Forecasting Brasileirão 2023:

A mathematical approach to the Brazilian soccer league in its current format

Lucas Fernandes

Writing in the Sciences

Jordon Smith

April 14, 2023

Abstract

In this article, I explored the development of a prediction model algorithm to forecast the Brasileirão 2023 final standings. Initially, I was unsure if my desired methodology was possible, so at first, I gathered data to build twenty different datasets, one for each club in Brazil's national soccer league, by searching for open-source websites with creditable information. With the data in my hands, I developed twenty different linear regressions, all of them with the same target variable, but each one built differently, and combined them into a final formula that weighted a historical feature to achieve the final forecast. The forecast had some clear and some not-so-clear limitations, such as the short data sample, a mean error for each regression, and inaccuracy that will be observed throughout the season. Furthermore, the data brought us to some results that were expected by those who keep up with the soccer scenario in Brazil, and some not so much, that included but were not limited to a repeat championship by Palmeiras and a Red Bull Bragantino relegation.

Introducing Data Science in Brazilian Soccer

Data Science and Analytics have been heavily brought to sports in recent years as a method of improving teams' performance. A clear example of how that has been used is Daryl Morey's Houston Rockets' impressive usage of the three-point line led to a complete change in the course of Basketball's playstyle.[1] Still, in this paper, the goal was to explore data science in the Brazilian soccer league with an outside perspective of the game and its standings. Instead of understanding the game of soccer itself, the aim was to try to forecast a regular twenty-team soccer league, in resemblance to the University of Oxford publication in November 2022 of their forecast for the 2022 FIFA World Cup, which went viral in social media.[2]

After going through *Analisando Dados do Brasileirão Série A* by Juvenal Fonseca, in which the author observed the historical performance of Brazilian clubs since 2003, I wondered if there was any pattern that would individually or collectively affect each team, and if this data could be gathered, observed, and applied on developing a forecasting model.[3] Still, most tournament forecasting models out there take into account scheduled games and possible outcomes for each match, as it makes sense since the results of a tournament are the product of all of its matches, but intending to try something different, Instead of looking at the team's calendars, since every team roughly has the same calendar in a twenty-club soccer league, this research looks at the individual performances of every club and how they replicate themselves through time. Furthermore, some wonderings went over what other data could have some impact

[1] Alex Wong, "A High-Scoring Revolution Has the Rockets Soaring", explanation of Daryl Morey's basketball.
[2] Oxford Mathematics, "A Mathematician's Guide to the World Cup", University of Oxford World Cup modelling
[3] Juvenal Fonseca, "Analisando Dados Do Brasileirão Série a - Dados Ao Cubo - Python."

on a club's final number of points each year, where could this data be found, and how it complements what Fonseca had already got, each club's scoring and defending.

Similarly to most major national soccer leagues around the globe, Brasileirão consists of a single round-robin in which all of its twenty teams play each other twice, with a home and an away match. A win represents a gain of three points, a draw gives each team a point and a loss simply does not add points. By the end of the tournament, the team that accumulates the highest number of points is the champion and the full standings follow in descending other of acquired points, with the number of wins, goal difference, and goals scored coming as the most relevant tie breakers. The main difference between the Brazilian league when compared to some of the others is the relegation system, in which most leagues relegate three clubs, often followed by a play-off between an ascending and a relegated team, while in Brazil, there are four direct relegation spots, with no post-tournament play-off. This relegation system was one of the major obstacles to bypass throughout developing the prediction model, as only four teams in Brazil have played every single first-division tournament in the selected period for the research, 2003 to 2023. But working with data analytics in soccer, especially in Brazil, has one extra cultural barrier, that I had to fight myself. Soccer is not forecastable. The beauty of this game is that the outcome usually does not make sense. Beyond that, this study is trying to rationalize something that millions of people are passionate about and thus struggle to enjoy making soccer something more rational than heartfelt, but for this experiment, I will put my feelings aside and explore some data science inside the beautiful game.

