Solano Caffarena

November 25, 2023

Argentinian Soccer League: Binance Cup 2022

Introduction:

Soccer is one of the most-watched sports in the world, with over 3.5 billion fans worldwide (World Atlas, 2023). England was the first country that create a professional association back in 1863. In Argentina, soccer is the most popular sport. Families and friends create plans around the soccer schedule so they can hang out with friends while watching their team. Watching your team at a coffee shop with friends is an event that many people do. Soccer's popularity has kept a lot of businesses open.

Apart from that, Argentina has always been recognized for its soccer players. For instance, Lionel Messi, who won 8 Ballon d'Or (The most recognized individual award), Angel DiMaria, and Diego Maradona, among others. Argentina won the FIFA World Championship last year. The goal of this paper is to attempt to ascertain what factors need to be present to win a soccer match. This paper will focus on the question "How does scoring the first goal change the probability of winning the match?". Also, “Is a team more likely to win more matches at home? This question was always around, and many soccer managers would rather play at home than away. But is the outcome of the match going to change depending on this condition?

The Argentine first-division soccer league, Copa de La Liga, is the league we will be analyzing. There are 28 teams in it. In terms of structure, there were two groups in the tournament, each with 14 clubs. There were thirteen matches overall, with each team playing every other team in each group. The Argentine Football Association's (AFA) official website has all the data used in this study. I gathered every data point related to all the games in 2022 from

the official page. Also, the independent variables that I created to conclude this research were taken from this website as well.

Articles:

The paper "Football is becoming more predictable; network analysis of 88 thousand matches in 11 major leagues" written by Victor Martins Maimone and Taha Yasseri, talks about a study from 1993 to 2019. The study collected 87,000 data points (soccer matches). The researchers omitted tied matches for simplicity**.** Another finding from Martins and Yasseri is that matches are more predictable than before because not all teams have the same economic resources. Good soccer players can make millions, and not many clubs can afford their salaries, leading to an economic gap between clubs that can recruit high-skilled players and those that cannot. In the article, we can see they have used a Gini coefficient, which is a statistical measure of income inequality. Besides that, The Area Under the Curve (AUC) shows the probability of random outcomes. We can see in the graphs below that in countries where soccer is highly competitive like England, Germany, Portugal, and Spain, where the most recognized players are, the Gini coefficient has been increasing, also the AUC line has been increasing as well. As this gap of inequality is increasing, the authors state that soccer is becoming more predictable. Also, they have presented a table with the p-values of these models. We can see that in many countries the p-value is below the typical 5% significance level. If we check the column "Gini: t" we can see that the Gini coefficient is highly significant in countries like England. They argue that richer leagues are more likely to have higher predictability. The column AUC:t tells us that the expected average value of AUC is the same for both parts of the sample. Or if there is a significant difference in the average of the AUC values in the sample. A lower p-value will state that there is a significant difference in the average of the AUC so we will reject the null

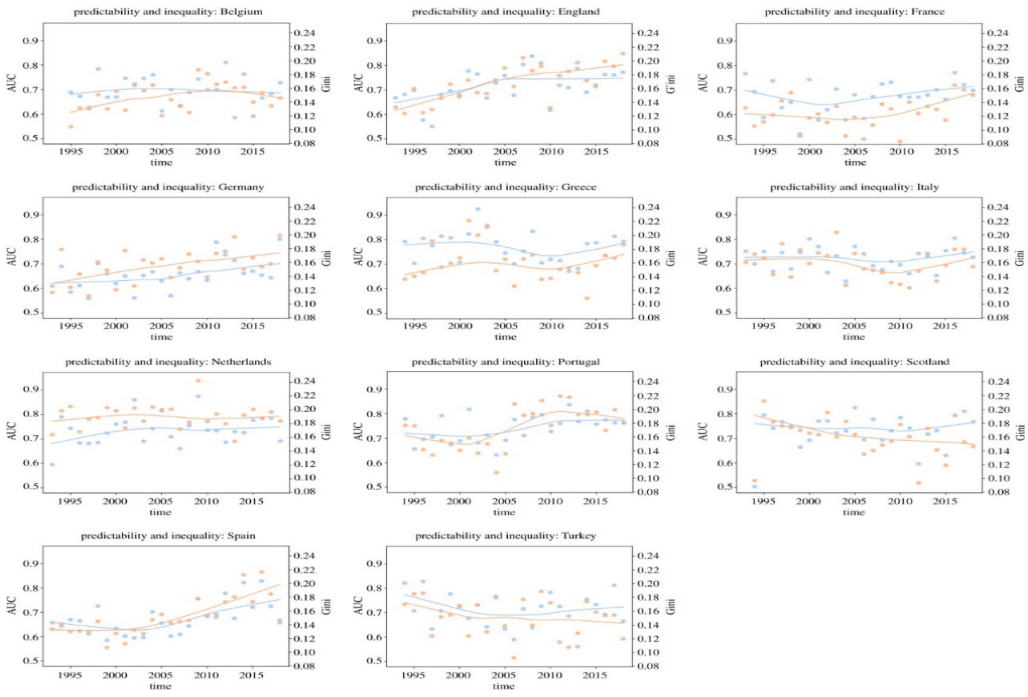
hypothesis. Again, England, one of the biggest and most watched leagues in the world rejects the null hypothesis. But in this column, Portugal, Spain, and Germany also reject the null hypothesis at a significance level of 5%.

