Investigating the impacts of COVID-19 on team wins and social media engagement in the EPL and Manchester United*

Swarnadeep Chattopadhyay

26 April 2022

Abstract

The recent developments of the COVID-19 pandemic has had a tremendous impact on the world. Albeit school, work or daily life, there has been changes to the way that we conduct our daily lives. While a lot has changed, the intensity of the major soccer competitions all around the world has not. With teams playing on a regular schedule with advanced COVID protocols, we wanted to assess if teams in the English Premier League have been severely impacted by the change in environment. This report investigates the impact of COVID from the standpoint of team wins and team engagement in the English Premier League(EPL) and uses Manchester United which is a club playing the league to prove the point. R Core Team (2020) is the software used to perform this analysis.

Contents

1 Introduction **2**

2 Scope of the Report **2**

3 Data Collection **2**

4 Exploratory Data Analysis **2**

4.1 Team Wins Data Analysis (1st Question) 3

4.2 Twitter Engagement Analysis for Manchester United (2nd Question) 3

5 Results and Methodology **4**

5.1 Methodology 4

5.2 Aanlytic 1: Results from Team Wins affected by COVID 4

5.2.1 Pythagorean Analysis 5

5.3 Visual 1: A comparison between Pre COVID and During COVID Win Rates 7

5.4 Analytic 2: Twitter Engagement 8

5.4.1 Pre Covid Engagement 8

5.4.2 During COVID Engagement 8

5.4.3 Difference in Average Engagements vs Predicted Engagement 8

5.5 Visual 2: Change in Engagement vs Change in Actual Win% 9

*Code and data are available at: https://github.com/WanteXecutioN/Final_Project.

6 Conclusion **10**

References **11**

1 Introduction

The English Premier League is renowned to be one of the most watched soccer leagues in the world. To date, the league is broadcasted in over 200 countries which spans over 4 billion people (Richards (2020)). With such a large market, teams in the EPL are continually able to attract some of the best players in the world. This is pivotal because, not only will the league be attracting more engagement, but the level of competition will be more intense. At the start of 2020, the COVID-19 pandemic had halted a lot of major activities around the world. With the virus spreading quick, the organizers of the EPL halted the premier league football between March to June. After the initial suspension of the games, teams were allowed to play again but behind closed doors meaning no attendance and stricter protocols. We chanced upon an article on ESPN, that essentially highlights the loss of home advantage arising from games with no fans. Pivotally, they pointed out that Liverpool, who were the 2019-2020 season winners, had lost 6 games at home during the COVID era. Prior to that, they were on a 68 games win streak at home (Hamilton (2021)). We use Wickham et al. (2019), (**citegrid?**), (**citeExtra?**), Rudis (2020), Delignette-Muller and Dutang (2015), Wickham and Bryan (2022) and Harrell Jr (2021) to prove how Covid has affected teams in the EPL both on and off the field.

2 Scope of the Report

We wanted to assess what we deem here, as the "COVID Effect," while building on the fundamental impact that COVID has had on teams in the EPL. With that, our research questions are:

```
1) How has the pandemic affected team wins for English Premier League teams?
2) How has the pandemic affected Twitter engagements for Manchester United?
```

Central to this investigation, we have chosen to use Manchester United because they are to date, the 4th most valuable club in the world at US$4.2 Billion. In the EPL, they are currently the most valuable club. A big part of the reason behind this valuation is, in the early 2000's, Manchester United had been one of the most dominant teams in the EPL, consistently ranking 1st in the league. In that time span, they have generated a large following in person and on social media. The team has fans living all around the world who constantly contribute by purchasing the team's merchandise.

3 Data Collection

In terms of collecting the data, there were numerous websites used to conduct this paper. Firstly, the 2019-2020 season data was generated from "2019-2020 Premier League Stats" (n.d.).

4 Exploratory Data Analysis

This investigation will consist of 5 data sets and focus on teams in Premier League (in England). To assess the first question, we will be working with data sets 1-2. For reference, the 2019/2020 season will be associated with Pre-Covid period and 2020/2021 season will be associated with COVID period and the average season has 38 games.

