Abstract

With the growing spectator base of Formula 1, a significant opportunity exists to develop predictive models that enhance both betting activities and in-depth analysis within the sport. This paper aims to create a model for accurately predicting the final positions of the top 20 drivers in each race. The approach involves adapting a rating system model, previously employed in football, to estimate drivers' strengths and predict outcomes based on that. This exploits the gap in having single algorithms useful for different sports. The study's results demonstrate the effectiveness of this approach, surpassing alternative models that employed supervised machine learning techniques. In contributing to the field of sports prediction, this research specifically addresses sports with multi-class outcomes beyond the simple win/loss/draw scenarios.

Contents

CHAPTER 1

Introduction

Sports result prediction poses a fascinating and challenging problem due to the inherent unpredictability of sports and the multitude of factors influencing outcomes [4]. The crux of this challenge lies in the uncertainty that defines sports prediction. The ability to forecast results and extract valuable insights appeals not only to sports professionals but also to a broader audience. Accurate prediction supports informed decision-making for team management, thereby enhancing the prospects of winning leagues and tournaments. This appeal extends to sports enthusiasts, particularly those engaged in sports betting. Implicit in this endeavor is the utilization of available science, data, and statistical analysis techniques, including result prediction, player performance assessment, player injury prediction, sports talent identification, and game strategy evaluation [11]. The objective of this research is to adapt and develop a predictive model for Formula 1 race outcomes—a multi-outcome sport—based on proven algorithms designed for win/loss/draw outcomes.

Horvat and Job [8] conducted a review about machine learning in sport prediction, analyzing over 100 papers from 1996 to 2020. Despite neural networks being the most commonly used machine learning (ML) model for game outcome prediction, they don't always yield the highest accuracy, as revealed in the research. ML algorithms aim to predict observed process outcomes and build decision-making models. Result comparison is challenging due to different datasets and league competitiveness. Low-scoring or highly uncertain sports tend to provide lower accuracy. The authors identified a challenge: the need for a unified method predicting outcomes across multiple sports. This raises the question: can models be adapted to different sports? The authors suggest that employing alternative ML algorithms can yield positive prediction outcomes, and in some cases, even superior results—a statement particularly relevant to the present model objective.

A model highlighted in [9], boasting an impressive 93% accuracy, underscores the significance of the Player Performance Index (PPI), assigning it a higher weight compared to other features. This emphasis on PPI greatly influences the present model. In another notable review, [4] advocates for a comprehensive approach to model comparison, with a focus on the impact of feature selection on predictive models. Their analysis reveals that in almost 90% of surveyed studies, researchers emphasize the importance of incorporating advanced features through engineering. Furthermore, the review emphasizes the evolution of new model features, such as performance indicators (PIs), often extending beyond the sports domain to encompass external elements.

In Formula 1, a sport marked by recent technological breakthroughs, each car, equipped with 120 sensors, generates 3 GB of data per race, producing 1500 data points per second [12]. Unlike many other sports, predicting outcomes in Formula 1 faces unique challenges such as penalties, accidents, mechanical failures, and in-season technology advancements. Data scientists tackle these challenges using deep-learning models trained on a 65-year dataset, extracting race statistics, generating forecasts, and offering insights into split-second decisions and strategies employed by teams and drivers.

Despite Formula 1's widespread recognition, there are limited publications on result predictions. While aspects like Tyre Strategy Prediction [16] and analysis of key factors influencing overall points [15] have been explored, few focus on predicting the race outcome from $20! \approx 2.4 \times 10^{18}$ possible scenarios. Predicting Formula 1 results involves two crucial factors: race performance and car competitiveness. The latter, challenging due to frequent motor updates, introduces a research gap. Constructors' constant development renders the car almost brand new, with no prior recorded data. Coupled with an average of around three incidents per race, some predictions become inadequate. Yet, this shouldn't discourage mathematicians and statisticians from designing more accurate models; the broad scope for improvement should be embraced.

In the literature, [19] conducted a study investigating supervised learning techniques for predicting the championship standings of the 2021 Formula 1 season based on historical data, achieving an accuracy of approximately 35% for predicting the top 10±1 results. Contributing to the discourse on Formula 1, another paper by [7] compared three neural network models, resulting in accuracies of 17%, 55%, and 58%. Finally, [20] also explores neural networks but concludes that they cannot definitively assert that it is the best method for race prediction, suggesting the consideration of the car brand as an additional factor.

Regarding the Elo Algorithm, [6] asserts that the model surpasses existing systems in prediction accuracy, and [1] confirms that no other algorithm based solely on past results could be expected to outperform Elo. Analyzing the algorithm's application in Formula 1, solutions are presented by [17] and [14]. The former anticipates issues when considering competitions over longer periods, where competitors' abilities change substantially, while the latter acknowledges the need for a long journey to improve accuracy. To address unexpected results, [18] suggests improving the Elo Algorithm. Additionally, [10] discusses widening the Elo algorithm by including more factors.

This study aims to develop a precise predictive model for Formula 1, predicting the final grid of 20 places for each race. The inspiration for this endeavor comes from the success of the *Soccer Power Index (SPI)* [2] in the 2022 World Cup, adapted for tennis predictions as well [13]. The predictions are based on the Elo rating system, initially created by Arpad Elo to estimate the strength of a team or player in a game. It hopes to solve the issues encountered before by adapting an existing algorithm, addressing the question raised at the beginning of this section: Can some models be adapted to different sports? The evidence from other sports suggests this might be advantageous.

This sets three clear objectives: first, to derive a comparable *Formula 1 Power Index (PI)*, as detailed in the methodology, serving as the foundation for the model; second, to assign appropriate weights to significant variables in the computation of PI; and lastly, the third objective is to assess the accuracy of the model and evaluate its performance. This study aims to contribute to the sports prediction body of knowledge by designing a novel model for Formula 1 race results.

The goal is to successfully adapt an existing predictive algorithm, demonstrating that certain algorithms can be implemented across multiple sports, encompassing both single-class outcomes and multiple ones. Focusing on Formula 1 is particularly opportune due to the sport's expanding viewership, with a recorded 36% increase in attendance in 2022 compared to 2019 [3]. This research addresses the current shortage of studies on predicting results in Formula 1 and opens a new scope of investigation for outcomes beyond the traditional single-class (win/loss/draw) scenario.

The innovative aspect of this work is the approach taken towards predictions. It's not only an adaptation of the model employed by [2], which was inspired by the mentioned Elo Ratings, but it goes further by considering the probabilities of winning or losing positions, taking an additional step in the algorithm. This way, it directly contributes to the set of algorithm adaptations for different sports with multiple outcomes.

CHAPTER 2

Dataset Exploration

The data for this project comes from the R package, `f1dataR` [5]. Table 2.1 represents the top three and bottom three lines from the dataset with which the model has been developed. The dataset consists of 880 rows of information spanning 10 different columns, centered around the 2021 and 2022 Formula 1 seasons. The data from 2021 just served as the basis for parameter tuning, details will be explained later. And the 2022 dataset was dedicated to in-depth analysis and testing.