Table 1:

country	AUC: KS	AUC: *t*	Gini: KS	Gini: *t*	ELO-AUC: KS	ELO-AUC: *t*
Belgium	0.9985	0.7383	0.2558	0.0752	0.8690	0.4805
England	0.1265	*0.0330*	*0.0126*	*0.0009*	*0.0443*	*0.0133*
France	0.1265	0.1437	0.2999	0.3426	0.1265	0.0895
Germany	0.1265	*0.0326*	0.2999	*0.0367*	0.5882	0.0787
Greece	0.2558	0.0634	0.5361	0.1190	*0.0079*	*0.0019*
Italy	0.2999	0.7370	*0.0443*	0.0547	0.9992	0.8292
Netherlands	0.8978	0.7371	0.5882	0.9488	0.8978	0.8320
Portugal	*0.0079*	*0.0044*	*0.0000*	*0.0000*	*0.0314*	*0.0071*
Scotland	0.9985	0.9128	*0.0314*	0.0771	0.5361	0.2864
Spain	*0.0443*	*0.0097*	*0.0005*	*0.0001*	*0.0029*	*0.0010*
Turkey	0.9985	0.6429	0.8690	0.4610	0.2558	0.3108

Table 1, https://royalsocietypublishing.org/doi/10.1098/rsos.210617

Figure 1



The blue line and dot correspond to the AUC, and the orange lines and dots correspond to the Gini coefficient (https://royalsocietypublishing.org/doi/10.1098/rsos.210617 Figure 1)

Apart from that, the research found that home teams' advantage has decreased since the 90s. They have created graphs about the English league where the number of points that the home and away teams scored each season from 1995 to 2015. They have talked about an article from Richard Pollard where he argues that factors like fans, travel effects; familiarity with the field, and referee bias, cause the home team to have a greater advantage over the visitor team. On the Taylor and Francis website, Pollard states that soccer teams win 64% of their points at home. The author states "Home advantage has changed very little since the formation of the League in 1888 and there are only small variations between the four Divisions of the League". The graphs created by Maimone and Yasseri show a trend that the home factor is decreasing.

Another study concluded by Aravind Surumpudi analyzes soccer with a multiple regression model. In this example, he collected 102 soccer matches. But it was not the Argentinian or European league. Instead, he gathered data from FIFA matches, which are between countries rather than clubs. These games were exclusive to World Cup competitions, as these events provide greater motivation for professional soccer players.

Surumpudi analyzed different independent variables. He included: T1-Shots on target, T1-Shots, T2-Shots on target, T2-Shots, T1-Posession %, T2-Posession %, T1-Fouls, T2-Fouls, T1-Red Cards, T2-Red Cards. With these variables, he created three separate models, one with six variables that include shots, shots on target, and shot %. The second and third models only included four variables. The first one is a non-comparison model which includes Shot %, Possession %, Fouls, and red cards. In the last model, he compares both teams, the winning and losing teams, using the factors of Shot %, Possession %, Fouls, and red cards. Then, Aravind decided to use the 4-factor comparison model since it gave him the highest R-squared.