4.1 Team Wins Data Analysis (1st Question)

On initial observation, both data sets consists of 20 teams and 15 variables. Some key observations from the 2019/2020 season, the average goals scored by a premier league team playing at home is: 51.70 or 1.36 goals per game. With the 1.36 goals per game, the average number of wins stands at 14 which constitutes a 37% win rate (not very good). Mean attendance for all 20 teams is: 29796.

With the 2020/2021 season, the average goals scored by home team stands is: 38 goals or 1.31 goal per game. With the 1.31 goal per game, the average number of wins is 38%! Mean attendance in the 2020/2021 season stands at a mere 82. Note here that, the 2020/2021 season is not finished, the average wins and average goals were computed over 29 games. (Change)

Initial comparison, in spite of the discrepancy in the unequal number of games, there are two observations to point out:

1) The win rate in the 2020/2021 season has improved by 1% from 37% to 38%.
2) The goals scored per game by home team has decreased by 0.05 from 1.36 to 1.31

A 99.7% decrease in attendance has attributed to a 1% increase in team wins but a 0.05 decrease in goals scored per game!

In order to run a logistic regression, the response variable needs to be of value 0 <y<1. We need to obtain a new response variable consisting of Wins. To do so, we will create new variable: Prop, which is the number of wins over matches played. Given that a team can't possibly win greater than 38 games, we are assured that our response will be less than 1. Another key change was to assign attendance according to: "High," "Medium" and "Low" which we have assigned according to the upper, median, and lower quantile of attendance variable. Same can be said for the 2020/2021, which we will assign "none" and "Yes" which by its name, represents teams with no fans and teams with some fans.

4.2 Twitter Engagement Analysis for Manchester United (2nd Question)

3rd data set is the twitter engagement data for Manchester United. To clear up the data to account for periods of COVID, we have divided the data set into 3.

```
1) Pre Covid: July - December 2019
The month with the highest engagement was in December, which is 5,019,237. In retrospect,
the month with the lowest engagement is in September, which is, 3,031,491.

2) During Covid from 2019/2020 season: Jan- June 2020
The month with the highest engagement is June 2020 at 7,479,142. The low was during April
2020 at, 2,982,343

3) During COVID from 2020/2021 season: July- December 2020
The month with the highest engagement is July 2020, at 9,546,656 with the low during
November, at 4,908,278
```

Within a year of the pandemic, there is a reasonable difference in the total engagements by month. Specifically, the highest total engagement by month has increased significantly by 90%. This statistic does make sense considering that a lot now EPL watcher's are working from home, this means that they have more freedom and time to engagement in various social media.

5 Results and Methodology

5.1 Methodology

To answer the first question, we will aim to predict the expected wins of the teams using logistic regression. The dependent variable will be "prop" which is the proportion of Wins with respect to MP (formula = W/MP). The key differentiate of the two data set is trivially the huge loss of the fans, we want to include that as a key variable along with variables: GF, GA, Points + attendance. Here attendance is mutated to be categorical variables (low,medium and high) which we assigned according to the summary statistics (25%, 50%, 75%). On top of the regression, we will also use the Pythagorean formula to predict wins given the number of goals scored. By the end, we will assess the validity of both predictions.

For the second question, we will aim to assess the key variables affecting twitter engagement of the team. Here obviously, we regress with dependent variable being Total Engagements that is subdivided into each day of every month in the constituent periods listed in 3.2. And key independent variables being: Number of Posts , Month/Year , Day, Page Likes (number of page likes on a given day). For the bridge of the assignment, we will compute the summary wins for Manchester United from the 4th and 5th data set and create a data frame involving average engagement and average wins. This data frame allows us to create a visualization linking average wins to the average engagement.