The decision to employ an entire season for training and testing stems from the fact that teams use the off-competition periods to advance their technologies and fine-tune their machinery. This rationale supports the need for just one season's data to develop the model effectively. Moreover, as elaborated in the next chapter, the model is self-contained and adaptable to incorporate updates from every season. At the start of a new season, the model is reset, and it continuously adapts in real-time alongside the teams and drivers as they learn and gather new information about the impact of seasonal updates and changes on race outcomes.

The 20 drivers from the 2022 are represented by *driver_id*. Secondly, the dataset is composed of the constructor, *constructor_id*, associated to the driver.

Following, four factor levels in the dataset are found. The first, *round* factor, denotes the 22 different races that took place in the 2022 season. The second factor, *season*, distinguishes the 2021 from the 2022 season. The third, *grid*, is a value from 1 to 20 that states the position on which the driver starts the race. This is a crucial piece of information because it can highly impact the race, being the defining number of places that the driver would need to overtake

driver_id	constructor_id	round	season	grid	position	status	pos_gained	round_points	accum_points	PI	PI_adj
<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
hamilton	mercedes	1	2021	2	1	Finished	1	24.26	24.26	13.26	14.00
max_verstappen	red_bull	1	2021	1	2	Finished	-1	21.27	21.27	11.68	13.16
bottas	mercedes	1	2021	3	3	Finished	0	18.65	18.65	10.19	19
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
hamilton	mercedes	22	2022	5	18	Hydraulics	-13	2.59	320.69	8.03	8.03
latifi	williams	22	2022	20	19	Collision damage	1	2.27	76.00	1.91	1.91
alonso	alpine	22	2022	10	20	Water leak	-10	1.99	178.91	4.48	4.48

Table 2.1: Formula 1 dataset from the 2021 and 2022 seasons.

in order to win. The fourth and final factor is the *position*, once again being a value in the range from 1 to 20, this number indicates the result of the race. There is an average of three incidents per race, meaning that usually, the last three places do not refer to official results. Therefore, the position needs to be taken into consideration together with the status variable.

The *status* variable explains the condition in which the drivers finalized the race. The focus is around two main concerns: whether the driver successfully finished the race, which is denoted by three specific conditions: *Finished*, *+1 Lap*, *+2 Laps*, or if an incident was presented, for which there are many different types of outcomes including, inter alia, *Collision damage*, *Engine* and *Hydraulics*.

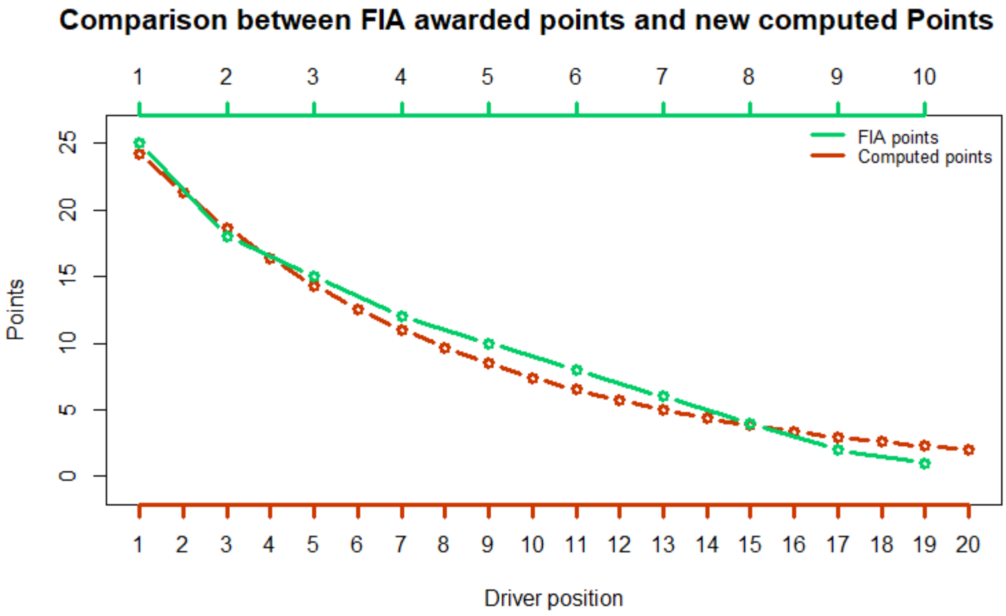
Positions gained, *pos_gained*, refers to the variable computed as *position - grid*. The value ranges from -19 to 19 and states the final number of places the driver gained (or in some cases, lost) at the end of the race. The measure is relevant for the model because it will be taken into account as the mean of total number of overtakes that the driver executes (or receives). It is an average that is considered to calculate the probability that, in a future career, the driver advances or regresses the position in which is starting.

The points awarded according to the race result are denoted by the variable *round_points*. Since the official awarded points are just for the top ten pilots, new points were computed. Inspired in a similar distributed form, the values were extracted from an exponential decreasing function shown by plot 2.1 together with the comparison of official awarded points. Let $\mathbf{p} \in \mathbf{R}^{20}$ represent the vector corresponding to the points[1] such that for $i = 1, 2, \ldots, 20$, the relation $p_i \propto f_{gamma}(x; \alpha, \beta)$ holds, where α represents the shape parameter and β represents the rate parameter of the gamma distribution function denoted as $f_{gamma}(x; \alpha, \beta)$. Fixing $p_i = 40 \times f_{gamma}(x; \alpha = 1, \beta = 1)$, the used distribution is defined. The *accum_points* variable

[1] The new point values are $\mathbf{p} = (24.26, 21.27, 18.65, 16.35, 14.33, 12.57, 11.02, 9.66, 8.47, 7.42, 6.51, 5.71, 5.00, 4.39, 3.85, 3.37, 2.96, 2.59, 2.27, 1.99)$, awarded for all 20 positions.

is the accumulated value of awarded points up to the i_{th} race, computed simply by adding up every race points.

Figure 2.1: Awarded points comparison. The x-axis denotes the different drivers positions which differ on the two classes. On one side, for the FIA, there are just 10 positions that are awarded the points (on green) and the new computed points are for the whole grid composed of 20 positions (on red). The y-axis denotes the points awarded.



Finally, the Formula 1 Power Index (PI) denoted by *PI* and adjusted Formula 1 Power Index, *PI_adj*, describe the percentage of available points and an the adjustment associated. These two values are crucial for understanding the model; they are its foundation and support. The math behind them and their contribution to the model, will be explained in the following section.

