

Datamining   
F1 forecasting model

29-10-2023

Version 1

Hewaar al koerdi 1619307

Table of Contents

[1. Introduction 2](#_Toc149501540)

[2. Data Acquisition and Preprocessing 3](#_Toc149501541)

[2.1 Business Understanding 3](#_Toc149501542)

[2.2 Data Understanding 3](#_Toc149501543)

[3. Exploratory Data Analysis (EDA) 4](#_Toc149501544)

[3.1 Data Overview 4](#_Toc149501545)

[3.2 Feature Engineering 4](#_Toc149501546)

[3.3 Initial Findings 4](#_Toc149501547)

[4. Issue Resolution and Project Adaptation 5](#_Toc149501548)

[4.1 Issue Analysis and Diagnosis 5](#_Toc149501549)

[4.2 Debugging and Troubleshooting 5](#_Toc149501550)

[4.3 Model Refinement 5](#_Toc149501551)

[4.4 Adaptation and Realignment 5](#_Toc149501552)

[4.5 Future Steps 5](#_Toc149501553)

[5. Self-Reflection and Future Prospects 6](#_Toc149501554)

[5.1 A Journey of Learning 6](#_Toc149501555)

[5.2 Collaborative Endeavors 6](#_Toc149501556)

[5.3 Overcoming Hurdles 6](#_Toc149501557)

[5.4 The Exploration of FastF1 6](#_Toc149501558)

[5.5 Conclusion 6](#_Toc149501559)

[References 7](#_Toc149501560)

# 1. Introduction

Data mining has become an indispensable tool in modern data analysis, providing the ability to extract valuable insights and predictions from vast datasets. In the realm of sports, data mining techniques have found application in diverse areas, from player performance analysis to game outcome prediction. This report delves into the fascinating world of Formula 1 (F1) racing, where we harness the power of data mining to develop an F1 prediction model using Python.

As an enthusiastic fan of Formula 1, I embarked on this data mining journey with a specific goal in mind - to create an F1 prediction model that could provide valuable insights into the outcome of F1 races. To achieve this, I leveraged various tools and resources at my disposal, including the powerful ChatGPT, and combined them with historical F1 race data.

The primary objective of this project is to utilize machine learning techniques to provide a preliminary forecast before each race weekend. By analyzing historical data encompassing various race factors, such as circuit characteristics, team performance, and driver statistics, we aim to make accurate predictions about the winners and other key outcomes.

In this report, I will outline the process I followed to construct this F1 prediction model, including the data collection, preprocessing, feature engineering, and the machine learning algorithms employed. Moreover, I will discuss the significance of Formula 1 as a subject of interest for data mining, and how my passion for the sport motivated me to choose this specific domain for analysis.

Formula 1, known for its cutting-edge technology, high-speed races, and the fierce competition between world-class drivers and teams, serves as the ideal backdrop to explore the capabilities of data mining. This project encapsulates not only the excitement of F1 racing but also the power of data mining to unravel hidden patterns and insights from a treasure trove of historical data.

As we navigate through the intricacies of developing an F1 prediction model, this report will shed light on the methods, tools, and techniques employed to harness the past and make informed predictions for the future in the dynamic world of Formula 1.

Join me on this journey into the world of data mining in F1, where we aim to blend the thrill of racing with the precision of data analysis.

# 2. Data Acquisition and Preprocessing

In the realm of Formula 1, where split-second decisions can make or break a race, data mining emerges as a powerful ally. This chapter sets the stage for our journey into the world of data-driven insights and predictions within the fast-paced arena of Formula 1.

## 2.1 Business Understanding

Our journey into data mining begins with a clear understanding of the business objectives we aim to achieve. The ultimate goal is to create a predictive model capable of providing insights and forecasts before each Formula 1 race weekend. This model, if successful, could prove invaluable for fans and stakeholders alike. By anticipating race outcomes, we aim to add a new layer of excitement to the F1 experience and provide data-driven insights for teams and enthusiasts.

## 2.2 Data Understanding

With our business objectives in focus, we transition to the data understanding phase, where we delve into the extensive Formula 1 dataset sourced from Kaggle. This dataset encapsulates the journey of Formula 1 from its inception to the midpoint of the 2023 season. Before any analysis can begin, a thorough understanding of the data is essential. (Kaggle , sd)

To facilitate this understanding, we loaded the 11 CSV files into dataframes, structured data containers, allowing for easier manipulation and analysis. This transformation enhances our ability to interact with the data efficiently. However, the process doesn't stop here. To ensure the quality and reliability of our data, we embarked on the vital journey of data cleaning. Missing values, often a hurdle in data analysis, were addressed diligently, either through imputation or removal, ensuring the dataset remains coherent and ready for the rigorous demands of predictive modeling.

In these initial phases, we align with the first two stages of the CRISP-DM (Cross-Industry Standard Process for Data Mining) model, focusing on business understanding and data understanding. These preparatory steps are essential in laying a robust foundation for the data mining adventure that awaits.

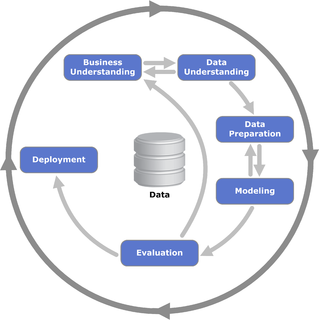


Figure 1: CRISP-Model

# 3. Exploratory Data Analysis (EDA)

In this chapter, we delve into the Exploratory Data Analysis (EDA) phase, bridging the gap between data preprocessing and predictive modeling. By dissecting the historical Formula 1 dataset sourced from 11 CSV files, including circuits, constructor results, constructor standings, constructors, driver standings, drivers, lap times, pit stops, qualifying, races, and results, we uncover vital insights to guide our journey.

## 3.1 Data Overview

Our exploration begins with a comprehensive overview of the dataset. We have assimilated 11 distinct CSV files, each contributing a unique facet of historical F1 data. These files chronicle the evolution of the sport from its inception to the midpoint of the 2023 season, capturing a diverse range of attributes and attributes. Understanding the dataset's structure and dimensionality is the cornerstone of our analysis.

## 3.2 Feature Engineering

Within our dataset, the "results" data stands as a crucial feature in our quest for predictive accuracy. By engineering features from this dataset, we aim to derive insights that directly contribute to our ability to forecast F1 race outcomes.

## 3.3 Initial Findings

Having traversed these analytical realms, we emerge with our initial findings. The data showcases an intricate tapestry of F1 history, from circuit specifics to driver achievements. These discoveries, rooted in EDA, provide the bedrock for our predictive modeling efforts, steering us toward our ultimate goal of forecasting F1 race outcomes.

# 4. Issue Resolution and Project Adaptation

In this chapter, we focus on addressing a significant setback that emerged during the implementation of our regression model. The project's progression was temporarily stalled due to unforeseen issues, and our journey towards creating a reliable F1 prediction model demanded adaptation and resolution. This chapter outlines our experience in facing and managing these challenges.

## 4.1 Issue Analysis and Diagnosis

The journey took an unexpected turn as we confronted a challenging error message:

ValueError: Found input variables with inconsistent numbers of samples: [3844, 26080]

This error revealed a fundamental inconsistency in the data samples, indicating a critical issue that needed immediate attention. The data used for training and testing appeared to be misaligned, with different numbers of samples. This inconsistency posed a significant obstacle in our quest for a functional regression model.

## 4.2 Debugging and Troubleshooting

In the face of this error, extensive debugging and troubleshooting efforts ensued. Code reviews, log analysis, and consultation with fellow data scientists and domain experts were part of the quest for a solution. Despite our persistent efforts, the root cause of the inconsistency remained elusive.

## 4.3 Model Refinement

In an attempt to circumvent the issues encountered, adjustments were made to the regression model. This included refining hyperparameters, revisiting data preprocessing steps, and meticulously examining the model architecture. Furthermore, the ratio of the training and testing data was altered to better suit the dataset's reduced size. Unfortunately, these changes did not yield the desired outcome, and the inconsistent samples error persisted.

## 4.4 Adaptation and Realignment

Given the persistent nature of the issue and the constraints imposed by time, a decision was made to adapt the project's direction. It became evident that the initial approach, which aimed for a traditional 80/20 train-test split, was not suitable for the dataset's reduced size and the unique challenges posed by the Formula 1 data. A strategic realignment of project goals and methodology became imperative.

## 4.5 Future Steps

The path ahead is not without its challenges, but our resolve to create a reliable F1 prediction model remains unwavering. The next steps in our project will involve revisiting the fundamental data preprocessing steps, reassessing the choice of regression model, and exploring alternative methodologies that better suit the characteristics of the dataset. Hyperparameter tuning and a more exhaustive model evaluation will also be crucial components of this phase.

By addressing these issues and adapting our approach, we aim to steer the project back on course. While challenges are an inherent part of the data science journey, they offer valuable opportunities for learning and growth. We remain committed to the realization of our F1 race prediction objectives and are determined to overcome any obstacles that come our way.

# 5. Self-Reflection and Future Prospects

In this concluding chapter, we take a moment to reflect on the valuable lessons learned and the path forward for this project. Our journey through the world of Python and data modeling has been transformative, marked by challenges, learning, and the support of others.

## 5.1 A Journey of Learning

The journey embarked upon for this project has been one of profound learning. Python, with its versatility and power, has revealed its capabilities through hands-on experience. The ability to code, manipulate data, and create predictive models has grown significantly, and the project served as a platform for honing these skills. One noteworthy influence in this journey was the guidance of Tijmen Weber in the "Introduction to Modeling" course. The insights gained from these lessons proved invaluable in the pursuit of creating a robust F1 prediction model.

## 5.2 Collaborative Endeavors

Collaboration was a cornerstone of this project. Seeking guidance and mentorship from individuals well-versed in data science was pivotal. A significant contribution came from my sister, an accomplished bioinformatician. Her expertise extended to assisting in the utilization of Visual Studio Code (VSCode) and outlining the necessary steps for this project. While the allure of VSCode's interface was compelling, the familiarity and ease of use in Spyder ultimately brought me back. Despite the seamless functionality of VSCode, Spyder's comfort and familiarity have become indispensable in my workflow.

## 5.3 Overcoming Hurdles

The looming project deadline did not deter our commitment to completing this endeavor. The prospect of delivering a functional F1 prediction model remained the goal, and time constraints only served to intensify the project's pursuit. The completion target has been extended to align with the 2024 Formula 1 season, granting us a more generous timeframe for refinement and thorough validation.

## 5.4 The Exploration of FastF1

An intriguing development during the project was the discovery of the Python library FastF1. This library promised access to a wealth of additional data sources, potentially enhancing the model's predictive capabilities. However, its reliance on APIs introduced a new challenge, as working with them required additional proficiency. The inherent volatility of APIs, which FastF1 acknowledges, raised concerns about the stability of the code. The potential risks and rewards associated with adopting FastF1 remain under consideration and warrant further exploration. (FastF1, sd)

## 5.5 Conclusion

In conclusion, this project has been a journey characterized by learning, adaptation, and perseverance. The technical skills developed, lessons gleaned from mentors, and the ongoing commitment to the project's objectives collectively define this undertaking. As we continue to work toward the creation of a reliable F1 prediction model, the horizon is marked by new challenges and opportunities for growth. The resilience to overcome hurdles and the pursuit of data science excellence remain at the forefront of our journey, and the future holds the promise of exciting developments and insights in the world of Formula 1 prediction.

# References

*FastF1*. (sd). Opgehaald van https://docs.fastf1.dev/

*Kaggle* . (sd). Opgehaald van https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020