Professional Soccer Stadium Capacity Analysis

06/04/2022

Team Analytica
Mat Stanham, Dianne Pham, Jackie Tencza, Shane Conner, Rosa Ortiz

Table of Contents

Executive Summary

Team Analytica is a consulting company composed of Mat Stanham, Dianne Pham, Jackie Tenza, Shane Conner, and Rosa Ortiz. As a company, we provide expert analysis and recommendations based on the collective expertise represented by each associate.

In working with professional soccer team management, Team Analytica aims to explore the relationship between wins and fan attendance. The goal is to build a report and dashboard that summarizes the effect of fan attendance on wins for soccer teams.

For this project, Team Analytica analyzed datasets from soccer matches across five leagues for the past several years using Python. Next, the data was transformed and hypothesis testing was performed to ensure the independent variables were appropriate, and extract insights from the dataset. Next came the modeling process, where multiple linear regression models were developed to determine win percentage using R. Finally, an interactive dashboard was created using R Shiny to allow the professional soccer teams to interact with the final dataset and linear regression model.

As a result of this project, professional soccer teams can use the interactive dashboard and results to determine where to invest their money and energy in order to maximize wins.

Overview & Introduction

Professional Soccer Teams make significant capital investment decisions to build stadiums or renovate an existing structure as a means to increasing capacity and, ultimately, revenue from fan attendance. While increasing stadium capacity can increase potential revenue, such undertakings are slated as being expensive and a risk to the respective team's financial future. Arriving at a decision based on dependable data can potentially be the difference between a defunct sports team or a successful one. According to Albert et al., referencing a study by Boyko et al. (2007), it is stated “Crowd density had significant effects on goal difference and home team penalties” (407). Therefore, it is important to make judicious investments and understand the potential impacts to on-field performance from changes to stadium capacity.

In professional sports, measuring the impact of fan presence on in-game performance has historically been challenging due to the inability to control for a stadium's fan presence variable. The COVID-19 pandemic forced matches without in-person fan attendance and produced a dataset that controls for the presence of fans in a stadium. With this data, we can now explore and better understand the relationships between key sports variables, such as stadium capacity/utilization and home/away performance statistics.

Team Analytica Consulting will provide the league's and team's management with a tool to effectively direct off-field decisions related to stadium capacity using an objective analysis based on a data-driven methodological approach to minimize and/or eliminate impacts to fan attendance and its effect on in-game performance.

Proposed Objectives

Team Analytica Consulting aims to provide a concise report that summarizes the effect of fan attendance on wins for soccer teams. In conjunction, findings are supported by a web-based dashboard tool providing further exploration into the different variables that comprise team performance with relation to stadium capacity/utilization. The consulting team anticipates the deliverable will be used to assist professional soccer teams in making decisions regarding stadium improvements and renovations.

Data Source Review

The data comes from the FBREF website and is owned by SportsReference (FBREF). SportsReference compiles a variety of sports-related data from third-parties. The data from this dataset comes from the third party StatsPerform. Team Analytica Consulting collected every match played from the start of the 2016 season through the 2021 season. This data will be collected for the following leagues: Bundesliga, English Premier League (EPL), La Liga, Ligue 1, and Serie A. For each match, we have the season the match was played in, the date the match was played, the name of the home team and the away team, the number of fans in attendance, the venue, the final score, the referee, and a URL for that game's match report. Within each match report, are useful metrics for our analysis. The match report metrics that are the focus for this analysis are: pass completion, pass attempts, shots on target, shots, fouls, tackles, interceptions, yellow cards and red cards for both the home and away teams. There are two additional variables that are pulled from the FBREF website: stadium capacity and whether or not the match contained a Video Assistant Referee (VAR). Stadium capacity looks at the total number of fans that could be in attendance (the capacity of the venue the stadium is in).