CHAPTER 3

Prediction Model

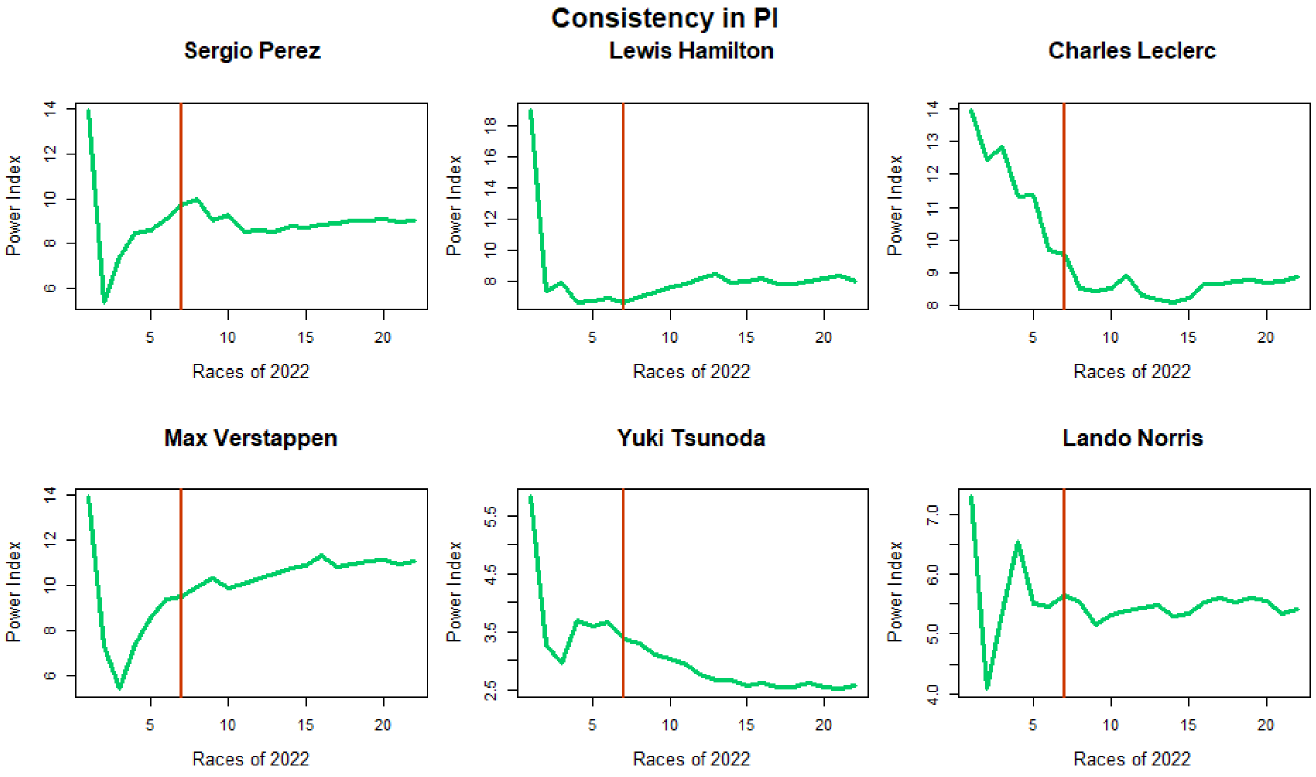
The model aims to develop an index called *Formula 1 Power Index (PI)*, which describes an estimate of the driver's strength and will later be used for race predictions. The index represents the percentage of total awarded points, as mentioned previously. The predictions are fundamentally based on the assumption that the portion of assigned points will remain constant throughout the season. In other words, the marginal change is expected to be minimal. Figure 3.1 represents a visual demonstration of this assumption. It is believed that a driver who is performing well during a certain season will continue to do so for the rest of it. On the other hand, it is also assumed that a driver who struggles during one season will continue to face challenges. Furthermore, the probability of gaining/losing places will also be considered.

To clarify ideas and extend the mentioned concept, the three primary pillars that sustain the model, are presented.

1. Power Index (PI).

   At the core of the model, lies the PI. This number represents the percentage of available points a driver owns and its purpose it to show the driver's strength in a specific season. Both, driver and constructor points, are considered in the calculation making this strength value somewhat more objective, meaning that in this way if a driver had a bad race, but has a good team, its strength value would not fall dramatically. The way it is computed allows the model to be self-contained, which means that, apart from the final constructors results from the previous season which are used as a kick-start,

Figure 3.1: Visual representation of PI consistency for drivers Sergio Perez, Lewis Hamilton, Charles Leclerc, Max Verstappen, Yuki Tsunoda, and Lando Norris during the 2022 season. The red line represents the 7th race, while the first 6 races are considered the *burn-in* period, characterized by higher uncertainty. The y-axis indicates the value of the Power Index, and the x-axis represents the races of the 2022 season.



no previous seasonal data is required; all the computations needed are contained in a one-season model. This design also enables the model to update every race, trying in this way to predict as accurately as possible next race results. The PI for the ith driver and jth constructor on the tth race is defined as:

$$PI_{it} = \frac{\gamma(accum\_points_i)_t + (1-\gamma)(team\_accum\_points_j)_t}{\gamma\sum_{k=1}^{20}(accum\_points_k)_t + (1-\gamma)\sum_{k'=1}^{10}(team\_accum\_points_{k'})_t} \times 100, \tag{3.1}$$

in the equation above, the variables draw their values from the sets *driver_id* and *constructor_id*, both specified in the dataset, along with the value of *accum_points*, which is derived from the dataset presented in the previous section.

The team (constructor) can easily be spotted in the dataset by looking at the two drivers who belong to the same constructor or it can also be seen as j such that $(team\_accum\_points_j) = (accum\_points_i) + (accum\_points_{i'} \mathbb{1}_{\text{constructor}_i = \text{constructor}_{i'}})$ where i' is another driver and the indicator function, $\mathbb{1}_{\text{constructor}_i = \text{constructor}_{i'}}$, takes a value of either 0 or 1, depending on whether the driver i and driver i' share the same constructor. Finally, t denotes the round number, ranging between 1 and 22, and $\gamma \in [0,1]$ is the parameter that optimizes the PI, which will be addressed in the subsequent section.

Equation 3.1 can be rewritten as shown below. Let $\mathbf{p}$ be a 20-dimensional vector denoting the awarded points and $\mathbf{1} \in \mathbb{R}^{20}$ be a vector made of ones.

