and nuances of our Formula 1 dataset.

1. **Data Extraction:** The data was sourced from the Kaggle dataset titled "Formula 1 World Championship (1950 - 2023)." We carefully selected specific CSV files relevant to our research question, namely 'results.csv', 'constructors.csv', 'drivers.csv', 'lap_times.csv', and 'races.csv'. These files were extracted into a dedicated 'F1_Dataset' folder, and we used Python's pandas library for efficient reading and handling of these datasets.
2. **Data Cleaning and Transformation:** The datasets, while rich in information, required extensive cleaning and preprocessing. We addressed missing values, labeled as '\N', and standardized the formats for analysis compatibility. A significant step was the conversion of mixed data types in the 'results' dataset, where attributes like 'positionText' contained both integers and strings. These values were mapped to numerical codes to maintain consistency across the dataset.
3. **Feature Engineering:** Given the project's focus on lap time progression and its correlation with points accumulation, we engineered several key features. This included segmented normalized lap times, consistency scores, and lap time grouping. These

features were designed to capture the intricate aspects of driver performance and were pivotal for the effectiveness of our Random Forest Regressor model.

4. **Tools and Techniques:** Python was our primary programming language, and we utilized pandas for data manipulation, scikit-learn for machine learning, and Matplotlib for data visualization. These tools were chosen for their versatility, robustness, and widespread usage in data science, making them ideal for our project requirements.

The data extraction and preparation stages were aligned with our project's objectives and were instrumental in building a robust foundation for our subsequent analysis.

E. Data Analysis Process

E.1 Data Analysis Methods

1. **Random Forest Regressor Model:** The core of the analysis was the random forest model that we created. This model was selected for its ability to handle the non-linear relationships in our complex dataset. The Random Forest model is known for its high predictive accuracy and robustness against overfitting, making it an excellent choice for our project, where we had to analyze a range of engineered features and intricate data.

2. **Data Cleaning and Preprocessing:** A significant part of our analysis involved engineering features from the raw lap time data. This included creating a consistency score, segmenting normalized lap times, and categorizing laps based on their speed. These features were essential in capturing the nuanced aspects of driver performance,

which were crucial for our model to accurately predict their potential for long-term success in Formula 1. The engineering of these features was informed by our observations on the quality and completeness of the dataset, where we handled missing values and mapped non-numeric data to numerical codes for better processing in our predictive model.

3. **Data Cleaning and Preprocessing**: Before applying the Random Forest Regressor, we undertook extensive data cleaning and preprocessing. This step was vital to ensure the reliability and accuracy of our model. We addressed missing values and standardized data formats, especially in the 'results' dataset, where we converted mixed data types into a uniform numerical format. These preprocessing steps made the data ready for analysis and were critical for the success of our model .

E.2 Advantages and Limitations of Tools and Techniques

1. **Advantages and Limitations of the Random Forest Regressor Model:**
    a. **Advantages:**
        i. **High Predictive Accuracy:** The Random Forest Regressor is renowned for its high predictive accuracy, which is crucial in a domain like Formula 1, where precision is key. This accuracy stems from the model's ability to handle complex, non-linear data and its robustness against overfitting .
        ii. **Interpretability:** Another significant advantage of the Random Forest Regressor is its interpretability. Unlike some more complex models,

Random Forest allows us to understand the influence of different features on the prediction, making it easier to draw actionable insights and develop tailored strategies for driver development and race planning .

b. Limitations:

i. **Computational Intensity:** While highly effective, Random Forest models can be computationally intensive, especially when dealing with large datasets. This might pose challenges in terms of processing time and resource allocation.

ii. **Model Complexity:** Despite its interpretability, the complexity of the Random Forest model, with multiple decision trees, can sometimes make it challenging to fully grasp the nuances of how certain predictions are made, especially for those without a deep understanding of machine learning.

2. **Advantages and Limitations of Engineered Features:**

a. **Advantages:** The engineered features like consistency score, segmented normalized lap time, and speed categorization were pivotal in capturing the nuanced performance aspects of rookie drivers. These features provided a deeper understanding of drivers' abilities and growth potential, enhancing the model's predictive power.

b. **Limitations:** Relying heavily on engineered features means that the effectiveness of the model is closely tied to the accuracy and relevance of these features. Any misinterpretation or error in feature engineering could potentially lead to less accurate predictions.

3. **Advantages and Limitations of Data Preparation Techniques:**
    a. **Advantages:** Our thorough data cleaning and preprocessing ensured that the model was trained on high-quality data, leading to more reliable predictions. This step was crucial in handling the diverse range of data types and formats found in our dataset.
    b. **Limitations:** The extensive data preparation required significant time and effort. In some cases, the transformation of data might also lead to the loss of certain nuances or information that could have been valuable for analysis.

E.3 Application of Analytical Methods

1. **Feature Engineering and Selection:** Initially, we engineered features that are crucial for our analysis, such as consistency scores, segmented normalized lap times, and categorization of laps into fast, medium, or slow. These features were derived from the raw lap time data, and their selection was based on their relevance to assessing rookie drivers' performances.
2. **Preprocessing and Data Cleaning:** Prior to model implementation, we executed extensive data cleaning and preprocessing. This included addressing missing values and standardizing data formats. Special attention was paid to the 'results' dataset, where mixed data types were converted into a uniform numerical format. This step was essential for ensuring data compatibility with our machine learning model.
3. **Model Implementation - Random Forest Regressor:** The core of our data analysis was the implementation of a Random Forest Regressor model. This model was selected for its

ability to handle complex, non-linear relationships present in our dataset. We trained this model on our engineered features, using Python's scikit-learn library.

4. **Model Validation and Testing:** Post-model training, we conducted thorough validation and testing. We used standard metrics such as accuracy, R^2 score, Mean Absolute Error (MAE), and Mean Squared Error (MSE) to evaluate the model's performance. These metrics were chosen as they collectively offer a comprehensive view of the model's accuracy and predictive power.
5. **Assumption and Requirement Verification:** Throughout the process, we continuously verified our assumptions and requirements. This included ensuring that our data cleaning methods did not distort the underlying data and that our model was not overfitting or underfitting. The Random Forest model inherently offers a feature importance metric, which we used to validate the relevance of our engineered features.

F. Data Analysis Results

F.1 Statistical Significance

The Random Forest Regressor model we developed is a supervised regression model. We evaluated its performance using key metrics: the R^2 score, Mean Absolute Error (MAE), and Mean Squared Error (MSE). According to our criteria set in Task 2, an R^2 score above 0.8 was the target for success.

Our model achieved this target, with an R^2 score surpassing 0.8. This indicates that the model explains a significant portion of variance in a rookie driver's performance. The low MAE and

MSE values also confirm the accuracy of our model in predicting future success of rookie drivers. Thus, the results support our hypothesis that rookie drivers showing consistent improvement in lap times are likely to succeed in Formula 1.

F.2 Practical Significance

The practical value of our model is in its application to real-world scenarios in Formula 1 teams. It serves as a tool for identifying rookie drivers who have a high potential for success, based on their lap time progression. This is particularly useful for teams in making decisions about driver selection and development strategies. By identifying drivers who consistently improve their lap times, teams can make more informed choices, optimizing their training and race planning. The model provides actionable insights for teams, helping them to focus their efforts on drivers who are more likely to achieve success.

F.3 Overall Success

The project is a success, not only in terms of meeting the statistical benchmarks set in Task 2, but also in its relevance to real-world Formula 1 racing scenarios. Our approach, from data collection and feature engineering to model development, was thorough and systematic. The end result is a predictive model that is both accurate and useful for Formula 1 teams. It shows the effectiveness of data-driven decision-making in sports and offers a valuable tool for teams to evaluate and strategize their driver selections and development.

G. Conclusion

G.1 Summary of Conclusions

In this project, we aimed to develop a predictive model that links the progression of rookie Formula 1 drivers' lap times with their end-of-season points tally. This focus is critical because both driver and constructor standings are directly tied to the points accumulated throughout the season. Recognizing that faster lap times often translate to higher points, we concentrated on engineering features from lap time data to explain this points variable, essential in determining a driver's success.

Our Random Forest Regressor model, evaluated using R^2 score, MAE, and MSE, successfully achieved this objective. The analysis substantiated our hypothesis: consistent lap time improvement across the season correlates with a rookie driver's increased point accumulation, indicating greater success in Formula 1. This finding, derived from extensive data collection, targeted feature engineering, and thorough model validation, highlights the significant impact of data analytics in assessing sports performance, especially in a high-stakes context like Formula 1 racing.

G.2 Effective Storytelling

During our project, we used graphical tools and descriptive markdown cells to communicate our findings clearly. The markdown cells provided insights into our analytical thought processes and conclusions after each significant step in the data analysis. Visually, we created a correlation matrix using graphs to identify the interrelationships among various engineered features. Additionally, we depicted the differences in lap speed categorizations

between rookie and non-rookie drivers, among other things. These visualizations, developed with Python's Matplotlib library, played a crucial role in making our complex data analysis accessible and engaging. They helped us transform intricate analytical results into a narrative that was both informative and easily understandable.

G.3 Recommended Courses of Action

Based on the insights obtained from our project, here are two recommended actions for Formula 1 teams:

1. **Integration of Data-Driven Strategies in Driver Evaluation**: We advise Formula 1 teams to incorporate the insights from our model into their driver evaluation and development strategies. Focusing on rookies whose lap time improvements correlate with higher points accumulation allows teams to identify promising talents more effectively and optimize their training and race strategies.
2. **Ongoing Refinement and Application of the Model**: For the model's continuous improvement, we suggest incorporating additional crucial data that was not included in the initial model, such as weather conditions during races. This data can provide deeper insights and potentially explain aspects of driver performance that the current model couldn't capture. Regular updates and refinements of the model with new data each season, including such variables, will ensure that the insights remain relevant and comprehensive. This adaptive approach will enable teams to evolve their strategies in line with the dynamic and multifaceted nature of Formula 1 racing.

H. Panopto Presentation

Panopto link: [redacted]

References

Rao, R. (2023). *Formula 1 World Championship (1950 - 2023).* Retrieved from Kaggle: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020