Once he processed the data, the model got an R-squared of 25% approximately, the result was not very informative since 75% of the variability is still not explained by the model. Surumpudi argued that “these may be factors that cannot be recorded, such as team morale, whether the team was home or away, weather, etc.”. Finally, Surumpudi ran a simple regression model with each variable of the 4-factor comparison model to check the R-squared of each variable separately. The R-squared values for possession, shots, fouls, and red cards were 12.09%, 7.61%, 3.63%, and 1.01%, respectively. This shows that possession is the most crucial factor in determining a soccer game's outcome out of these four factors.

The equation for Surumpudi model is:

$$Score = 0.1209 * Possession + 0.0761 * Shot\% + 0.0363 * Fouls + 0.0101 * Red\ cards.$$

Where:

- *“Possession”* is the amount of time the team T1 had the ball during the match compared to T2. (You get this value by subtracting T1 possession and T2 possession)
- *“Shot%”* is the difference between total kicks in the whole match and the kicks from T1 compared to T2.
- *“Fouls”* is the number of faults from T1 minus the number of faults from T2.
- “*Red cards”* is the number of red cards from T1 minus the number of faults from T2.

Model development:

The variables presented in this model were obtained from the studies and literature reviewed for this project. This model tries to explain the questions from above using variables

that were significant in other models and applying new ones that I believe can explain the variability of the data.

An example of the data is shown below.

Game	HomeT	AwayT	WinTeam	Possession_H	Possession_V	Goal_H	Goal_V	FirstGoal	YellowCards_H	YellowCards_V	RedCards_H	RedCards_V	Penalties_H	Penalties_V
1	Barracas	Central C	N	44.9	55.1	1	1	H	2	5	0	0	1	0
1	Independiente	San Lorenzo	N	46.2	53.8	1	1	A	1	2	1	0	0	0
1	At Tucuman	Colon	N	53.3	46.7	1	1	H	5	4	1	0	0	0
1	Banfield	Newells	A	59.8	40.2	1	2	A	1	3	0	0	1	0
1	Patronato	Velez	N	39.1	60.9	1	1	A	1	2	0	0	1	0
1	Racing	Huracan	H	62.6	27.4	2	0	H	5	4	0	1	0	0
1	Platense	Godoy Cruz	H	51.5	48.5	0	1	A	2	2	0	0	2	0
1	Talleres	Sarmiento	H	58.2	41.8	2	0	H	3	4	0	1	0	0
1	Union	Tigre	A	53	47	1	2	H	2	2	0	0	0	0
1	Estudiantes	Gimnasia	N	52.9	47.1	1	1	H	2	3	0	0	0	0
1	Boca	Arsenal	H	63.2	36.8	2	1	H	2	2	0	0	0	0
1	Defensa	River	N	35.9	64.1	0	0	N	3	2	0	0	0	0
1	Ros Central	Lanus	N	38.3	61.7	0	0	N	2	2	1	0	0	0
1	Argentino	Aldosivi	H	50.4	49.6	2	1	H	1	3	0	0	0	1
2	Newells	San Lorenzo	N	50.6	49.4	0	0	N	1	4	1	0	0	0
2	Gimnasia	Patronato	H	46.2	53.8	2	0	H	3	4	0	1	1	0
2	Aldosivi	Estudiantes	A	51.6	48.4	0	1	A	4	0	1	0	0	1
2	Lanus	Defensa	N	63.3	36.7	1	1	H	2	1	0	1	0	0
2	Huracan	Ros Central	H	42.7	57.3	2	0	H	2	6	0	0	0	0
2	Independiente	Talleres	H	39.8	60.2	1	0	H	2	2	0	0	0	0

However, because Possession_H and Possession_V have perfect collinearity we decided to drop Possession_V. Also, the transformation of the data was one of the main issues. There were two dummy variables, *FirstGoal* and *WinTeam,* that had three different values. Because of this, we decided to exclude tight matches. After doing this, the data became smaller, with 263 observations. As you can see in the picture below, WinTeam and FirstGoal variables only have zeros and ones now.