Therefore, $\mathbf{p}^{\mathrm{T}}\mathbf{1} = \sum_{k=1}^{20} p_k = \sum_{k=1}^{20}(accum\_points_k)_1$. Multiplying the latter quantity by t, the total accumulated points up to the race t is obtained by $t\mathbf{p}^{\mathrm{T}}\mathbf{1} = \sum_{k=1}^{20}(accum\_points_k)_t$. Similarly, $2t\mathbf{p}^{\mathrm{T}}\mathbf{1} = \sum_{k=1}^{20}(team\_accum\_points_{k'})_t$ are the points accumulated by all teams. Therefore, by simplifying the denominator, we can express it as $\gamma t\mathbf{p}^{\mathrm{T}}\mathbf{1} + (1-\gamma)2t\mathbf{p}^{\mathrm{T}}\mathbf{1} = t\mathbf{p}^{\mathrm{T}}\mathbf{1}(2-\gamma)$. This simplification allows us to rewrite the equation in the following form:

$$PI_{it} = \frac{\gamma(accum\_points_i)_t + (1-\gamma)(team\_accum\_points_j)_t}{t\mathbf{p}^{\mathrm{T}}\mathbf{1}(2-\gamma)} \times 100. \tag{3.2}$$

2. Adjusted Power Index (PI_adj).

The adjusted power index serves as a supportive component for the PI. Given that Formula 1 often witnesses numerous incidents, which may not accurately reflect a driver's skill, it becomes imperative to address and handle these incidents with care. This adjustment is computed by blending the new PI (calculated for each race) with the PI_adj from the previous race. The weighting process differs depending on the driver's status; specifically, it varies if a driver encounters any form of incident, represented by *status* $\notin$ {*Finished, +1 Lap, +2 Laps*}, or if none were presented. The following adjustment for the i$^{\text{th}}$ driver and j$^{\text{th}}$ constructor on the t$^{\text{th}}$ race is defined as:

$$PI\_adj_{it} = \begin{cases} \zeta PI_{it} + (1-\zeta)last\_year\_results_j, & \text{if } t = 1 \quad \& \quad \text{no incident} \\ \eta PI_{it} + (1-\eta)last\_year\_results_j, & \text{if } t = 1 \quad \& \quad \text{incident} \\ \theta PI_{it} + (1-\theta)PI\_adj_{i(t-1)}, & \text{if } t \geq 2 \quad \& \quad \text{no incident} \\ \nu PI_{it} + (1-\nu)PI\_adj_{i(t-1)}, & \text{if } t \geq 2 \quad \& \quad \text{incident} \end{cases} \tag{3.3}$$

where $(\gamma, \zeta, \eta, \theta, \nu)^T \in [0,1]^5$ are the tuning parameters which play a crucial role as they significantly impact on the prediction accuracy, effectively managing and controlling the impact of incidents on the PI.

3. Probability of gained places.

   It is important to address the probability that a certain driver will gain or lose a specific number of places because not all pilots are able to overtake all of their opponents every single race. Therefore, the given number of events (places gained/lost) occurring in a set interval of time (every race) is computed according to a Poisson distribution.

   The probability is only computed if the grid position is close (less than one standard deviation) to the usual driver starting position (grid position's mode) because otherwise it would not be accurate to take it into account, since unusual starting positions lead to unusual results.

   Given that overtakes are often perceived as rare events, and since they are discrete and can only take non-negative values, it appears sensible to model the data using a Poisson random variable. The Poisson distribution is characterized by the lambda parameter, and thus, employing the maximum likelihood estimator for this distribution family seems reasonable.

   Two lambda parameters are considered: Let $X_{it\delta} \sim Pois(\lambda_{it\delta})$ represent random variables signifying the places gained or lost by the i^{th} driver in the t^{th} race. Here, $\delta \in \{p, n\}$ indicates whether the probability being computed corresponds to gaining places (p) or losing places (n). The parameter $\hat{\lambda}_{MLE} = \hat{\lambda}_{itp} = \sum_{t=1}^{t^*} \frac{x_{it}}{m_p}$ serves as the estimate for places gained, where $x_{it} \geq 0$ denotes the places gained, and m_p represents the count of x_{it} for driver i up to race t^*. Similarly, owing to the invariability property of the MLE, $\hat{\lambda}_{MLE} = \hat{\lambda}_{itn} = \sum_{t=1}^{t^*} \frac{|x_{it}|}{m_n}$ is the estimation for places lost. Here, $x_{it} < 0$ signifies places lost, and m_n is the count of such instances for driver i up to race t^*.

   It's crucial to emphasize that, given the Poisson distribution's range over positive values, a linear transformation involving absolute values is applied to account for the negative representation of places lost.

After completing the explanation of the main concepts, it is straightforward to expose the steps for the results predictions. It is a matter of computing the PI_adj for all 20 grid places and finding the place such that the change in computation is the minimum, having in this way the least marginal change in the PI. This can be rewritten as:

$$\underset{x \in 1,2,\ldots,20}{\operatorname{argmin}} \quad |(PI\_adj_x)_{it} - (PI\_adj)_{i(t-1)}|[1 - \mathbb{P}(X_{it\delta} = |x - grid|)], \tag{3.4}$$

where $(PI\_adj_x)_{it}$ denotes the projected adjusted PI associated to the finishing place x for driver $i \in$ *driver_id* on race $t \in [2, 22]$. This is computed, as stated before, for all the 20 different places. The previously computed PI_adj, $(PI\_adj)_{i(t-1)}$, for race $t-1$ is taken into account and $\mathbb{P}(X_{it\delta} = |x - grid|)$ represents the probability of gaining places, indicated by $\delta = p$ if $x - grid \geq 0$, or losing places denoted by $\delta = n$ if $x - grid < 0$. The inverse probability is taken because a minimization problem is being solved and thus the lower the value, the better. This is how predictions are done. It's essential to emphasize that these forecasts are made with only one day's notice because the starting position, referred to as the *grid*, is information available, and the qualifying sessions occur just one day before the race.

After having clarified both PI and PI_adj, the next step involves parameter tuning. This process aims to directly enhance the precision and accuracy of the model. The criteria for selecting the parameters that define the adjustments is the minimization of the mean squared error (MSE), which, measures the average of the squares of the errors or deviations—differences between the estimator and what is estimated. In this context, it's used as a criterion for parameter selection.

To formulate this as a minimization problem, let's denote it as follows: By letting $f(x \mid i, t) = [(PI\_adj_x)_{it} - (PI\_adj)_{i(t-1)}][1 - \mathbb{P}(X_{it\delta} = |x - grid|)]$, the second problem can be rewritten as follows:

$$\begin{aligned} & \underset{(\gamma,\zeta,\eta,\theta,\nu)^T \in [0,1]^5}{\operatorname{argmin}} && \frac{1}{n}\sum_{k=1}^{n}(\hat{y_k} - y_k)^2 \\ & \text{s.t.} && \hat{y_k} = \underset{x \in \{1,2,\ldots,20\}}{\operatorname{argmin}} \quad f(x \mid i, t) \end{aligned} \tag{3.5}$$

In the constraint of the optimization problem, $\hat{y_k}$ represents the predicted finishing position for a given driver and race, while y_k denotes the actual result. The minimization process optimizes the model's accuracy by determining how the weights should be distributed among the five parameters. Here, γ plays a crucial role in PI_it as defined in Equation 3.2, and it's later used in Equation 3.3 to calculate $(PI\_adj)_{it}$, present in $f(x \mid i, t)$, where the parameters ζ, η, θ, and ν come into play.

The significance of parameters η and ν lies in their ability to address issues stemming from unexpected events, ensuring that the occurrence of an incident does not adversely impact future predictions. Additionally, parameters ζ and η help gauge the relevance of information from the previous year's constructor's championship. Finally, γ establishes the weight between the driver and the constructor in the PI calculation, contributing to precision by striking the right balance between being a skilled pilot and having a competent team. By carefully fine-tuning these parameters, the model can compute predictions with the utmost precision.

