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

[redacted]

Western Governors University

Table of Contents

A. Proposal Overview

A.1 Research Question or Organizational Need

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?

A.2 Context and Background

Formula 1 is a highly competitive sport that requires drivers to have a wide range of skills, including speed, consistency, and racecraft. Teams invest heavily in scouting and developing talent, with the understanding that a proficient driver can be a game-changer. Traditionally, teams have evaluated rookie drivers based on subjective assessments of their driving style and performance. However, this approach can be prone to bias and may not accurately reflect a driver's true potential.

A data-driven approach to evaluating rookie drivers can provide a more objective and reliable method. By analyzing lap time progression via engineered features, teams can gain insights into a driver's adaptability, consistency, and potential for growth. This information can be used to make informed decisions about driver selection, development, and race strategy.

A.3 and A3A Summary of Published Works and Their Relation to the Project

Review of Work 1

The Motorsport.com article presents an evaluation of the first season performances of three Formula 1 rookies, offering insights into their progression, the challenges they faced, and

areas for improvement. The analysis emphasizes the importance of balance between aggression and consolidation in races, the translation of underlying pace into qualifying results, and closing performance gaps to more experienced teammates. These elements are fundamental to understanding the dynamics of rookie development in the high-pressure environment of Formula 1, where every fraction of a second and every strategic decision can have significant implications for a driver's career trajectory. (Straw, 2023)

The article details specific instances where the drivers' performances reflect their adaptability and potential for growth, paralleling the data-driven approach I've used for this project. By highlighting key areas for improvement, such as adaptability and consistency, the article supports the premise that a systematic analysis of performance and strategic progression is critical for predicting future success in Formula 1. Consistency is one of the features I've tried to account for in this project, having engineered it via a coefficient of variation ratio based on lap times. As such, this attention to detail that's described in the article to gauge performance is the same level of detail we're doing in this project.

Review of Work 2

In the world of racing, driver performance and the ability to improve over time are critical indicators of future success. This is exemplified in the article detailing NASCAR driver Sheldon Creed's career trajectory. Creed's initial challenges in the Xfinity Series, where he missed the playoffs and finished 14th in the championship standings, contrast sharply with his subsequent improvements. Notably, in his second year, Creed demonstrated significant progression, achieving a series of top-10 finishes. His crew chief was quoted in the article,

stating “If you look at his progression in stock-car racing in general, it took him a year to understand what he needed… When we move up to a series together, it’s not just him – it’s us together figuring out what we can do to make him more comfortable and give him what he wants. It’s a group effort for all of us. If history repeats itself, next year should be lights out for him.” (Albino, 2023)

This article’s story aligns with the research question of the capstone project. The article provides an example of how longitudinal performance data can be vital for strategizing and driver development in motorsports, a principle that is central to our objective. Therefore, Creed's experience in NASCAR, although a different series, offers a relevant parallel to Formula 1, validating the methodology of utilizing a data-driven approach to evaluate rookie drivers' potential.

Review of Work 3

The article from Joey Barnes on Motorsport.com focuses on the rigorous testing schedule of a 19-year-old rookie driver transitioning through multiple racing series, highlighting the meticulous planning by Chip Ganassi Racing to prepare him for a future in IndyCar. The piece underscores the importance of cross-series experience, noting the direct crossover from Formula 2 to IndyCar in terms of power-to-weight ratio and tire degradation. It also captures the driver’s personal growth, citing specific improvements in adaptability, feedback to engineers, and racecraft, bolstered by mentorship from more experienced teammates. In the article, Joey also gives a quote from the driver Simpson, stating the following: “I do think I've improved a lot over

the year," said Simpson, the 2021 Formula Regional Americas champion. "Improved lots of specific parts of my driving, just being more adaptable.**"** (Barnes, 2023)

This article relates to the capstone project as it emphasizes the significance of comprehensive testing and varied racing experiences in a driver's development, similar to the analysis of lap time progression for Formula 1 rookies. The driver's reflection on his own growth and the detailed observations from his manager provide qualitative insights that complement the quantitative analysis proposed in this project. It exemplifies the multifaceted approach to driver development, which is a key consideration for our project's objective of predicting rookie success. The meticulous preparation and diversified exposure are equal to the capstone's aim of using data analytics to evaluate rookie drivers, further supporting the hypothesis that a combination of consistent performance improvement and strategic development opportunities can be indicative of long-term success in motorsports.