	Game	HomeT	AwayT	WinTeam	Possession_H	Possession_V	Goal_H	Goal_V	FirstGoal	YellowCards_H	YellowCards_V	RedCards_H	RedCards_V	Penalties_H
1	1	Banfield	Newells	0	59.8	40.2	1	2	0	1	3	0	0	1
2	1	Racing	Huracan	1	62.6	27.4	2	0	1	5	4	0	1	0
3	1	Platense	Godoy Cruz	1	51.5	48.5	0	1	0	2	2	0	0	2
4	1	Talleres	Sarmiento	1	58.2	41.8	2	0	1	3	4	0	1	0
5	1	Union	Tigre	0	53.0	47.0	1	2	1	2	2	0	0	0
6	1	Boca	Arsenal	1	63.2	36.8	2	1	1	2	2	0	0	0
7	1	Argentino	Aldosivi	1	50.4	49.6	2	1	1	1	3	0	0	0
8	2	Gimnasia	Patronato	1	46.2	53.8	2	0	1	3	4	0	1	1
9	2	Aldosivi	Estudiantes	0	51.6	48.4	0	1	0	4	0	1	0	0
10	2	Huracan	Ros Central	1	42.7	57.3	2	0	1	2	6	0	0	0
11	2	Independiente	Talleres	1	39.8	60.2	1	0	1	2	2	0	0	0
12	2	Sarmiento	Argentino	1	27.6	72.4	1	0	1	3	1	0	0	0
13	2	Velez	Platense	0	63.0	37.0	0	1	0	1	1	1	0	0
14	2	Godoy Cruz	Racing	1	34.9	65.1	2	0	1	4	1	1	1	0
15	2	Central C	Boca	1	30.1	69.9	1	0	1	3	0	0	0	0
16	3	Patronato	Aldosivi	1	37.5	62.5	1	0	1	5	3	0	1	0
17	3	Talleres	Newells	0	59.3	40.7	0	1	0	4	2	1	0	0
18	3	Estudiantes	Sarmiento	1	65.4	34.6	2	1	1	3	5	1	0	1

The model is about the dependent variable WinTeam. All the independent variables will be used to explain the regressor. The model is shown below:

$$WinTeam = Possession\_H + Goal\_V + Goal\_H + FirstGoal + YellowCards\_V + RedCards\_V + FirstGoal + Penalties\_V + + Penalties\_H + u.$$

Where:

WinTeam is the regressed variable. It is one when the home team from the sample has won and zero when the away team won.

Possession_V is the percentage of possession from that team.

Goal_V is the number of goals scored by the away team.

YellowCards_V is the number of yellow cards obtained by the away team.

RedCards_V is the number of red cards obtained by the away team.

FirstGoal is the team that makes the first goal. It could be 1, meaning home team, 0 away team.

Penalties_V is the number of penalties that the away team had.

U is the error term.

Model application:

The result of our model is expressed below.

```
Call:
lm(formula = WinTeam ~ Possession_H + Goal_H + Goal_V + YellowCards_V +
    YellowCards_H + FirstGoal + RedCards_H + RedCards_V + Penalties_H +
    Penalties_V, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.57365 -0.12915 -0.02074  0.12269  0.77929

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.5644350  0.0751091   7.515 9.94e-13 ***
Possession_H  -0.0042290  0.0011550  -3.661 0.000306 ***
Goal_H         0.1790281  0.0138755  12.902  < 2e-16 ***
Goal_V        -0.1710396  0.0161277 -10.605  < 2e-16 ***
YellowCards_V  0.0018302  0.0090145   0.203 0.839277
YellowCards_H -0.0005622  0.0095164  -0.059 0.952942
FirstGoal1     0.3236582  0.0395840   8.176 1.44e-14 ***
RedCards_H    -0.0539757  0.0350606  -1.539 0.124937
RedCards_V     0.0372935  0.0286075   1.304 0.193551
Penalties_H    0.0212871  0.0305106   0.698 0.486010
Penalties_V    0.0122401  0.0371485   0.329 0.742057
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2006 on 252 degrees of freedom
Multiple R-squared:  0.8358,    Adjusted R-squared:  0.8292
F-statistic: 128.2 on 10 and 252 DF,  p-value: < 2.2e-16
```

As we can see, our model includes all the variables from above but not all of them have the same impact on the regressed variable WinTeam. If we look at the p-value of Possession_H, Goal_H, Goal_V and FirstGoal1, we can see that their values are below the significance level of 5%. This means that these are the variables that can explain better the variability of WinTeam than the rest of the data. One of the relevant coefficients is Possession_H since it has a negative value, meaning that as Possession_H increases WinTeam decreases by 0.0042290. Also, the p-value is 0.000306 which is below the typical 5% significance level. Surumpudi's model concluded that possession was his most informative variable, and we can see that, at this point, we got a negative coefficient of -0.003 meaning that one increase in possession will decrease WinTeam by 0.003.

```
Call:
lm(formula = WinTeam ~ Possession_V + Goal_H + Goal_V + YellowCards_V +
    YellowCards_H + FirstGoal + RedCards_H + RedCards_V + Penalties_H +
    Penalties_V, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.57339 -0.12937 -0.01906  0.12277  0.77978

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.1420591  0.0721234   1.970 0.049972 *
Possession_V   0.0042084  0.0011491   3.662 0.000304 ***
Goal_H         0.1790886  0.0138754  12.907  < 2e-16 ***
Goal_V        -0.1711305  0.0161274 -10.611  < 2e-16 ***
YellowCards_V  0.0018787  0.0090151   0.208 0.835093
YellowCards_H -0.0003511  0.0095191  -0.037 0.970604
FirstGoal1     0.3237087  0.0395812   8.178 1.42e-14 ***
RedCards_H    -0.0542308  0.0350712  -1.546 0.123285
RedCards_V     0.0378815  0.0286191   1.324 0.186821
Penalties_H    0.0210653  0.0305087   0.690 0.490535
Penalties_V    0.0121650  0.0371474   0.327 0.743577
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2006 on 252 degrees of freedom
Multiple R-squared:  0.8358,    Adjusted R-squared:  0.8292
F-statistic: 128.2 on 10 and 252 DF,  p-value: < 2.2e-16
```

Talking about the other variables, FirstGoal has a coefficient of 0.3236582. This indicates that if FirstGoal is 1 (meaning the team scored the first goal), the predicted value of WinTeam

increases by approximately 0.3237, ceteris paribus. The coefficient for Goal_V is -0.1710396. As Goal_V increases by 1 unit, the predicted value of WinTeam decreases by approximately 0.171, holding other variables constant. The coefficient for Goal_H is 0.1790281. This suggests that for a one-unit increase in Goal_H, the predicted value of WinTeam increases by approximately 0.179, ceteris paribus. Then if we check the p-values from the rest of our variables, we will see that they do not contribute to the explanation of our regressed variable. Apart from that, the R-squared of the model is 0.83. This means that 83% of the variability on WinTeam is explained by our model. Also, the p-value is highly significant below the 1% level. One of the questions that we wanted to address in this paper was whether the home team had some advantage. The coefficient of Goal_H is bigger than the Goal_V. This implies that a goal scored by the home team has a greater impact on the probability of winning than a goal scored by the visiting team. A positive coefficient indicates that the proportional chances of the home team winning increases together with the home team's goals. The total number of goals scored in the league was 645. Which, 269 were scored away and 376 were scored at home. This indicates that 58% of the total goals were scored in home matches. Probably, there is a positive advantage for home teams, but we cannot argue with Maimone and Yasseri since their study has more evidence than one tournament with 378 matches.

The next step was to create a restricted model with only significant variables.

$$WinTeam = Possession\_H + Goal\_H + Goal\_V + FirstGoal + u$$

The model gave us the following output.

```
lm(formula = WinTeam ~ Possession_H + Goal_H + Goal_V + FirstGoal,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55519 -0.13717 -0.01893  0.12737  0.82414

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.544544   0.069132   7.877 9.37e-14 ***
Possession_H -0.003834   0.001132  -3.388 0.000813 ***
Goal_H        0.182061   0.013627  13.360  < 2e-16 ***
Goal_V       -0.171224   0.015930 -10.748  < 2e-16 ***
FirstGoal1    0.329381   0.039138   8.416 2.71e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2 on 258 degrees of freedom
Multiple R-squared:  0.8329,    Adjusted R-squared:  0.8303
F-statistic: 321.4 on 4 and 258 DF,  p-value: < 2.2e-16
```