Lastly, the evaluation criteria of the model, needs to be described. It is relevant to mention that, as previously stated in the limitations of the method, this predictive model is not taking into account the possibility of an incident happening, meaning that for the predictions it is assumed that all drivers will finish the race even though very likely, it is a false statement. Having said this, all the error computations are just calculated for drivers whose *status* $\in$ {*Finished, +1 Lap, +2 Laps*}. As mentioned above, not only the MSE was computed to tune out the parameters, but it also helps us determine the precision of the model. A measurement of how precise every single race was, was obtained. Furthermore, as another way of quantifying and evaluating the model, an accuracy rate was calculated. This compared the exact prediction, the prediction ±1 place and ±2 places. Here is the generic form of this assessment proposed:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Observations}} \times 100, \tag{3.6}$$

By computing the evaluation methods described, the season predictions were done. And by doing so, a relevant limitation of the method was found: the model is not subject to predicting that each driver will be finishing in a different place, that is, even though it correctly assigns each driver a result, each result is not assigned a driver. Thus, the model can be thought as being a surjective function and not a bijective one, as it should be.

Table 3.1: Algorithm for Formula 1 Power Index (PI) Prediction

```
Algorithm Steps

Step 1: Initialize
Set tuning parameters $(\gamma, \zeta, \eta, \theta, \nu)^T \in [0, 1]^5$ according to Equation 3.5 using last season's data
Set race count $t^*$
Step 2: For each race $t$ from 1 to $t^*$ (or up to the most recent race)
    Step 2.1: Calculate Power Index (PI)
        Use Equation 3.2 to compute $PI_{it}$ for each driver
    Step 2.2: Calculate Adjusted Power Index (PI_adj)
        Use Equation 3.3 to compute $PI\_adj_{it}$ for each driver
Step 3: For each driver $i$ at each race $t$
    Step 3.1: Calculate Probability of gained places
        Use Poisson distribution calculating parameters $\hat{\lambda}_{itp}$ and $\hat{\lambda}_{itn}$
    Step 3.2: Predict finishing position
        Use Equation 3.4 to find $\hat{y_k}$
Step 4: Model Evaluation
    Calculate Mean Squared Error (MSE) and Accuracy using Equation 3.6
```

CHAPTER 4

Data Application and Analysis

To showcase results, we've chosen the last race, round 22nd, expecting improved model performance due to increased training throughout the season. Figures 4.1 present the results compared to the actual values, with errors calculated per driver. The error indicates prediction accuracy: positive if the actual result is better, negative if worse, and 0 if accurate. For instance, if Verstappen started 1st, the model predicted 2nd, but he finished 1st, the error is -1.

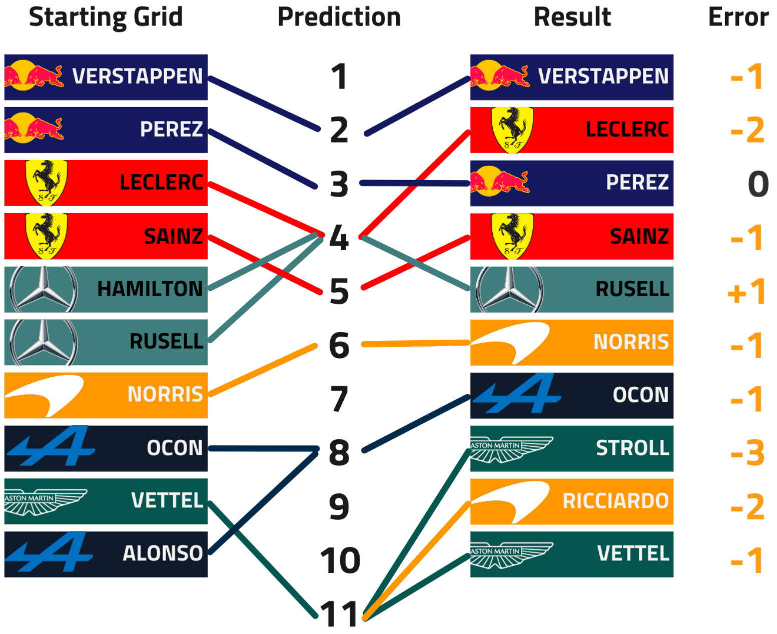


Figure 4.1: Display of the starting grid, predictions, and final results for the top 10 drivers in the Abu Dhabi Grand Prix.

Moreover, we can analyze the results by calculating the root mean squared error (RMSE), providing insight into the accuracy of the race predictions. Figure 4.2 illustrates these results, displaying the RMSE for each race with $t \in [2, 22]$.

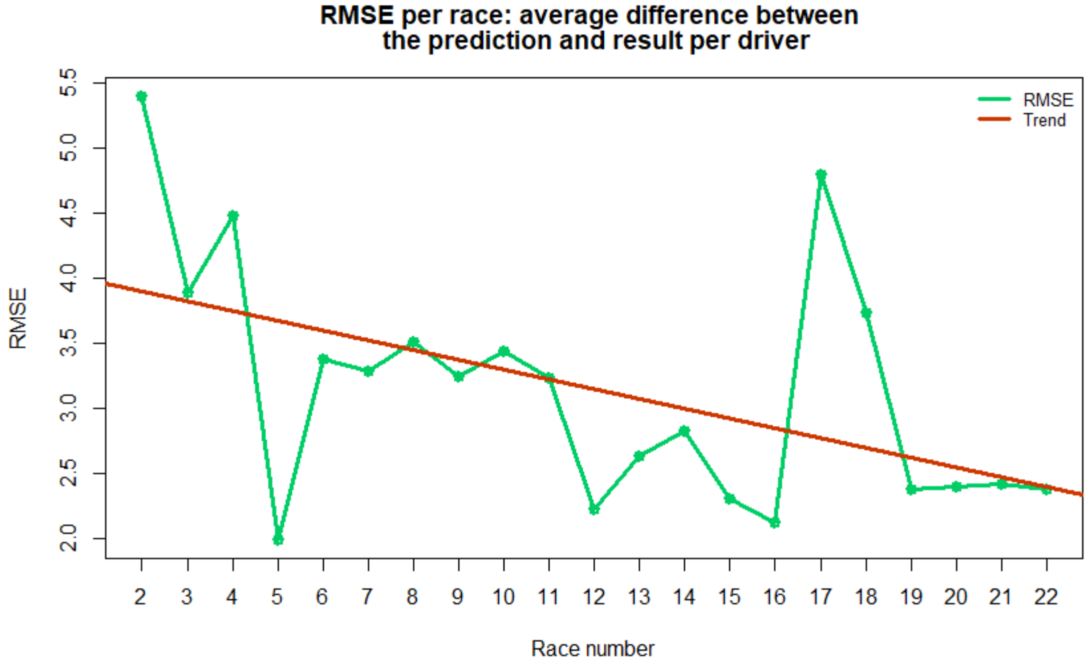


Figure 4.2: Exact values and trend of RMSE for races 2 to 22 from the predicted results of the races in season 2022 of Formula 1.