VAR is a possible confounding variable that was introduced in the German Bundesliga for the 2017/18 season, in the English Premier League for the 2018/19 season, in the Spanish La Liga for the 2018/19 season, in the French Ligue 1 for the 2017/18 season, and in the Italian Serie A for the 2017/18 season. VAR was written into the laws of the game by the International Football Association Board-IFAB in 2018. According to the IFAB website, "The referee may receive assistance from the VAR only in relation to four categories of match-changing decisions/incidents" (Video). In all these situations, the VAR is only used after the referee has made a (first/original) decision (including allowing play to continue), or if a serious incident is missed/not seen by the match officials. The referee`s original decision will not be changed unless there was a 'clear and obvious error' (this includes any decision made by the referee based on information from another match official e.g. offside). The categories of decision/incident which may be reviewed in the event of a potential ‘clear and obvious error’ or ‘serious missed incident’ are:

a. Goal/no goal
b. Penalty kick/no penalty kick
c. Direct red cards (not second yellow card/caution)
d. Mistaken identity (red or yellow card)

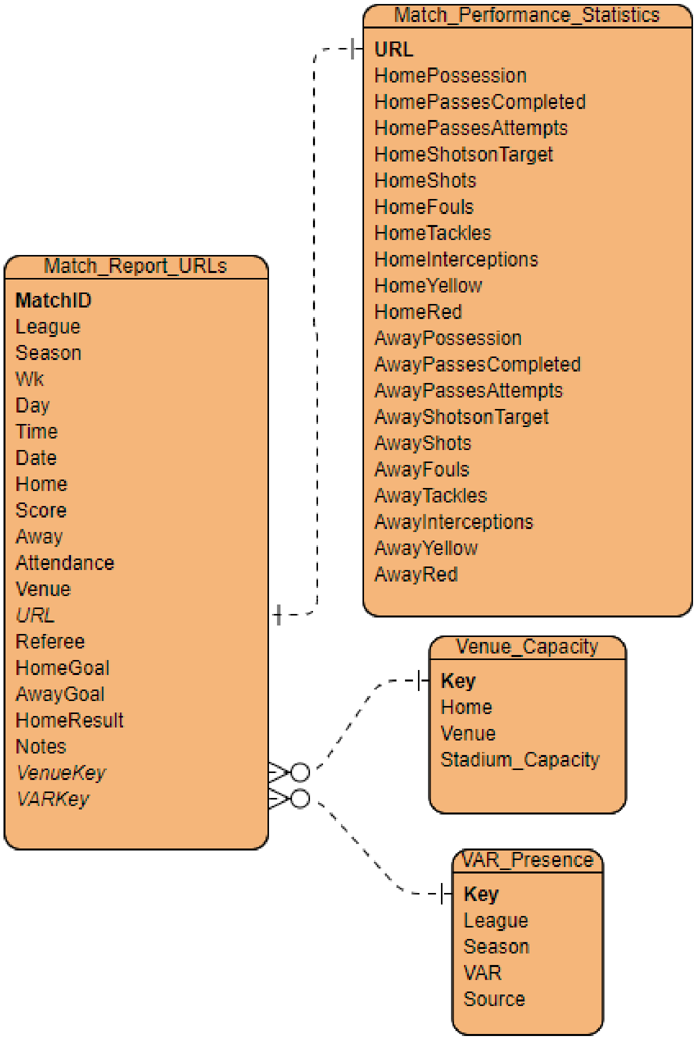
If the referee penalizes an offense and gives the wrong player from the offending (penalized) team a yellow or red card, the identity of the offender can be reviewed; the actual offense itself cannot be reviewed unless it relates to a goal, penalty incident or direct red card” (Video).

It is important to take into consideration the impact that VAR has on team performance before analyzing the impact of fans on team performance. It is possible that the changes to in-game win percentage and other metrics may be due to VAR and not to the absence of fans.

Using the BeautifulSoup package in Python, we are able to pull in the data from FBRef using web scraping. This was inspired by Christopher B. Martin’s Premier League history web scraper (Martin). The data is stored in an Azure Data Lake Gen1 environment. To prevent the site's bot from blocking our code, the source data was pulled 100-150 records at a time using three second intervals. Approximately 9130 match records were stored throughout the several seasons.