5.2 Aanlytic 1: Results from Team Wins affected by COVID

```
##   Rk           Squad MP  W  D  L  GF GA GD Pts   xG  xGA  xGD xGD.90 Attendance
## 1  1       Liverpool 38 32  3  3  85 33 52  99 71.5 40.0 31.5   0.83      41955
## 2  2 Manchester City 38 26  3  9 102 35 67  81 93.0 34.7 58.3   1.53      37097
## 3  3  Manchester Utd 38 18 12  8  66 36 30  66 59.4 37.4 22.0   0.58      57415
## 4  4         Chelsea 38 20  6 12  69 54 15  66 66.6 37.9 28.6   0.75      32023
## 5  5  Leicester City 38 18  8 12  67 41 26  62 61.6 44.5 17.1   0.45      25312
## 6  6       Tottenham 38 16 11 11  61 47 14  59 46.1 52.0 -6.0  -0.16      43757
##   attendance      prop ExpectedWins    ExtraWins EW diff
## 1       High 0.8421053    0.8288912  0.013214015 31    1
## 2     Medium 0.6842105    0.6806827  0.003527783 26    0
## 3       High 0.4736842    0.4822787 -0.008594526 18    0
## 4     Medium 0.5263158    0.5482127 -0.021896887 21   -1
## 5     Medium 0.4736842    0.4595705  0.014113698 17    1
## 6       High 0.4210526    0.4254338 -0.004381155 16    0
```

```
##   Rk           Squad MP  W  D  L GF GA GD Pts   xG  xGA  xGD xGD.90 Attendance
## 1  1 Manchester City 38 27  5  6 83 32 51  86 73.3 31.4 42.0   1.10        526
## 2  2  Manchester Utd 38 21 11  6 73 44 29  74 60.2 42.2 18.0   0.47        526
## 3  3       Liverpool 38 20  9  9 68 42 26  69 72.6 45.3 27.3   0.72        837
## 4  4         Chelsea 38 19 10  9 58 36 22  67 64.0 32.8 31.2   0.82        526
## 5  5  Leicester City 38 20  6 12 68 50 18  66 56.0 47.7  8.3   0.22        421
## 6  6        West Ham 38 19  8 11 62 47 15  65 53.9 48.3  5.6   0.15        632
##   attendance1      prop ExpectedWins1   ExtraWins1 EW diff PREDWIN
## 1        Some 0.7105263     0.7017707  0.008755639 27    0      27
## 2        Some 0.5526316     0.5976109 -0.044979293 23   -2      23
## 3        Some 0.5263158     0.5395330 -0.013217248 21   -1      21
## 4        Some 0.5000000     0.5220472 -0.022047184 20   -1      20
## 5        Some 0.5263158     0.5151775  0.011138301 20    0      20
## 6        Some 0.5000000     0.5083774 -0.008377420 19    0      19
```

```
## # A tibble: 3 x 2
```

```
##   attendance    TW
##   <chr>      <int>
## 1 High          90
## 2 Low           48
## 3 Medium       150
```

```
## # A tibble: 2 x 2
##   attendance1    TW
##   <chr>       <int>
## 1 Some          270
## 2 Very Few       27
```

We are working with the following logistic regression as follows:

```
prop ~ GF + GA + Pts + attendance
```

Prop here represents the proportion of team wins with respect to total matches played by the team. GF and GA represents the goals scored and conceded by the team Attendance here is a categorical variable that assigning, "High,""Medium," "Low" according to the 25%,50% and 75% of variable attendance. Likewise is applied for the 2020/2021 season, where the only difference is that, "Yes" and "None" is applied to teams playing with fans and teams playing without fans in the stadium.

Results from the regression output of the Pre COVID Data allows us to assess the predicted wins of each team in the EPL. Notably, we can see that Liverpool, ranked 1st, would have the highest predicted wins of 0.8288. The actual proportion of wins that Liverpool has is, 0.84 for a marginal difference of 0.0112. Using the predicted expected wins, we computed EW which represents the amount of predicted Wins the team would have at the end of the season (over 38 games). Computing the difference between actual wins and predicted wins at end of season shows, the difference is ranged to be -1 and 1. The regression fits relatively well. As for attendance, it can be seen that attendance does not have much of an impact on actual wins. Teams characterized with medium attendance had 60 more total wins that teams characterized in the high attendance.

Results from the regression output of the COVID Data shows a compelling result.The current league leaders, Manchester City, would have the highest predicted win percentage of 76%. In retrospect, they had won 73% of their total games to date for a marginal difference of 3%. Similar to the 2019/2020 season, Manchester City, like Liverpool, would have a 1 more predicted win (over total matches played) than their actual wins to date. To add on, we predicted the number of games that each team would win at the end of the season. Manchester City is predicted to win 29 games over 38 games.Likewise for the COVID season, the attendance does not have a impact on the team wins. Teams characterized with some attendance, has 2 more wins than teams without any attendance.