The subtle trend observed is a result of greater variability at the beginning of the year, which decreases as the season progresses, directly impacting the magnitude of the RMSE. The decline in RMSE is attributed to the model acquiring more information and, consequently, becoming more accurate as the season unfolds. For the presented results, the model exhibits an RMSE of 2.38, approximately equivalent to a 2-error-place prediction per driver.

In the 2022 season there were a total of 72 incidents, this number plus the race predicted for Nyck de Vries, a driver who raced only once, gives a total of $420 - 73 = 347$ races to predict.

$$\text{Accuracy For Exact Places} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Observations}} \times 100 = \frac{48}{347} \times 100 = 13.83\% \tag{4.1}$$

The accuracy for ±1 place and ±2 places was also carried out. Having a total of 150 correct predictions considering a wider range of ±1 place, giving an accuracy of 43.23%. For the range of ±2 places, 214 predictions were correct, providing the model with a 61.67% accuracy.

The shown accuracy should be attributed to the tuning parameter, which, as stated in the previous chapter, there are five parameters, $\gamma, \zeta, \eta, \theta$ and ν that must be adjusted to reach the best model. This is accomplished by solving the minimization problem proposed in equation 3.5. The following Figure 4.3 displays the results of the minimization.

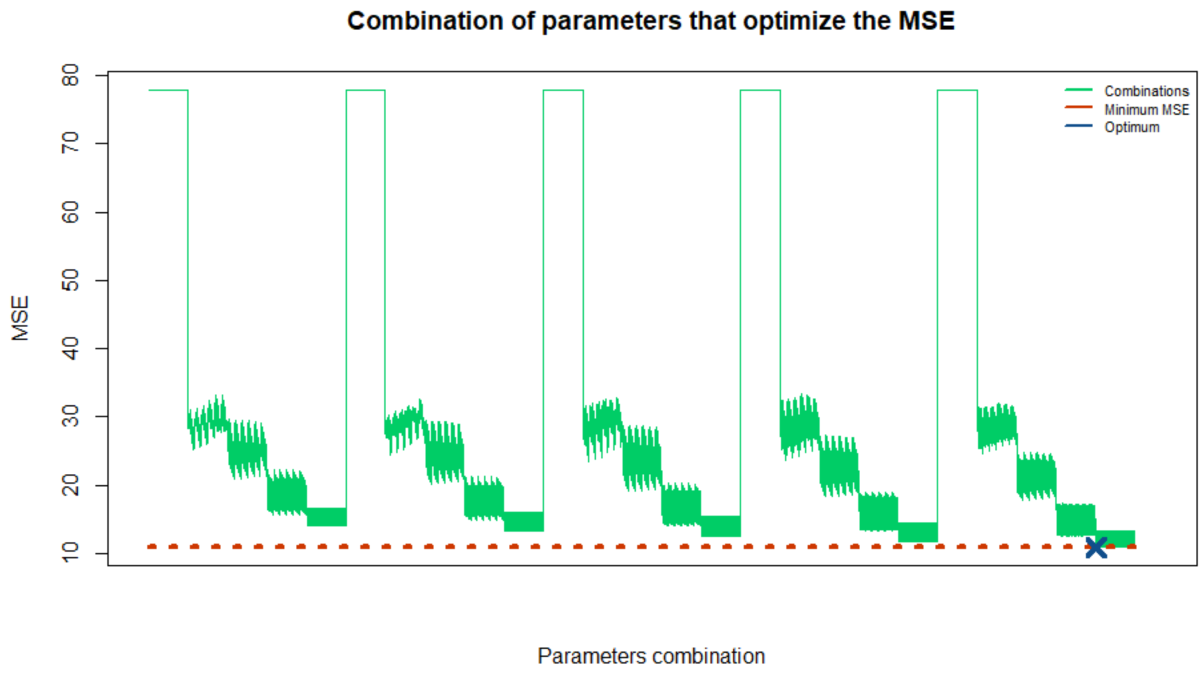


Figure 4.3: Minimization MSE by choosing the right combination of the parameters $\gamma, \zeta, \eta, \theta$ and ν. Each point on the graph represents a different combination of values, always within the region $[0, 1]^5$. The red dotted line represents the minimum value of the MSE, and the blue cross is where the min MSE value is reached, hence, the optimal value of the parameters.

Here, over 40,000 distinct parameter combinations, all falling within the interval $[0, 1]^5$, were subjected to testing and assessment. The goal was to identify the combination that resulted in the lowest MSE. The result obtained was an optimal set of values, specifically $(\gamma, \zeta, \eta, \theta, \nu)^T = (0.9825, 0.5, 0.5, 0.985, 1)^T$ that led to the MSE value of 8.58.

The value of $\gamma = 0.9825$ means that to compute the PI, the driver must be considered almost on a 100%. Even though the driver is embedded in the constructor, individual participation is the only relevant factor when it comes to predicting results using this model. The parameters $\zeta = 0.5, \eta = 0.5, \theta = 0.985, \nu = 1$ correspond to situations in which the *status* considers distinctions when an incident is presented, and the *race* number treats the first race $t = 1$ as a specific case. We can interpret it as follows: for the initial race $t = 1$, where $\zeta = 0.5$

and $\eta = 0.5$, equal relevance is given to both last year's results and the PI, irrespective of whether incidents were presented.

For all subsequent races with $t \geq 2$, if no incident is presented, the optimal value of $\theta = 0.985$ indicates that only a minimal adjustment in PI is required. However, in the case of a presented incident (for $t \geq 2$), having a value of $\nu = 1$ signifies that the PI must be considered in its entirety, accounting for 100% of the modification.

CHAPTER 5

Conclutions

The central focus of this study was to devise a model capable of predicting the final positions of the top 20 drivers in each race. The results suggest that the adaptation employed from [2] can be utilized and adjusted for multiple sports; we are confident that this algorithm, with necessary adaptations, would be effective in any sport. The analysis showed that the model had an accuracy of 13.83% for exact predictions, 43.23% for ±1 place, and 61.67% for ±2 places. Notably, these results demonstrate better accuracy than published alternative models employing supervised learning techniques, including neural networks.

In the realm of sports prediction, this model contributes to the body of knowledge by specifically addressing Formula 1 race outcomes. Unlike sports with simpler outcomes (win/loss/draw), Formula 1 presents a multi-class outcome structure, and this project helps filling in the existing gap in the literature. This paper not only proposes a model tailored to such complexities but also supports the increasingly popular practice of betting on Formula 1, driven by its growing spectator base. Confirming in this way the possibility of obtaining models capable of adapting to other sports.

In terms of future research directions, an opportunity lies in incorporating race incidents into the model. It could involve simulating scenarios that accounts the occurrence of such unforeseen events potentially leading to a solution proposal for the prediction's assignment restriction, wherein the potential for multiple drivers being predicted to end up in the same position is present. Therefore, it is thought that computing simulations could align more closely to what truly happens in real life.