As we can see, the p-value and R-squared of the new model are almost the same. There is no need to keep the variables from the previous model. But to be sure about this hypothesis, the ANOVA test will tell us if we need to keep those variables.

```
Model 1: WinTeam ~ Possession_H + Goal_H + Goal_V + YellowCards_V + YellowCards_H +
    FirstGoal + RedCards_H + RedCards_V + Penalties_H + Penalties_V
Model 2: WinTeam ~ Possession_H + Goal_H + Goal_V + FirstGoal
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    252 10.139
2    258 10.318 -6  -0.17887 0.741 0.6171
```

The results of this ANOVA test reinforce that there is no significant difference in fit between the restricted and unrestricted models. The p-value helps us to check whether the inclusion of additional variables in Model 1 significantly improves the explanation of the regressed variable compared to the restricted Model 2. The p-value (0.6171) is greater than the commonly used significance level of 0.05. Therefore, we would not reject the null hypothesis. There is no strong evidence to suggest that Model 2, is significantly worse in explaining the dependent variable compared to Model 1.

After, I created another model including two more variables. The variables are the fitted values of our previous regression (ols2). Creating these two variables helps our model to capture more data points that are non-linear with our independent variables, higher-order terms create a larger number of smooth shapes that help us to explain non-linear data.

```
Call:
lm(formula = WinTeam ~ Possession_H + Goal_H + Goal_V + FirstGoal +
    y_sq + y_cb + y_4 + y_5, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32792 -0.04292  0.00544  0.02841  0.95251

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2096354  0.0557715  -3.759 0.000212 ***
Possession_H  0.0030677  0.0007228   4.244 3.07e-05 ***
Goal_H        0.0001075  0.0184277   0.006 0.995351
Goal_V       -0.0089455  0.0160123  -0.559 0.576884
FirstGoal1   -0.4865094  0.0529666  -9.185  < 2e-16 ***
y_sq          3.2895840  0.3938464   8.352 4.38e-15 ***
y_cb          2.2142783  1.1784372   1.879 0.061390 .
y_4          -6.4242686  1.1523220  -5.575 6.31e-08 ***
y_5           2.5214213  0.3587950   7.027 1.93e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1033 on 254 degrees of freedom
Multiple R-squared:  0.9561,    Adjusted R-squared:  0.9547
F-statistic: 691.8 on 8 and 254 DF,  p-value: < 2.2e-16
```

Now, if we check the p-value from Possession_H, we can observe a smaller p-value. It is under the 1% level now. Also, we can see that the R-squared of our model increased to 95% by including these 4 variables. The overall p-value from our model is strong meaning that the new variables explain more our dependent variable. Also, one of the variables that were significantly important for us, ‘FirstGoal’, became negative in this model. One of the questions that was proposed in this paper was if scoring the first goal has any significance in the output of the match. We can observe in the models from above that FirstGoal is highly significant in

predicting the WinTeam. However, in our first model, it had a positive coefficient and in our last model became negative. So, there is no strong evidence that FirstGoal is a determinant factor in predicting a soccer match in the Argentinian League of 2022.

Then, a RESET helped to test if we needed to include more non-linear terms in our model (ols3). In the first RESET, the RESET Statistic result was 205.08 as shown below. This means that the fitted values are far from the real values of our data. However, the p-value is highly significant at 2.2×10^{-16}. The p-value tells us that we should reject the null hypothesis since the significance of our dependent variables is strong.

```
        RESET test

data:  ols3
RESET = 205.08, df1 = 3, df2 = 253, p-value < 2.2e-16
```

The last ANOVA test, which compares the new OLS model, which includes fitted values raised to the second, third, fourth, and fifth powers, with the excluded model, that previously gave us the best fit. The null hypothesis was rejected since the p-value is less than the 1% significance level. This outcome suggests that there is sufficient evidence to conclude that at least one of the recently added variables, which are created by exponentiating the fitted values, makes a substantial contribution to the explanation of the variance in the regressed variable.

```
Analysis of Variance Table

Model 1: WinTeam ~ Possession_V + Goal_H + Goal_V + FirstGoal
Model 2: WinTeam ~ Possession_H + Goal_H + Goal_V + FirstGoal + y_sq +
    y_cb + y_4 + y_5
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    258 10.3197
2    254  2.7088  4    7.6109 178.42 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After completing the first RESET, we checked if increasing our exponential variables were going to explain better our model, so I created another OLS model and included the fitted values from our OLS2 model at the 2nd, 3rd, 4th, and 5th powers to verify if our model was not able to explain the data because of non-linearity problems. The last RESET, with a p-value of less than 1% level provides support for the presence of omitted variables or nonlinearity in our model. In other words, the current model structure may not adequately capture the non-linear relationships in the data and there may be additional explanatory variables or a nonlinear relationship that could improve the model's fit.

```
        RESET test

data:  ols4
RESET = 113.99, df1 = 3, df2 = 251, p-value < 2.2e-16
```

To address possible heteroskedasticity problems, a Breusch-Pagan test was run on the OLS model 4. The goal was to look for signs of heteroskedasticity, which is a situation where the variance of the errors varies with the independent variable levels.

The results were:

```
        studentized Breusch-Pagan test

data:  ols4
BP = 11.593, df = 8, p-value = 0.1703
```

The BP result is not enough to state that Heteroskedasticity is present in our model. The p-value is above the 10% level, so we fail to reject the null hypothesis, meaning that there is not sufficient evidence of Heteroskedasticity in the model ols4.

Conclusion:

To sum up, this study examined several variables impacting soccer match results, with a special focus on the Copa de La Liga, the first-division soccer league in Argentina, in 2022. This essay examined the effects of possession from the home team, first goal scoring, and team attributes on match outcomes by examining data from official sources and deriving conclusions from prior research. Regression models, RESET tests, and ANOVA tests were used in the statistical analysis to examine the models' suitability and relationships. The results indicated that team goals, possession, and first-goal scoring all played a substantial role in explaining the variation in the game results. Our most significant variables in the last model (ols4) were y_sq + y_cb + y_4 + y_5. This shows that our data has non-linear relationships with the independent variables.

The investigation also addressed the presence of possible heteroskedasticity. Even though the Breusch-Pagan test did not reveal any evidence of heteroskedasticity in the model, it is still important to carefully review the model specifications and any potential missing variables. The complexity of the variables affecting game outcomes is highlighted by this research, which also advances our understanding of soccer match behavior.

Work cited:

Soccer Is Getting Too Predictable. This Math Proves It - Popular Mechanics, www.popularmechanics.com/science/math/a39503543/math-proves-soccer-is-getting-predictable/. *Accessed 10 Dec. 2023.*

Un-Building: A Utopia of Receding Construction - Taylor & Francis Online, www.tandfonline.com/doi/full/10.1080/13602365.2023.2242876?scroll=top&needAccess=true. *Accessed 10 Dec. 2023.*

Richard Pollard (1986) Home advantage in soccer: A retrospective analysis, Journal of Sports Sciences, 4:3, 237-248, DOI: 10.1080/02640418608732122 tandfonline.com/doi/epdf/10.1080/02640418608732122?needAccess=true

Football Is Becoming More Predictable; Network Analysis of 88 Thousand ..., royalsocietypublishing.org/doi/10.1098/rsos.210617. *Accessed 10 Dec. 2023.*

"The Most Popular Sports in the World." WorldAtlas, 13 Sept. 2023, www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html.

"Reasons Why Soccer Is the Most Popular Sport in the World." KU Sports, www2.kusports.com/sponsored-content/2023/apr/20/reasons-why-soccer-is-the-most popular-sport-in-the-world/. *Accessed 10 Dec. 2023.*

Football Is Becoming More Predictable; Network Analysis of 88 Thousand ..., royalsocietypublishing.org/doi/10.1098/rsos.210617. *Accessed 10 Dec. 2023.*

Wilson, Bill. "European Football Is Worth a Record £22bn, Says Deloitte." BBC News, BBC, 6 June 2018, www.bbc.com/news/business-44346990.

Surumpudi, Aravind. "What Is the Most Important Factor to Winning a Soccer Game?" Medium, Medium, 8 Jan. 2019, medium.com/@ar.surumpudi/what-is-the-most-important-factor-to-winning-a-soccer-game-129a4a2c1d2c.

Home Advantage in Football: A Current Review of an Unsolved Puzzle, opensportssciencesjournal.com/contents/volumes/V1/TOSSJ-1-12/TOSSJ-1-12.pdf. *Accessed 10 Dec. 2023.*