A.4 Summary of Data Analytics Solution

The proposed data analytics solution will use a random forest regressor model to predict rookie driver performance based on progression in their first season via the following engineered features: a consistency score, segmented normalized lap time, and whether a lap qualifies as fast, medium, or slow.

The Random Forest Regressor model used in the study serves as the analytical engine, converting raw data into actionable insights. This model offers both high predictive accuracy and easy interpretability, making it a viable tool for real-world applications. The model will be

trained on a dataset of historical lap time data and engineered features for Formula 1 drivers. Once the model is trained, it can be used to predict viability of current rookie drivers. The predictions from the model can be used by teams to identify rookie drivers with the most potential and to make informed decisions about driver development and race strategy.

A.5 Benefits and Support of Decision-Making Process

The proposed process offers several benefits to Formula 1 teams:

- **Improved driver selection and development:**
    - By identifying rookie drivers with the most potential, teams can make better decisions about driver selection and development. This can lead to a more competitive driver lineup and improved team performance.
- **Assessment of driver development programs:**
    - When tracking the progress of rookie drivers over time, teams can assess the effectiveness of their driver development programs. This information can be used to make improvements to these programs and ensure that they are providing the best possible support to rookie drivers.
- **Tailored training plans and race strategies:**
    - By having an understanding of the strengths and weaknesses of each rookie driver, teams can develop training plans and race strategies that are tailored to each individual driver. This can help rookie drivers to reach their full potential and maximize their performance.

In addition to these benefits, we can also help to improve transparency and accountability in the driver selection and development process. By providing teams with a more objective and reliable

method for evaluating rookie drivers, this can help to reduce the risk of bias and ensure that the most talented drivers are given the opportunity to succeed.

B. Data Analytics Project Plan

B.1 Goals, Objectives, and Deliverables

- Goal 1: To develop a predictive model that can assess the long-term viability and success of rookie Formula 1 drivers based on their first-season performance.
    - Objective 1.1: Conduct thorough data collection, cleaning, and integration to create a comprehensive dataset.
        - Deliverable 1.1.1: A cleaned and integrated dataset comprising lap times and other relevant metrics.
        - Deliverable 1.1.2: Documentation detailing the cleaning methods and integration steps.
    - Objective 1.2: Engineer features that capture the nuanced aspects of driver performance.
        - Deliverable 1.2.1: Engineered features such as "Segmented Normalized Lap Times" and "Clustered Lap Times."
        - Deliverable 1.2.2: An explanation of the rationale behind each engineered feature.
    - Objective 1.3: Implement and validate a predictive model.
        - Deliverable 1.3.1: A trained Random Forest Regressor model.

B.2 Scope of Project

- Included in Project Scope:
    - Data collection
    - Data cleaning
    - Feature engineering
    - Model development
    - Model validation
- Not included in Project Scope:
    - Real-time data ingestion
    - In-season updates
    - Deployment of the predictive model in a live Formula 1 racing environment

B.3 Standard Methodology

The project will adopt the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. This methodology is well suited for projects that require a structured yet flexible approach, which is the aim of this project's process.

Standard Methodology Used: CRISP-DM

- **Business Understanding**: In this initial phase, a thorough understanding of the organizational need is developed. The primary focus is to predict the long-term viability of rookie Formula 1 drivers based on metrics extracted from their first-season performance to points obtained.

- **Data Understanding**: This phase is instrumental for comprehending the quality, structure, and limitations of the dataset. Exploratory data analysis is performed to identify key features and to understand the characteristics of the data.
- **Data Preparation**: Here, the raw data undergoes a series of cleaning and transformation steps. Missing values are handled, outliers are assessed, and the data is formatted to be compatible with subsequent analytical models.
- **Modeling**: Various machine learning algorithms and techniques are employed in this phase. The predictive model is constructed using features identified as significant during the Data Understanding phase.
- **Evaluation**: The final phase is dedicated to assessing the model's effectiveness in meeting the business objectives. Metrics such as accuracy, R^2 score, Mean Absolute Error, and Mean Squared Error are used to validate the model's predictions.