Appendix A
Example

Now, we will be detailing the computational process for the result prediction of Lewis Hamilton (*hamilton*), the seven-time world champion, during the 10th race: the iconic British Silverstone race.

The data used as for the example is presented in Table A.1. Two rows of data are shown since the results for race 10 are treated as unknown, utilizing only the starting grid information and considering only the data up to race 9. The prediction for the race outcome requires computation of Equation 3.4. From the table, it is known that $(PI\_adj)_{\text{hamilton9}} = 7.30$ and $grid = 5$ thus it leaves the following equation:

$$\underset{x\in 1,2,\dots,20}{\operatorname{argmin}} \quad |(PI\_adj_x)_{\text{hamilton10}} - 7.30|[1 - \mathbb{P}(X_{\text{hamilton10}\delta} = |x-5|)]. \tag{A.1}$$

To calculate this, the first step is to obtain the projected adjusted PI. For explanatory purposes, only one such calculation will be performed: $(PI\_adj_1)_{\text{hamilton10}}$. This calculation involves straightforward basic operations, and the remaining values can be acquired by

Table A.1: Data rows of Lewis Hamilton in races 9 and 10 extracted from the full dataset shown in Table 2.1.

driver_id	constructor_id	round	season	grid	position	status	pos_gained	round_points	accum_points	PI	PI_adj
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
hamilton	mercedes	9	2022	4	3	Finished	1	18.65	118.98	7.31	7.30
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
hamilton	mercedes	10	2022	5	3	Finished	2	18.65	137.62	7.58	7.58
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

following the equations outlined in the methodology. Initially, to compute the projected PI, it is needed the substitution of the values from the table A.1 into the equation 3.2, resulting in:

$$(PI_1)_{\text{hamilton10}} = \frac{\gamma(accum\_points_{\text{hamilton}})_{10} + (1-\gamma)(team\_accum\_points_{\text{mercedes}})_{10}}{10\mathbf{p}^{\mathsf{T}}\mathbf{1}(2-\gamma)} \times 100 \tag{A.2}$$

$$= \frac{0.9825(118.98 + 24.26) + 0.0175(266.96 + 24.26 + 16.44)}{10 \times 182.62 \times 1.0175} \times 100 \tag{A.3}$$

$$= \frac{146.11}{1858.16} \times 100 = 7.86 \tag{A.4}$$

Where, apart from the values retrieved from Table A.1, the value $p_1 = 24.26$ is equivalent to the points awarded if *hamilton* happen to end up in first place, 16.44 equals Hamilton's teammate (*russell*) average points, $266.96 = 118.98 + 147.98$ represent the sum of accumulated points for *hamilton* and *russell*, respectively, and $\gamma = 0.9825$ is the optimal value for the parameter. Given that the optimal value for θ is $\theta = 0.985$, it follows that

$$(PI_1)_{\text{hamilton10}} = 0.985(PI\_adj_1)_{\text{hamilton10}} + 0.015(7.30) = 0.985(7.86) + 0.015(7.30) = 7.85 \tag{A.5}$$

According to equation 3.3. Both parameters', γ and θ, optimization will be shown further down in this chapter, after the presentation of results prediction. Therefore, as the result A.4 shows, $(PI\_adj_1)_{\text{hamilton10}} = 7.85$ which is the projected adjusted PI for Hamilton if he ended up in first place.

Continuing from the calculation presented in Equation A.1, it is observed that, for $x \in [1, 5]$, the probability $\mathbb{P}(X_{\text{hamilton10}p} = 5 - x)$ is computed. Similarly, for $x \in [6, 20]$, the corresponding probability is $\mathbb{P}(X_{\text{hamilton10}n} = x - 5)$. Retrieving from the dataset, on Table 2.1, the positions gained by *hamilton* up until the race 9, it is found that he gained $2, 5, 1, 1, 0, 1, 0, 3, 1$ positions in races 1 to 9, respectively. These makes $\lambda_{\text{hamilton10}p} = \frac{2+5+1+1+1+3+1}{9} = \frac{14}{9} = 1.56$ and $\lambda_{\text{hamilton10}n} = 0$. Now, the prediction can be done.

Table A.2 shows the procedure for the finishing places $x = 1, 2, \ldots, 20$. It is not hard to spot that the value of x, where the function minimizes as shown in equation A.1, is $x = 5$. Therefore, the prediction for Hamilton on race 10, is to finish in 5th place.

This exact course of action is needed to predict the results of any other driver for whichever race. Following with the example, the same race is taken as an exemplification to predict the finishing places for all the pilots. The results can be found in Table A.3. This table displays the results in the actual finishing position order, where, in addition to the variables that indicate the driver (*driver_id*) and the team (*constructor_id*), it presents the *position* column which

Table A.2: Procedure of the solution to the Equation A.1 the rows show the calculation involved in obtaining the difference between the projected adjusted PI and the last adjusted PI, times the inverse probability of finishing in position x.

Position (x)	**Procedure**
1	$\|7.85 - 7.30\|[1 - \mathbb{P}(X_{hamilton10p} = 4)] = 0.55 \times 0.95 = 0.53$
2	$\|7.70 - 7.30\|[1 - \mathbb{P}(X_{hamilton10p} = 3)] = 0.39 \times 0.87 = 0.35$
3	$\|7.56 - 7.30\|[1 - \mathbb{P}(X_{hamilton10p} = 2)] = 0.26 \times 0.74 = 0.19$
4	$\|7.44 - 7.30\|[1 - \mathbb{P}(X_{hamilton10p} = 1] = 0.13 \times 0.67 = 0.09$
5	$\|7.33 - 7.30\|[1 - \mathbb{P}(X_{hamilton10p} = 0)] = 0.03 \times 0.79 = 0.02$
6	$\|7.24 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 1)] = 0.07 \times 1 = 0.07$
7	$\|7.15 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 2)] = 0.15 \times 1 = 0.15$
8	$\|7.08 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 3)] = 0.22 \times 1 = 0.22$
9	$\|7.02 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 4)] = 0.28 \times 1 = 0.29$
10	$\|6.96 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 5)] = 0.34 \times 1 = 0.34$
11	$\|6.91 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 6)] = 0.39 \times 1 = 0.39$
12	$\|6.87 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 7)] = 0.43 \times 1 = 0.43$
13	$\|6.83 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 8)] = 0.47 \times 1 = 0.47$
14	$\|6.80 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 9)] = 0.50 \times 1 = 0.50$
15	$\|6.77 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 10)] = 0.53 \times 1 = 0.53$
16	$\|6.75 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 11)] = 0.55 \times 1 = 0.56$
17	$\|6.73 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 12)] = 0.58 \times 1 = 0.58$
18	$\|6.71 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 13)] = 0.60 \times 1 = 0.60$
19	$\|6.69 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 14)] = 0.61 \times 1 = 0.62$
20	$\|6.67 - 7.30\|[1 - \mathbb{P}(X_{hamilton10n} = 15)] = 0.63 \times 1 = 0.63$

indicates the real value of the finishing place, *prediction* is the value obtained by computing all the calculations previously shown and the *status* variable indicates if the driver successfully finished the race.