The major difference between the Pre COVID season and the COVID season shows that the current leaders would have a 7% drop in win. Over the 38 games, we predict the number of win proportions for the COVID season would be 76% as well for Manchester City. In a more notable difference, Liverpool only won 45% of their games to date in the current season. That is a 38% drop in wins for the previous season winners! For a team like Liverpool, the COVID effect with the lack of fans, has been significant. This supports the theory of ESPN as introduced in the introduction.

5.2.1 Pythagorean Analysis

```
##   Rk           Squad MP  W  D  L    GF   GA GD Pts   xG  xGA  xGD xGD.90
## 1  2 Manchester City 38 26  3  9 102.5 35.5 67  81 93.0 34.7 58.3   1.53
## 2  1       Liverpool 38 32  3  3  85.5 33.5 52  99 71.5 40.0 31.5   0.83
## 3  3  Manchester Utd 38 18 12  8  66.5 36.5 30  66 59.4 37.4 22.0   0.58
```

```
## 4  5  Leicester City 38 18  8 12  67.5 41.5 26  62 61.6 44.5 17.1   0.45
## 5  4         Chelsea 38 20  6 12  69.5 54.5 15  66 66.6 37.9 28.6   0.75
## 6  6       Tottenham 38 16 11 11  61.5 47.5 14  59 46.1 52.0 -6.0  -0.16
##   Attendance attendance      prop ExpectedWins    ExtraWins EW diff      wpct
## 1      37097     Medium 0.6842105    0.6806827  0.003527783 26    0 0.8946599
## 2      41955       High 0.8421053    0.8288912  0.013214015 31    1 0.8690161
## 3      57415       High 0.4736842    0.4822787 -0.008594526 18    0 0.7707006
## 4      25312     Medium 0.4736842    0.4595705  0.014113698 17    1 0.7275527
## 5      32023     Medium 0.5263158    0.5482127 -0.021896887 21   -1 0.6201641
## 6      43757       High 0.4210526    0.4254338 -0.004381155 16    0 0.6274874
##   expwin     GFpg      PPw
## 1     34 2.697368 1.828036
## 2     33 2.250000 1.524849
## 3     29 1.750000 1.185994
## 4     28 1.776316 1.203829
## 5     24 1.828947 1.239498
## 6     24 1.618421 1.096822
```

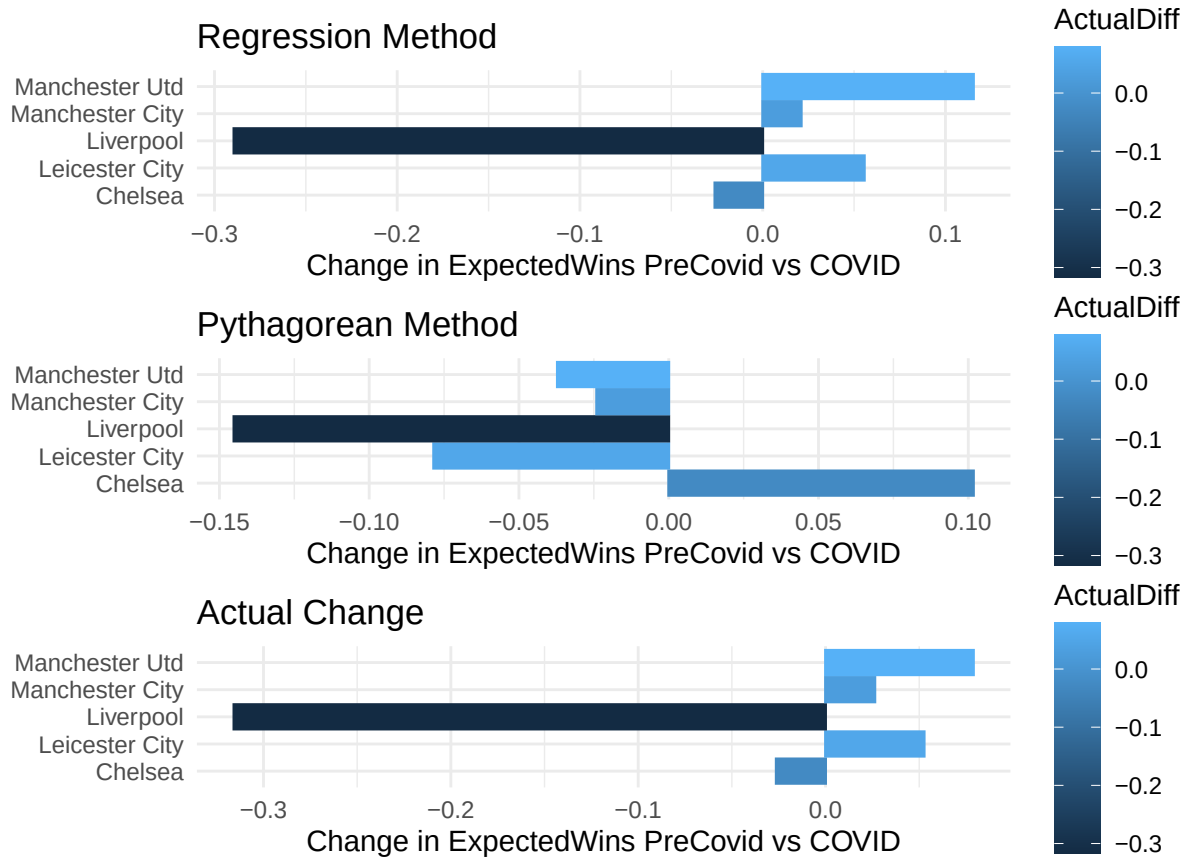
Noteworthy observations from the Pythagorean analysis:

1) Based on goals scored by the home team, Manchester City would have won the 2019/2020 season with a predicted 89% win rate vs the 68% ACTUAL win rate.

2) In the 2020/2021 season, Manchester City would have a predicted win rate of 90% vs their actual 73% win rate

3) The pythagorean analysis shows major variance in the actual vs predicted and may not accurately make for a good model.

5.3 Visual 1: A comparison between Pre COVID and During COVID Win Rates



1) From Regression Method:

The useful information obtained from this plot shows that, majority of teams have actually improved during COVID. Of the same 5 teams that were sampled from two different seasons, 4 of the 5 teams have a higher predicted win percentage during COVID. Assessing for actual difference in the proportion of wins and predicted wins, there is a consistency with the classification. Namely, the team with the highest change in predicted wins before and during COVID has the same highest actual difference in actual wins vs predicted wins

2) From Pythagorean Method:

Shows that Chelsea would have the highest change in predicted wins. But the validity of this model is questioned because the classification is not consistent. Namely, given Chelsea has the highest change in predicted wins, they are supposed to have the highest actual difference. Yet, we actually see the opposite.

3) From the actual change:

We can formally observe that the regression method is preferred over the pythagorean method due to the closeness in match in terms of change in expected wins and classification of actual difference in wins.

5.4 Analytic 2: Twitter Engagement

5.4.1 Pre Covid Engagement

```
##   Year.Month AverageEngagements AverageWins
## 1    2019-08             174896        0.25
## 2    2019-09             101050        0.33
## 3    2019-10             108980        0.33
## 4    2019-11             112641        0.33
## 5    2019-12             161911        0.57
```

Here the regression that we are interested in is:

```
Engagements.per.post ~ Year.Month + Week.Day + Page.Likes
```

Engagements per post is intuitively the engagements over the number of posts per month made by Manchester United on Twitter. Year.Month represents the month and year respectively and page likes, essentially represents the number of twitter page likes that Manchester United have on a certain day.

Noted here, we create a data frame that represents the Average Engagements (per post) and the Average Actual Wins for each level of engagement. There is no clear indication that a higher average engagement will result in an increase in the average wins for Manchester United. Evidently, when the Average Engagements was at its highest (174896), the average wins for that month was 0.25 or 25%.