B.4 Timeline and Milestones

Milestone or deliverable	Duration	Projected start date	Anticipated end date
Data Collection	1 day	*11/01/2023*	*11/02/2023*
Data Cleaning & Integration	1 day	*11/02/2023*	*11/03/2023*
Feature Engineering	3 days	*11/03/2023*	*11/06/2023*
Model Implementation	1 week	*11/06/2023*	*11/13/2023*
Model Validation	1 week	*11/13/2023*	*11/20/2023*

B.5 Resources and Costs

- Kaggle dataset: Free
- Python: Free
- Python libraries: Free

- Jupyter Notebook: Free

B.6 Criteria for Success

- Successful collection and cleaning of data from multiple sources, resulting in a comprehensive dataset.
- Effective feature engineering that results in at least two new insightful features.
- A predictive model with an R-squared value above 0.8 on the validation set.

C. Design of Data Analytics Solution

C.1 Hypothesis

Rookie drivers who demonstrate consistent improvement in their lap times, as measured by various engineered features such as consistency scores and normalized times, are more likely to achieve greater success in Formula 1, measured by points obtained.

C.2 and C.2.A Analytical Method

The chosen analytical method for this project is the Random Forest Regressor model. This model will be trained on engineered features like "Segmented Normalized Lap Times", "Clustered Lap Times" and “Consistency Score” to predict metrics associated with long-term success for rookie drivers.

The choice of Random Forest is driven by several reasons. First, Formula 1 is a complex ecosystem where driver performance is influenced by a multitude of non-linear factors, from car mechanics to weather conditions, and Random Forest is more than capable of capturing such

complexities. Second, it offers an inherent feature selection mechanism, crucial for this project since several features have been engineered. Finally, its robustness to overfitting and ability to handle missing values make it a practical choice for real-world data.

C.3 Tools and Environments

The project uses Python as the programming language, with specific libraries such as pandas for data manipulation, scikit-learn for machine learning, and Matplotlib for data visualization. The project is done in a Jupyter Notebook file.

Note: no third party code has been used for this project.

C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance

- Model Evaluation:
    - Type of Model: The project employs a supervised regression model.
    - Algorithm(s) Used: The Random Forest Regressor algorithm has been implemented to develop the model.
    - Metric(s) Used to Assess Performance:
        - Mean Squared Error (MSE)
        - Mean Absolute Error (MAE)
        - R-Squared (R^2) Score
    - Benchmark for Success:
        - The model will be considered successful if the R-Squared score is higher than a baseline model, indicating the model explains a substantial portion of the variance in the outcome.

- A low MSE and MAE relative to the scale of the dataset's target variable would indicate high prediction accuracy.
- If the R-Squared score is close to 1, and the MSE and MAE are minimized, we can conclude that the model is capable of predicting a rookie’s future performance with high accuracy.

- **Justification for Chosen Methods and Metrics (C4A):**
  - Random Forest Regressor: This model is appropriate as it is capable of handling the non-linear relationships and interaction effects commonly present in complex datasets such as Formula 1 performance metrics.
  - Performance Metrics: The chosen metrics are standard for regression problems.
    - MSE is used to measure the average of the squares of the errors, i.e., the average squared difference between the estimated values and the actual value.
    - MAE measures the average magnitude of the errors in a set of predictions, without considering their direction.
    - R-Squared indicates the proportion of the variance for the dependent variable that's explained by the independent variables in the model.

These metrics were chosen because they collectively provide a comprehensive assessment of the model's predictive accuracy. The R-Squared score is particularly useful for understanding the model's explanatory power, while the MSE and MAE offer a clear picture of the model's prediction error.

The benchmark for success was set by considering the scale of the target variable (points, the main way to define success in F1), and the historical context of the dataset. The thresholds for MSE and MAE were determined based on the distribution of the target variable to ensure they are meaningful, while the R-Squared threshold was set to reflect a high level of variance explained by the model. These benchmarks ensure that the model's predictions are not only statistically significant but also practically relevant for F1 teams making decisions based on the model's output.