A good way to measure how accurate were this race predictions is the root mean square error (RMSE). This is simply done by taking the squared root of the MSE, helping in this way with the interpretation of the measurement. By computing this evaluation method, the differences between the real values and the predictions will be exposed. In other words, the RMSE will serve to uncover the average of the magnitudes of the errors in predictions. The way to obtain the value is by:

$$\text{RMSE} = \sqrt{\frac{4^2+2^2+2^2+0+5^2+2^2+(-4)^2+7^2+4^2+2^2+1+3^2+(-2)^2+(-2)^2}{14}} = 3.34 \tag{A.6}$$

This number means that, on average, for the 10^{th} race, an error of $3.34 \approx 3$ places is presented on each prediction. An important limitation of the method must be considered: it does not account for incidents, and as a result, they are excluded from error calculations, since calculating errors for unaccounted occurrences would be illogical.

Continuing with the application of the method, mirroring the presented calculations of the 10^{th} race that exemplified the proper utilization of the equations, predictions were carried out for the whole season.

Table A.3: Predictions corresponding to the iconic British race in Silverstone, race 10, for all the 20 pilots in the grid. This includes information related to the driver (*driver_id*), the associated constructor (*constructor_id*), the real *position*, the *status* and the *error* of the predictions.

driver_id	**constructor_id**	**position**	**prediction**	**error**	**status**
sainz	ferrari	1	5	4	Finished
perez	red_bull	2	4	2	Finished
hamilton	mercedes	3	5	2	Finished
leclerc	ferrari	4	4	0	Finished
alonso	alpine	5	10	5	Finished
norris	mclaren	6	8	2	Finished
max_verstappen	red_bull	7	3	-4	Finished
mick_schumacher	haas	8	15	7	Finished
vettel	aston_martin	9	13	4	Finished
kevin_magnussen	haas	10	12	2	Finished
stroll	aston_martin	11	12	1	Finished
latifi	williams	12	15	3	Finished
ricciardo	mclaren	13	11	-2	Finished
tsunoda	alphatauri	14	12	-2	Finished
ocon	alpine	15	8	-	Fuel pump
gasly	alphatauri	16	11	-	Collision damage
bottas	alfa	17	7	-	Gearbox
russell	mercedes	18	4	-	Collision
zhou	alfa	19	13	-	Collision
albon	williams	20	12	-	Collision

Bibliography

[1] David Aldous. “Elo Ratings and the Sports Model: A Neglected Topic in Applied Probability?” In: *Statistical Science* 32.4 (2017), pp. 616–629. DOI: 10.1214/17-STS628. URL: https://doi.org/10.1214/17-STS628.

[2] Jay Boice. *How Our Club Soccer Predictions Work*. Tech. rep. FiveThirtyEight, 2020. URL: https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/.

[3] Maury Brown. *Inside The Numbers That Show Formula 1's Popularity And Financial Growth*. Tech. rep. Forbes, 2023. URL: https://www.forbes.com/sites/maurybrown/2023/03/29/inside-the-numbers-that-show-formula-1s-popularity-and-financial-growth/.

[4] Rory Bunker and Teo Susnjak. “The Application of Machine Learning Techniques for Predicting Results in Team Sport: A Review”. In: (Dec. 2019). DOI: 10.13140/RG.2.2.22427.62245.

[5] Santiago Casanova and Philip Bulsink. *f1dataR*. 2023. URL: https://scasanova.github.io/f1dataR/index.html.

[6] Aram Ebtekar and Paul Liu. “Elo-MMR: A Rating System for Massive Multiplayer Competitions”. In: *Proceedings of the Web Conference 2021*. WWW '21. Ljubljana, Slovenia: Association for Computing Machinery, 2021, pp. 1772–1784. ISBN: 9781450383127. DOI: 10.1145/3442381.3450091. URL: https://doi.org/10.1145/3442381.3450091.

[7] KIKE FRANSSEN. “COMPARISON OF NEURAL NETWORK ARCHITECTURES IN RACE PREDICTION”. PhD thesis. tilburg university, 2021.

[8] Tomislav Horvat and Josip Job. "The use of machine learning in sport outcome prediction: A review". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10 (June 2020), e1380. DOI: 10.1002/widm.1380.

[9] Chinwe Igiri. "An Improved Prediction System for Football a Match Result". In: *IOSR Journal of Engineering* 04 (Dec. 2014), pp. 12–020. DOI: 10.9790/3021-04124012020.

[10] Martin Ingram. In: *Journal of Quantitative Analysis in Sports* 17.3 (2021), pp. 203–219. DOI: doi:10.1515/jqas-2020-0066. URL: https://doi.org/10.1515/jqas-2020-0066.

[11] ACSIJ J. and Hamid Rastegari. "A Review of Data Mining Techniques for Result Prediction in Sports". In: 2 (Nov. 2013).

[12] Ravindu Kavishwara. *Formula 1 Machine Learning*. Tech. rep. Becoming Human: Artificial Intelligence Magazine, 2021.

[13] Stephanie Ann Kovalchik. "Searching for the GOAT of tennis win prediction". In: *Journal of Quantitative Analysis in Sports* 12.3 (2016), pp. 127–138. DOI: doi:10.1515/jqas-2015-0059.

[14] Ville Kuosmanen. *Predicting Formula 1 results with Elo Ratings*. Tech. rep. Medium - Sports Analytics, 2020. URL: https://towardsdatascience.com/predicting-formula-1-results-with-elo-ratings-908470694c9c.

[15] Ankur Patil et al. "A Data-Driven Analysis of Formula 1 Car Races Outcome". In: *Artificial Intelligence and Cognitive Science*. Ed. by Luca Longo and Ruairi O'Reilly. Cham: Springer Nature Switzerland, 2023, pp. 134–146. ISBN: 978-3-031-26438-2.

[16] Elena Loli Piccolomini, Davide Evangelista, and Massimo Rondelli. *The Future of Formula 1 Racing: Neural Networks to Predict Tyre Strategy*. Bachelor's Thesis. 2022.

[17] Ben Powell. In: *Journal of Quantitative Analysis in Sports* 19.3 (2023), pp. 223–243. DOI: doi:10.1515/jqas-2023-0004. URL: https://doi.org/10.1515/jqas-2023-0004.

[18] Luke Pritchard. *Predicting Formula 1 Races using Ordered Logistic Regression and Elo Ratings*. Tech. rep. The College Wooster, 2022. URL: https://wooster.edu/2022/04/26/luke-pritchard/.

[19] Horatiu Sicoie. “Machine Learning Framework for Formula 1 Race Winner and Championship Standings Predictor”. Bachelor’s Thesis. Tilburg University: Department of Cognitive Science Artificial Inteligence School of Humanities and Digital Sciences, 2022.

[20] Eloy Stoppels. *Predicting race results usingartificial neural networks*. 2017. URL: https://api.semanticscholar.org/CorpusID:67301053.