5.4.2 During COVID Engagement

```
##   Year.Month AverageEngagementsCOVID AverageWinsCOVID
## 1    2020-09                  214998             0.50
## 2    2020-10                  279741             0.33
## 3    2020-11                  163609             1.00
## 4    2020-12                  205917             0.67
## 5    2021-01                  281107             0.50
## 6    2021-02                  231911             0.40
## 7    2021-03                  182677             0.67
```

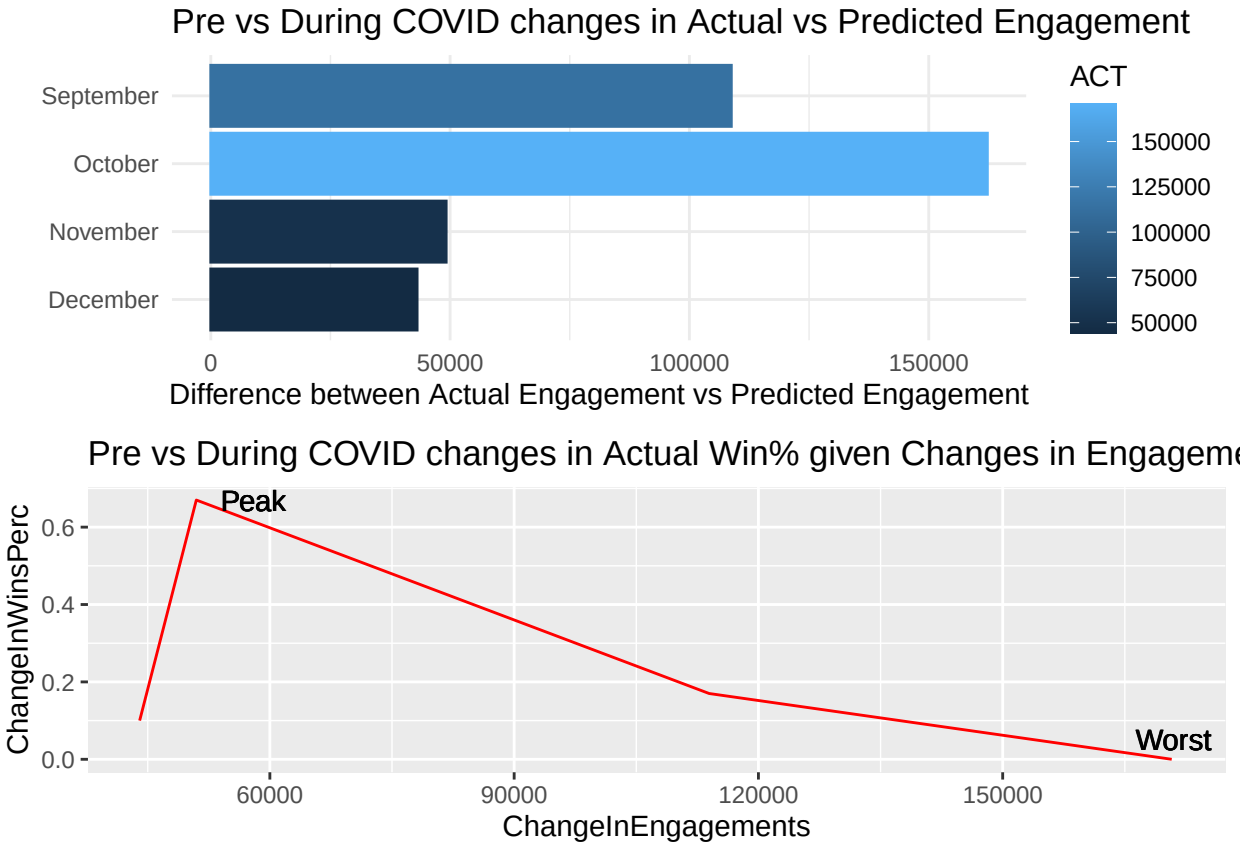
A similar tale. The number of average engagements on Twitter (per post) has no real effect on the average wins. Manchester United had the highest average wins (100%) when the average engagement was the lowest (163609). To support, when average engagements was the highest at 281107, the average wins was only 0.50 (50%)