Data gathering and Datasets constructions

When first idealizing the conduction of this experiment, the period to be analyzed was already clear since 2003 is the starting year of the single round-robin format in the Brazilian league. Still, the building of databases started in the year 2006, and the reason for that is the number of teams participating in the tournament. Starting in 2003, and repeating in 2004, the Brasileirão was held in a round-robin with twenty-four teams instead of the regular twenty clubs we observe today, and in 2005 the league counted twenty-two teams participating until the format that is still used to this date was finally established, in 2006. This constant change had an impact on our data since our sample is already considerably short, with only seventeen leagues played in the current format, and the data used to impact the forecast gathered from the previous seasons of the tournament. Nine features were gathered to build each club's dataset, those being the target feature Points, the moving average of points, goals scored, and goals received in the past three years, the performance on the same season state tournament, the average attendance per season, the club's roster's value, how long the current coaching staff has been in charge of the team, and a club's tradition and historical performances in the tournament.

Features

- Points (PONTOS): the target feature is also the simplest to gather, simply representing how many points a team ended the tournament in a given year.
- Moving average of points in the past three years (MM3_PTS): Simply the mean number of points of the last three years of a team. This feature has a slightly higher magnitude in the first three years of the data frame since the previous format included more teams and more matches in a season.
- Moving average of goals scored in the past three years (MM3_ATK): Similar to MM3_PTS, but considering goals scored. This feature also has a slightly higher magnitude in the first three years of the data frame for the same reason.

- Moving average of goals received in the past three years (MM3_DEF): Similar to the previous two features, considering goals received, being one of the two features intended to be inversely proportional to a club's performance. This feature also has a slightly higher magnitude in the first three years of the data frame.
- Performance on state tournament (ESTADUAL): Simply stating the team's final position in their respective state league that happens right before Brasileirão starts. Also intended to be inversely proportional to the target as the higher position, the worse performance.
- Average Attendance per season (MEDPUB): Another self-explanatory feature, but with a few observations. The data for 2006 over this is very blurry and not so trustworthy. Furthermore, between 2010 and 2013 multiple clubs had to play in alternate stadiums due to renovations being made on their stadiums for the world cup, leading to much lower attendance in those seasons. The same happened in the 2021 season due to the Covid-19 pandemic, while in 2020 every club's attendance was zero.[4][5][6]
- Roster Value (PRECO): The estimated valuation of a club's roster for a given season in millions of euros, gathered from Transfermarkt.[7] The data provided by Transfermarkt is not precise for the 2006 and 2007 seasons since most rosters had

[4] Perspectiva Online, "Média de Público Final Do Brasileirão 2008", for 2008 average attendance data

[5] Rsssf Brasil, "Médias de Público Dos Principais Clubes No Campeonato Brasileiro", for attendance data up to 2010

[6] Transfermarkt, "Mercado de transferências, rumores, valores de mercado e notícias", the data was collected from multiple different pages for each club attendance that was available in Transfermarket.

[7] Transfermarkt, "Mercado de transferências, rumores, valores de mercado e notícias", the data was collected from multiple different pages for each club roster valuation that was available in Transfermarket.

very few players with a determined market value. Data seemed more precise starting in 2008.

- Coaching staff longevity (TMPTECDIA): Time in days a coaching staff oversaw the team up to their first game in the season.
- Tradition (TRAD): A short formula I developed that will not be used in the individual forecast but in the collective forecast in the end. The idea behind this is a common superstition in Brazilian soccer that determined clubs have a "heavier jersey", meaning that their historical success could indirectly impact their future performance. To measure a club's "tradition" I simply added ten points for each championship, seven for a runners-up position, five for third place, and two for fourth place. For example, Athletico's tradition would be 10 * 1 + 7 * 1 + 5 * 1 + 2 * 1, totalizing 24, since the club has been in each spot of the top four once in their history.

Finally, on every data set, whenever a club did not play in the first division due to relegation, all the values in that year were set to 0. Thus, this would impact the moving averages features drastically, as it is expected that teams that were recently promoted will not perform well, even though there are exceptions such as Internacional in 2018 and Grêmio in 2006.

Data modelling and Forecast

With all data in my hands, it was time to advance on applying the idealized algorithm. The plan was to develop twenty separate regular linear regressions and use all the described features, except tradition, to fit into the target points. One by one, the teams' forecasts were produced, selecting the traits that would develop the best model found until a list, with the values of each club's expected points, was built.

Starting by treating the data, the first thing done was to drop every row in all data frames that represented a year in which the club did not play in the first division due to relegation. This data, which were all zeros, had already impacted the moving averages data, and it would not make sense to force the model to forecast years in which every feature's value was zero. Following that, all features were initially added to each regression, and the fitting was run for every year that a given club played in the first division starting in 2009. The reason for that was that the data for the first three years in the datasets were impacted by the format changing with more clubs earlier in the century. Furthermore, some clubs did not play in the first division in the period from 2004 to 2006, then their first participation was selected as the initial train sample independently of what year that was.

With the data treated in such a way, the regressions were fitted multiple times by increasing the sample train data by a year each time. The miss's absolute value for each test run was summed up, making it possible that an average miss per test could be obtained and used as a parameter to evaluate the quality of the prediction. One of the goals of each model was to achieve a mean miss of fewer than ten points, which was possible in all but three regressions. Moreover, the feature coefficients, which represent their direct impact on the forecast, were used as a parameter to assess the forecast, being evaluated by the logical assumption of a feature impact over the prediction. As mentioned before, only the performance in the state league and the goals received were intended to be inversely proportional features but many times the regression coefficients for the traits were paradoxical. Comparing a feature's correlation to the target value was a way of perceiving which coefficients were logical and which were not. Still, some of the forecasts kept antagonistic values for either an increase in performance or deeper reasons, which will be further explained. Lastly, the verisimilitude of a team achieving a certain number of points was the last parameter used as a quality standard of the model. Now, this being very subjective and assessed by my perception was a limiting factor in the model. Even though I tried to be as unbiased as I could, I can't fully measure how much impact this could lead to in the final forecast. Still, it was necessary to be done due to the short overall sample that led to some predictions not being true to the likelihood of the material reality.

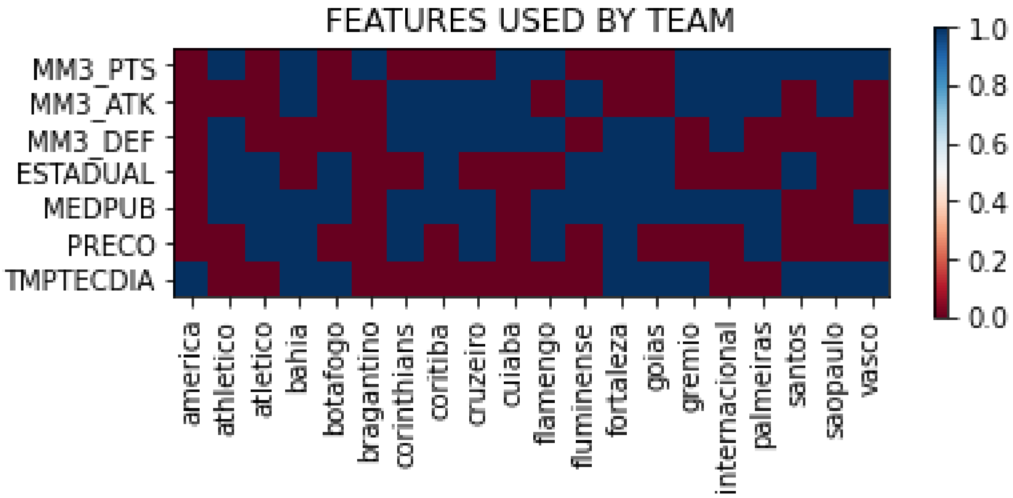
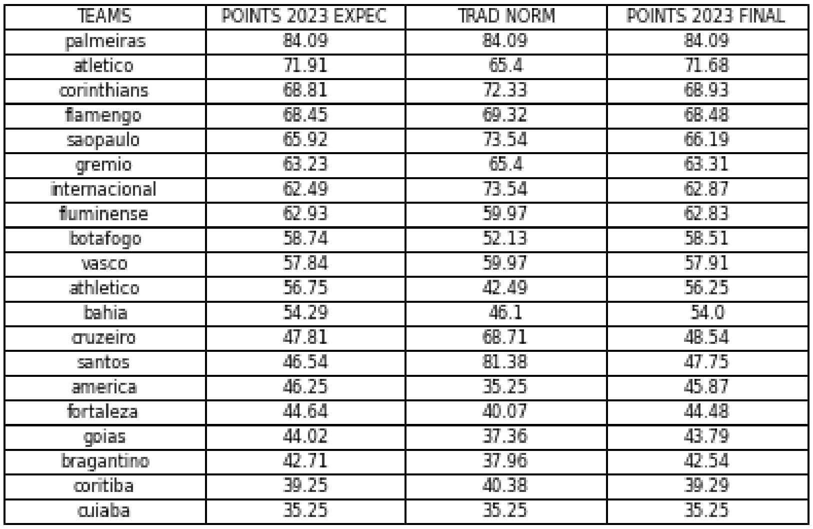


Table 1: Features used by team (plotted using Matplotlib in Python), 1 being used and 0 being not used.

By analyzing Table 1, we observe that the most used feature overall was the attendance average, while the least used was the roster value. Furthermore, we notice that América's and Bragantino's regressions only contain one trait. Those are two clubs that played Brasileirão a few seasons and had very short samples, which led to a much more inaccurate and unlikely model when compared to other clubs. This is also true for Fortaleza and Cuiabá, even though I was able to keep more features in their regressions. Corinthians' regression is the only one that goals scored is one a trait with a negative coefficient. Even though this may not seem right at first, the correlation between Corinthians scoring fewer goals and their success is observed throughout the years, as they are widely known for being a defensive team, so I decided to keep that feature.

Three forecasts the best designed were Goiás, Botafogo, and Santos, as their mean miss was noticeably low and generated very likely scenarios using three or more features.

Now there is much more to be taken from the feature selection but let's move on to applying the Tradition feature over the expected points. After comparing the Tradition value for every year and the performance, a 34.8% correlation between them is observed, proving it made sense to weigh it into the model. Still, it was probably already indirectly weighted throughout time, so to take the Tradition into account, it was applied a linear normalization to the feature values to fit into the points, and 10% of the correlation between the target and the trait was taken off from the original expected value, applying the weighed tradition into it in the following formula $P_F = \left(1 - \frac{C}{10}\right) \times P_E + \frac{C}{10} \times T$, in which P_F represents the final points for a team, C the correlation between TRAD and PONTOS, P_E the expected points from the twenty regressions and T the normalized value for the tradition. With this formula applied, it brought us the following results.



TEAMS	POINTS 2023 EXPEC	TRAD NORM	POINTS 2023 FINAL
palmeiras	84.09	84.09	84.09
atletico	71.91	65.4	71.68
corinthians	68.81	72.33	68.93
flamengo	68.45	69.32	68.48
saopaulo	65.92	73.54	66.19
gremio	63.23	65.4	63.31
internacional	62.49	73.54	62.87
fluminense	62.93	59.97	62.83
botafogo	58.74	52.13	58.51
vasco	57.84	59.97	57.91
athletico	56.75	42.49	56.25
bahia	54.29	46.1	54.0
cruzeiro	47.81	68.71	48.54
santos	46.54	81.38	47.75
america	46.25	35.25	45.87
fortaleza	44.64	40.07	44.48
goias	44.02	37.36	43.79
bragantino	42.71	37.96	42.54
coritiba	39.25	40.38	39.29
cuiaba	35.25	35.25	35.25

Table 2: Final forecast of Brasileirão 2023, plotted using Matplotlib in Python

The Tradition feature proved to have its value, but with the method used to apply it, the only significative change over the final standings was Internacional surpassing Fluminense.

Conclusions

Some of the most interesting results the forecast obtained were a repeat championship by Palmeiras, many points ahead of any other team, with Atlético, Corinthians and Flamengo completing the top four.[8] Beyond that, the relegation zone was composed of Goiás, Bragantino, Coritiba and Cuiabá. Grêmio demonstrated a surprising result after coming back from the second division, while Cruzeiro and Santos were the two most traditional teams with lower tier performances, as most would expect due to the given situation of both clubs. [9] Finally, Fluminense[10] being placed 8th was an odd prediction with the given scenario of the club.

Developing this prediction model was certainly a challenge as I am yet to find something similar to that, but also rewarding to understand how mathematics can comprehend and read through a sports competition. Acknowledging the model's limitations and weakness, its accuracy or not will be proven through the year 2023, and by the end of the tournament, there's margin to come back to this forecast and evaluate what went right and what did not, as well improve it for future versions of it. Conducting this experiment was also just a glance of what algorithms can do better than just fortune-telling a sports event, as it can certainly be replicated in any other soccer league around the globe, by gathering the right datasets and applying the proper methodology.

[8] Palmeiras is the current Brasileirão champion, winning its 11th title in 2022.

[9] Grêmio was playing in Brazil's second division in 2022. Cruzeiro and Santos have been off to a bad start in the 2023 season.

[10] Fluminense is taken by most soccer analysts as one of the, if not the strongest club in Brazil as of early 2023.

Bibliography

Arruda, Marcelo de. 2013. “Médias de Público Dos Principais Clubes No Campeonato Brasileiro.” The Rec.Sport.Soccer Statistics Foundation. December 17, 2013. https://rsssfbrasil.com/miscellaneous/mediaspub.htm.

“BOLA N@ ÁREA - O Arquivo Do Futebol.” n.d. Www.bolanaarea.com. Accessed April 11, 2023. https://www.bolanaarea.com.

Bull, Joshua. 2022. “A Mathematician’s Guide to the World Cup.” Www.youtube.com. Oxford Mathematics. November 17, 2022. https://www.youtube.com/watch?v=KjISuZ5o06Q.

Colorados Anônimos. 2022. Coloradosanonimos.com.br. 2022. https://www.coloradosanonimos.com.br.

Fonseca, Juvenal. 2021. “Analisando Dados Do Brasileirão Série a - Dados Ao Cubo - Python.” Dados Ao Cubo. July 19, 2021. https://dadosaocubo.com/analisando-dados-do-brasileirao-serie-a/.

“Grêmiopédia, a Enciclopédia Do Grêmio.” 2020. Www.gremiopedia.com. August 2020. https://www.gremiopedia.com/.

Kruse, André. n.d. “Grêmio1983.” Grêmio1983. Accessed April 11, 2023. https://gremio1983.wordpress.com.

Lemos, Carlos, Rodrigo Breves, and Leandro Silva. 2020. “Público Nos Estádios Do Brasil Em 2020.” GloboEsporte.com. March 23, 2020. http://app.globoesporte.globo.com/futebol/publico-no-brasil/index.html.

“Mercado de Transferências, Rumores, Valores de Mercado E Notícias.” n.d. Www.transfermarkt.com.br. https://www.transfermarkt.com.br.

Netto, Paulo. n.d. “Estatísticas E Jogos Do Clube de Regatas Do Flamengo.” Flaestatistica.com.br. Accessed April 11, 2023. https://flaestatistica.com.br.

Online, Perspectiva. 2008. “Média de Público Final Do Brasileirão 2008.” PERSPECTIVA ONLINE. December 10, 2008. https://perspectivabr.wordpress.com/2008/12/10/media-de-publico-final-do-brasileirao-2008/.

Rodrigues, Rodolfo. 2017. “Corinthians: Maior Média de Público Desde 1993.” UOL Esporte. June 27, 2017. https://futebolemnumeros.blogosfera.uol.com.br/2017/06/27/corinthians-maior-media-de-publico-desde-1993/.

“Verdazzo - O Site Da Torcida Palmeirense.” 2017. November 8, 2017. https://www.verdazzo.com.br.

Wong, Alex. 2017. “A High-Scoring Revolution Has the Rockets Soaring.” The Atlantic. November 30, 2017. https://www.theatlantic.com/entertainment/archive/2017/11/a-high-scoring-revolution-has-the-rockets-soaring/547103/.