C.5 Practical Significance

The model's practical significance will be primarily evaluated through its prediction accuracy, mainly the Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 score. The lower the MSE and MAE, and the closer R^2 is to 1, the more useful the model is at predicting a rookie's future performance.

An F1 team could apply these insights by comparing the predicted performance metrics against a threshold value to make informed decisions on rookie driver selections. If the model's predictions surpass the threshold, it could serve as a green flag for teams to invest more resources into the rookie.

C.6 Visual Communication

In this project, the visualizations aim to offer an intuitive understanding of a rookie driver's performance in Formula 1. Specifically, histograms are utilized to plot segmented normalized lap times across various races for individual drivers, offering a temporal view of their performance trends. A correlation matrix of the engineered features is used to gauge how the

features relate to one another. Additionally, histograms are used once again to represent the consistency scores of these drivers, thus providing insights into their reliability on the track. These visualizations were created using Python's Matplotlib library, a widely-used tool for generating graphs. Together, these graphical representations facilitate an effective communication of the project's findings, aiding both the analysis and decision-making processes for stakeholders.

D. Description of Dataset

D.1 Source of Data

The dataset, comprised of 14 csv files, was downloaded from the publically available download on Kaggle, provided by Rohan Rao. (Rao) A link to the dataset will be provided in the references section.

D.2 Appropriateness of Dataset

The CSV files contain a range of features such as lap times, points, and driver details, which are essential for analyzing rookie driver performance in Formula 1, thus making it an appropriate dataset for this project.

D.3 Data Collection Methods

The dataset was downloaded from Kaggle and each csv was placed into a folder titled F1_Dataset, inside the root folder of the project. Only 5 of these csv's (listed below) will be needed for the analysis, and they're read into their unique dataframes in the Python code via panda's read_csv method.

CSV’s used:

1. results.csv
2. constructors.csv
3. drivers.csv
4. lap_times.csv
5. races.csv

D.4 Observations on Quality and Completeness of Data

The datasets used in this project are rich but did require some cleaning and preprocessing to be suitable for analysis. Specifically, missing values denoted as '\N' across various columns were removed or replaced, depending on the context in which they appeared.

In the 'results' dataset, particular attention was paid to the positionText attribute. This column contained not only integer values representing finishing positions but also character strings like "R" for retired, "D" for disqualified, and so on. These were mapped to numerical codes to make the data ready for machine learning algorithms.

The grid attribute, with a value of '0' indicating the driver started from the pit lane, was also treated as a special case and handled accordingly. With this meticulous data cleaning, we’ve ensured the quality and completeness of the dataset, making it robust for subsequent analytical steps.

D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

- **Data Governance:** The project relies on publicly available lap time records and statistics from Formula 1 races. The integrity of this data will be preserved throughout the project.
- **Privacy:** No personally identifiable information is included in the data. All metrics are aggregated at the driver ID level, mitigating privacy concerns.
- **Security:** Given that the dataset is publicly available and doesn't contain sensitive information, security concerns are minimal.
- **Ethical, Legal, and Regulatory Compliance:** All data and code libraries used are open-source and publicly available.

References

Albino, D. (2023, September 6). *Sheldon Creed sees improvement in second Xfinity season, sets goals for Cup debut*. Retrieved from NASCAR: https://www.nascar.com/news-media/2023/09/06/sheldon-creed-improvement-second-xfinity-season-cup-debut-kansas-speedway/

Barnes, J. (2023, October 26). *Ganassi puts IndyCar rookie Simpson through testing gauntlet*. Retrieved from motorsport.com: https://us.motorsport.com/indycar/news/ganassi-indycar-rookie-simpson-testing-plan/10538087/

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

Straw, E. (2023, April 12). *F1'S ROOKIE TRIO RANKED SO FAR – AND WHERE THEY MUST IMPROVE*. Retrieved from The Race: https://www.the-race.com/formula-1/f1s-rookie-trio-ranked-so-far-and-where-they-must-improve/