5.4.3 Difference in Average Engagements vs Predicted Engagement

```
##      Months    ACT Year.Month TotalEPP count       AVG Year.Month.1   TEPP
## 1 September 113948  2020 - 09   435426    30 14514.200    2019 - 09 279465
## 2   October 170761  2020 - 10   489574    31 15792.710    2019 - 10 225630
## 3  November  50968  2020 - 11   301138    30 10037.933    2019 - 11 247355
## 4  December  44006  2020 - 12   305420    31  9852.258    2019 - 12 277357
##   count.1     AVGE   DiffAVG       OFF
## 1      30 9315.501 5198.6993 108749.30
## 2      31 7278.387 8514.3232 162246.68
## 3      30 8245.167 1792.7663  49175.23
## 4      31 8947.001  905.2571  43100.74
```

The main points from this comparison is to observe the predicted engagement per post vs the actual engagement per post. Here AVG, represents the predicted engagement per post during COVID and AVGE represents the predicted engagement per post PRE COVID. If we observe the extra engagement, the OFF column suggests that a very high deviance between the predicted engagement and actual engagement. We would expect to see some deviance in the plot.

5.5 Visual 2: Change in Engagement vs Change in Actual Win%



Here, when we refer to change in Win Percentage, a baseline would be: a) If the Change In Win Percentage > 0, the team has improved during COVID b) If change in Win percentage < 0, the team has NOT improved during COVID

Key Interpretations from this plot:

1) Manchester United were actually predicted to have more engagements (per post) than actual engagements (per post) in October. In the other months, this was the opposite.

2) When Manchester United Observed their lowest change in engagements(per post) they actually had a 67% change in win rate in that given month. Similarly, it can be said that when Manchester United had the biggest change in engagements (per post), they performed really poorly with a 0% change in win percentage.

3) Conclusion, while there has been increase in engagement, the increased engagement in Twitter does not necessarily correlate to more wins. This implies that the physical presence of the fans actually have a impact!

6 Conclusion

The premise of this paper was to: 1) investigate the effect of COVID on team wins in the EPL 2) investigate the effect of COVID on team engagement (MANCHESTER UNITED) in the EPL

From the results obtained, we assessed that the effect of COVID was significant for some teams in the EPL. Namely, the first observation was that the change in expected win% from the two periods was 7% between the leaders from both seasons. From classifying attendance into "High," "Medium" and "Low," it can be seen that teams classified with medium attendance did much better than teams with high attendance. Trivially, the 'COVID EFFECT' was most significant with Liverpool, who had a 38% drop in expected win percentage between the two seasons.

From the twitter engagement, it can be seen clearly that there is no clear association between the change in engagement with the team performance in terms of wins. Our evidence shows that, when Manchester United had the highest change in positive engagement in twitter, their team performance had not improved (0% change). But when the team had the lowest change in positive engagement, the team had had 67% change in wins between seasons.

Future improvements to the model:

1) Do a full model building assessing for the factors that actually affect the team wins/team engagements
2) Assess the model diagnostics for influential points
3) Having full data from the 2020/2021 season to assess a fair comparison between seasons.

References

"2019-2020 Premier League Stats." n.d. *FBref.com*. https://fbref.com/en/comps/9/3232/2019-2020-Premier-League-Stats.

Delignette-Muller, Marie Laure, and Christophe Dutang. 2015. "fitdistrplus: An R Package for Fitting Distributions." *Journal of Statistical Software* 64 (4): 1–34. https://doi.org/10.18637/jss.v064.i04.

Hamilton, Tom. 2021. "Premier League's Home Edge Has Gone in Pandemic Era: The Impact of Fan-Less Games in England and Europe." *ESPN*. ESPN Internet Ventures. https://www.espn.com/soccer/english-premier-league/story/4312130/premier-leagues-home-edge-has-gone-in-pandemic-era-the-impact-of-fan-less-games-in-england-and-europe.

Harrell Jr, Frank E. 2021. *Rms: Regression Modeling Strategies*. https://CRAN.R-project.org/package=rms.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richards, Sam. 2020. "Statistically Ranking the Biggest Football Leagues in the World Today." *The Exeter Daily*. https://www.theexeterdaily.co.uk/news/sport/statistically-ranking-biggest-football-leagues-world-today.

Rudis, Bob. 2020. *Hrbrthemes: Additional Themes, Theme Components and Utilities for 'Ggplot2'*. https://CRAN.R-project.org/package=hrbrthemes.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Jennifer Bryan. 2022. *Readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl.