A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

From Data to Podium: A Machine Learning Model for Predicting Formula 1 Compound Decisions

Max Leischner

Work project carried out under the supervision of:

Michail Batikas

20-12-2023

Abstract

This thesis explores the optimization of Formula 1 pit stop strategies, integrating advanced analytics and machine learning to predict tire compound decisions. A novel aspect of this study is the clustering of driver profiles based on performance, tactical, and behavioral metrics, which provides a deeper understanding of driver characteristics and their impact on race strategy. By analyzing data from the FastF1 API and employing various machine learning techniques, we developed predictive models that not only forecast compound decisions with higher accuracy but also highlight the significance of personalized strategies tailored to different driver clusters. The findings demonstrate the potential of combining driver clustering with predictive analytics to refine pit stop strategies, offering teams a competitive edge through data-driven decision-making.

Keywords

Data, Data Analytics, Sports Analytics, Formula 1, Strategy, Prediction Model, Machine Learning

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Contents

Table of Figures

List of Tables

1. Introduction

In the competitive world of sports, the strategic use of analytics has revolutionised how competition is understood and won. This transformation is particularly evident in Formula 1, a sport where the smallest decisions can have significant impacts; leveraging data effectively is not merely an advantage but a cornerstone of success. Central to these strategies are the critical decisions made during pit stops, which can significantly influence race outcomes. In this context, the role of advanced analytics has become increasingly prominent, offering new avenues to optimize these crucial decisions.

While research has been conducted on various aspects of Formula 1 strategy, a notable gap exists in exploring driver clustering and its potential impact on strategic decisions as well as the evaluation of analytical approaches. This thesis is motivated by the prospect of harnessing advanced analytics and driver clustering to add additional insights to pit-stop strategies. The aim is to explore how a refined understanding of driver behaviour can enhance race outcomes and the interpretation of analytical outcomes when integrated with data-driven approaches.

This study is guided by the research question: "*How can the application of advanced analytics and driver clustering in Formula 1 enhance the accuracy and effectiveness of strategic decision-making, particularly in the context of optimising pit stop strategies?*" The objectives are twofold: firstly, to develop a robust clustering model based on performance, tactical, and behavioural metrics; and secondly, to use the results of this model within analytical frameworks, like prediction models, to assess its impact on compound selection decisions.

Our methodology centres around leveraging data mainly from the FastF1 API, focusing on recent seasons to construct a clustering model, providing a multifaceted perspective on driver profiles. The methodology extends to applying this clustering within different analytical models, aiming to shed light on the intricacies of pit-stop strategies and potentially add optimisations or additional insights.

This research aims to contribute significantly to the fields of sports analytics and Formula 1 strategic planning. By offering a novel perspective on driver behaviour and its implications for race strategy, the findings of this study could potentially guide teams and drivers towards more informed and effective decision-making. The insights gained could not only enrich academic discourse but also provide practical tools for strategic optimisation in the high-pressure environment of Formula 1 racing.

To lay the foundation for our analysis, we begin with a comprehensive literature review that covers the broader spectrum of sports analytics, with a specific focus on data analytics in Formula 1. This review will also discuss the application of clustering in other sports, noting its relatively nascent presence in Formula 1 literature. Following this, we detail our methodology, including our full data collection process and the clustering of Formula 1 drivers based on their performance metrics. In subsequent chapters, we examine tire compound decisions made during races and explore how these elements interact with the identified driver behaviour clusters.

2. Literature Review

2.1. Evolution and Scope of Sports Analytics

Sports analytics refers to the application of scientific techniques for investigating and modelling sports performance. It entails organising historical data in a structured manner, applying predictive analytical models to the data, and utilising information systems to inform decision-makers (Morgulev, Azar, and Lidor 2018). This practice enables sports organisations to secure a competitive advantage through enhanced player performance analysis, game strategy formulation, health and injury prevention, fan engagement, and operational strategy optimisation (Nadikattu

2020; Tan 2023). Originating in the 1960s with notational analysis in sports like American football and basketball (Hughes and MFranks 2004), sports analytics has evolved significantly with technological advancements and the rise of big data. Today, its application extends across various sports domains, from individual sports like tennis to team sports such as football and has become central in the data-driven world of motorsports like Formula 1. This wide adoption underscores the broad reach and applicability of sports analytics across diverse sporting disciplines (Bai and Bai 2021).

As technology evolved, sports analytics underwent a significant transformation, unveiling new avenues for analysing and improving sports performance. A recent comprehensive review by Ghosh et al. (2023) classifies the technological advancements in sports analytics into three primary research fields: sensors, computer vision, and wireless and mobile-based applications. These areas form the bedrock of modern sports analytics, providing essential methods to collect, analyse, and interpret data. These technological advancements prove particularly pertinent in Formula 1, a sport renowned for its data-driven approach. Incorporating hundreds of sensors in racing cars facilitates real-time data collection, covering various parameters such as speed, temperature, and throttle percentage (Shapiro 2023). Effective analysis of this data offers invaluable insights for making informed decisions on pit stop strategies, car setups, and race strategies. The inclusion of artificial intelligence (AI) and machine learning (ML) algorithms amplifies the potential of these technological advancements, enabling more sophisticated analysis and predictive modelling (Ghosh et al. 2023; Dindorf et al. 2023). Integrating these novel technologies with sports analytics methodologies has opened and revolutionised research opportunities in Formula 1 racing.

2.2. Analytical Approaches in Formula 1 Racing

As established in the preceding section, sports analytics is central to enhancing competitive strategies across various sports disciplines. This becomes particularly evident in Formula 1, a sport where even the minutest decision can significantly alter race outcomes. In F1, where the stakes are high and the financial implications are vast, the synergy between data analytics and elite sporting performance exemplifies the comprehensive capabilities of sports analytics. The body of literature directly engaging with sports analytics in the context of Formula 1 is limited both in terms of the quantity of research and the range of thematic focus. Broadening the search parameters to include 'Circuit Racing Motorsports'—thereby encompassing NASCAR, Formula E, and similar series—yields an expanded body of work. Preliminary examination suggests that while there has been increasing interest in this field, it appears that the field has not yet reached a point of saturation, indicating sufficient opportunity for further research and contribution.

The existing literature can be divided into several overarching categories, however, two are predominant: lap time simulations and race simulations. As Heilmeier (2018) highlights, it is crucial to differentiate between race simulations and the more prevalent lap time simulations. The latter predominates in the literature and typically focuses on the physical or engineering aspects rather than on the holistic view of an entire race. Siegler (2000) identifies three distinct approaches to lap time simulations: Steady State, Quasi-Static, and Transient. Heilmeier (2019) published a study on quasi-static lap time simulation, applying it to both Formula 1 and Formula E to illustrate its utility. Colunga (2014) examined the modelling of transient cornering and suspension dynamics, along with the investigation of control strategies for an ideal driver within a lap time simulation framework. In a similar vein, Timings (2014) contributed to this body of work by aiming to develop a robust lap time simulation, referring to its comprehensive nature and its resilience in varying conditions.

However, for this thesis, race simulations and their components, specifically those that prioritize pit stop strategies as a key component in modelling or predicting race outcomes, are of greater relevance. Such simulations are instrumental in forecasting final standings by accounting for various factors, including driver interactions, empirical fuel consumption models, tire wear, and probabilistic effects. Bekker (2009) developed one of the earlier holistic race simulations to replicate key on-track activities in Formula 1, such as mechanical failures, overtaking manoeuvres, and pit stops. This model facilitates strategy planning by simulating the mechanical and physical dynamics of a race, thereby offering a team a potential advantage. More recently, Heilmeier (2018) outlined a simulation methodology for circuit motorsport racing strategies, which considers variables such as pit stops, tire choices, and tire degradation. The tool is designed to rapidly simulate races based on discrete lap data and adjustable strategy inputs. Building on this previous work, Heilmeier (2020) introduces a new further improving the simulation. This advanced version surpasses simple optimisation models by providing a comprehensive, automated simulation that responds in real-time to race dynamics. Heilmeier (2020) suggests that the current methodology for optimising pit stop decisions and associated tire compound selections might benefit from additional exploration in the future. Furthermore, he acknowledges the omission of complex strategies, such as the undercut—a tactic where a driver pits and switches to faster tires to gain time on rivals who pit later—from current models. He advocates for the integration of such tactics into subsequent models to enrich the decision-making process.

In addition to comprehensive racing simulations, focused research activities are directed at optimising specific components of the racing domain and simulations themselves, such as pit stops. These efforts aim to refine these aspects to their utmost efficiency. One research exemplifies this by employing machine learning algorithms to aid tire strategy decisions in the NASCAR series.

This work utilises predictive analytics, employing historical race data to forecast positional shifts consequent to variables such as tire change frequency and tire lifespan. Extensive feature testing has revealed that support vector regression and LASSO regression yield the highest accuracy in results (Tulabandhula and Rudin 2014). Additionally, Bell (2016) advances the analysis by conducting a comprehensive analysis of performance determinants in Formula 1, examining the evolving contributions of team dynamics and driver skills over time. The study results in a systematic ranking of drivers, offering insights into the qualifications of the 'potential best' based on a quantifiable set of criteria. Furthermore, Monte Carlo methods and analysis of probabilistic factors play a significant role in simulating the inherent variability present in lap times, pit stops, race incidents, and potential degradation of vehicle parts. The study of Heilmeier (2020) builds upon this foundation, providing a comparative analysis of the seminal works of Bekker (2009), Phillips (2014), and Salminen (2019), thereby extending the understanding of these stochastic elements in race simulations.

The existent body of research on Formula 1 is mainly based on the scope of publicly available data, which has been limited to lap times, and race outcomes. Such data has predominantly been sourced from the Ergast API, a privately maintained database for Formula 1 statistics (Ergast 2009). However, the introduction of the Fast F1 API marks a significant progression in data availability, offering not only the information provided by the Ergast API but also a more comprehensive set of F1 data. This includes telemetry data, official weather statistics, and track information (FastF1 2020). The introduction of the Fast F1 API thus promises to broaden the scope of current literature and models by facilitating the incorporation of mechanical details of the vehicle (such as current gear, RPM, and speed) and more granular data like positional coordinates, distances between drivers, and time specific air and track temperatures. The newly available data points, particularly telemetry data, which have received little attention in the scientific literature to date, present potential opportunities to enhance and broaden current analytical frameworks. The richer and more granular nature of this data allows for a more detailed examination of specific drivers' behaviours based on positional and telemetry information. A prospective methodological approach might include clustering drivers according to their behavioural patterns in various racing scenarios. Such clustering could yield valuable insights into performance differentiators and decision-making processes throughout a race and its strategic decisions.

2.3. Clustering Techniques in Sports Analytics and Formula 1

Clustering, a core technique in unsupervised machine learning, groups data points into distinct categories based on common attributes without predefined labels (Pedregosa et al. 2011). In sports analytics, clustering is a key tool, enabling teams and coaches to decode complex patterns and detect subtle correlations. This method can help identify performance trends and effective team compositions, which, in turn, can be leveraged to improve predictive models that forecast future sports outcomes based on the grouping of player or play characteristics. Clustering may uncover non-intuitive groupings or strategies, providing a strategic advantage in the competitive world of professional sports. Its utility and adaptability across various sports disciplines are well-documented, with a significant body of literature.

In Basketball analytics, player assessment and categorisation have evolved well beyond the confines of traditional position labels. One study by Duman, Sennaroğlu, and Tuzkaya (2021) applies hierarchical cluster analysis to a rich dataset of game-related statistics from 15 NBA seasons, uncovering four to six distinct playing styles within each traditional position. The clusters, characterised by unique attributes and skill sets, offer a multifaceted perspective on player capabilities, yielding strategic insights for player placement and team composition. Muniz and Flamand (2022) introduce an advanced clustering technique based on weighted networks. The methodology starts with k-means clustering to form preliminary groupings, which then inform a network where players are interconnected by weighted edges reflecting performance similarities. Employing the Louvain method for community detection, the study identifies eight player archetypes, surpassing the insight offered by the five traditional positions. They further enhance this method by using tracking data, adding a layer of depth to the analysis with precise measurements of player movements and interactions. For Football, a fuzzy clustering model that can handle mixed data types has been introduced. This model assigns objective weights to attributes such as player performance metrics, positional data, and physical characteristics, thereby uncovering clusters that the complexity of mixed-attribute data might hide. Such detailed analysis facilitates the identification of player clusters according to their on-field roles, skill sets, and physical profiles, which is instrumental in tactical team structuring and player market valuation (D'Urso, De Giovanni, and Vitale 2023). A study in tennis clustered 1188 Grand Slam players by analysing their physical and play style data, revealing four distinct profiles. Through two-step cluster analysis and further MANOVA and discriminant analysis, the research identified how factors like height and handedness correlated with performance, notably in serving and net play (Cui et al. 2019). Clustering in sports analytics reaches beyond mainstream sports, extending its application to disciplines like badminton. Sinadia and Murwantara (2022) apply k-means and hierarchical agglomerative clustering to analyse badminton athletes' performances, identifying clusters based on game results and consecutive scoring. K-means clustering discerns four distinct performance clusters, with one particularly strong cluster associated with high scores and consecutive points, supported by similar results from hierarchical clustering.