Data Processing & Transformation

The data is stored in four different tables found in teh Match Report URLs.xlsx file. Match_Report_URLs, Match_Performance_Statistics, Venue_Capacity, and VAR_Presence. Match_Report_URLs is the main table with a primary key titled MatchID. In Match_Report_URLs there are three foreign keys: URL, VenueKey, and VARKey. There is a 1 : 1 relationship between Match_Report_URLs and Match_Performance_Statistics joined by URL. There is a 1 : Many relationship between Match_Report_URLs and Venue_Capacity joined by the VenueKey. There is also a 1 : Many relationship between Match_Report_URLs and VAR_Presence joined by the VARKey.



There are seven columns which are transformed in the target tables. In the Match_Report_URLs table there is HomeGoal, AwayGoal, HomeResult, VenueKey, and VARKey. HomeGoal and AwayGoal are both based on the score of the match and are numerical. HomeResult is based on the score as well but is categorical (W or L or T). VenueKey is a concatenation between Home and Venue. VarKey is a concatenation between League and Season. In the Venue_Capacity table the Key field is a concatenation between Home and Venue. In the VAR_Presence table the Key field is a concatenation between League and Season.

Exploratory Data Analysis

The original dataset consists of 9,130 observations or matches and 41 variables including league, date, score, stadium capacity, and presence of Video Assistant Referee (VAR). In initial data cleansing, we replace 2,233 null attendance values with 0, remove 2 matches from Italian Serie A due to rule violations found in the Notes column, and remove 101 matches from the French Ligue 1 2019-2020 season because they were canceled and no match took place - which also removes the missing values in Time and Referee variables. This data cleanup left us with 9,027 games for analysis, a decrease of only 1% of the data, which fell under our reasonableness threshold. The remaining totals for the dataset can be found in Figure 1 in the Appendix.

In order to understand the impact of fans on match outcomes, we calculated stadium utilization and added it as an additional field to the dataset. Stadium Utilization is calculated by dividing match attendance by total stadium capacity. The results are the buckets in the Appendix (Figure 3). All leagues have a healthy quantity of matches where fans are not present. The same can be said for near max capacity. Interestingly, La Liga, Ligue 1, and Serie A have very few sold out games relative to the Bundesliga and EPL. This new variable revealed that we might not have had a large enough sample of games to measure stadium utilization at every rate for every league. However, we had enough games across all leagues combined to make generalizations on stadium utilization across the combined 5 leagues.

One of our biggest confounding variables, which we needed to control for, is the implementation of the VAR. Thankfully, there was a healthy quantity of games across all leagues where VAR is both present and not present. In total, we have 2,965 matches without VAR and 6,062 matches with VAR. The Home team win rate without VAR looks to be approximately 48%, while with VAR it looks to be 43%.

To identify potential outliers, we reviewed all of the variables in the dataset - boxplots can be found in Figure 6 in the Appendix. We removed 1,358 records from our dataset due to outliers in the explanatory variables. This represents less than 15% of matches in our original dataset. Although this represents a larger impact on matches we could use in our model, it ensures that outliers did not skew our results. As such, we removed matches where the value of those variables fell below the first quartile or above the third quartile of values. The largest bucket of outliers fell under the HomeGoal variable. We removed between 123 matches and 153 matches (a total of 713 matches) from each of the leagues due to these outliers.

Finally, we did not want to include in our final model variables that were highly correlated. We examine intercorrelations among explanatory variables with correlation matrix/heat map in Figure 7 in the Appendix. We removed all of the variables that were highly correlated with other variables of interest from our model and were left with the following: Stadium, Capacity, Stadium Utilization, League, GoalsScored, Home Pass Accuracy, Home Shot Accuracy, Home Tackles, Home Interceptions, Home Yellow, Home Red, Away Pass Accuracy, Away Shot Accuracy, Away Tackles, Away Interceptions, Away Yellow, and Away Red.

Data Analysis & Modeling

For our data analysis, our goal was to identify whether or not fan attendance had a statistically significant difference on the win rate of the home team. By focusing specifically on the home teams, we can control for both the competitive quality of various home teams, as well as ensure we have enough observations to make generalizations.

In this case, fan attendance is denoted using a binary variable, based on the attendance field: if the attendance is equal to zero, then the match is classified as “no fans”, otherwise it is classified as “fans.”

In order to test for the impact of fans, the team deployed a technique known as two sample proportion z-test. This technique will allow us to compare the proportions between two sample populations. This testing process allows us to see if there is a difference in win rates between when fans are present and when fans are not present.

That being said, before we can test for the effects of fans, there is an assumption that needs to be tested: for the two sample proportion z-test to be effective, our data needs to hold all other variables constant, otherwise we will not be able to claim whether the effects are from fan attendance or some other ‘confounding variable’. The variable we are interested in accounting for is the presence of the Video Assistant Referee. Matches where the video assistant referee is present can be found in the binary variable VAR: games with a zero in this field denote a match without VAR, and “1” in this field denotes a match with VAR. In addition to this variable, we would like to see if the impact of fan attendance is consistent between different leagues.

Therefore, first we performed the two sample proportion z-test both on the entire match dataset and then on matches split by leagues. If we find that VAR does have a statistically significant effect on the home team win rate, then rather than testing for the significance of fan attendance on the entire set of matches, the testing on fan attendance will be done on a subset of matches where VAR is present. This will allow us to ‘control’ for the potential influence of VAR.

By the end of our testing our group will be able to answer the following questions:

- Does VAR generally have a statistically significant effect on the home team win rate?
- Does fan attendance have a statistically significant effect on the home team win rate, having accounted for the potential confounding effects of VAR?

- Is the impact of VAR and fan attendance on home team win rates consistent between leagues, or are they different between leagues?

In an article on testing testing a difference in population proportions, Pattnaik writes that the two sample proportion z-test is a statistical inference test that requires both the set-up of a null and alternative hypothesis, a set level of significance, and that both of the following assumptions are met:

1. That all the matches are independent.
2. That each sample group has at least 10 observations in each proportion respectively.

All of the Null and Alternative hypothesis will have a similar structure:

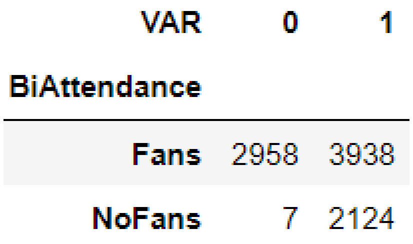
H_0: Pi -Pj = 0

H_A: Pi -Pj ≠ 0

Where P is the win rate for a given proportion.

As for a significance threshold, our team has selected a constant level of significance to be 0.05. The result of testing will be a P-value. This P-value identifies the probability that the difference we are observing between the two win rates is the same given what the win rates are. Below are details on all 12 tests conducted:

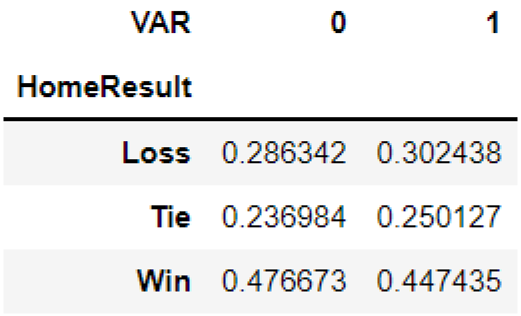
The first is the General Test across all matches:



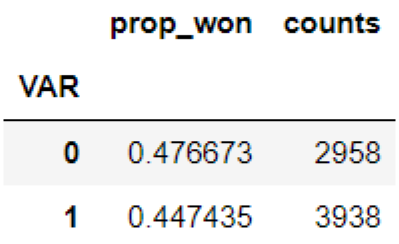
VAR	0	1
BiAttendance		
Fans	2958	3938
NoFans	7	2124

Here are the null and alternative hypothesis for this test
H_0: VAR has no impact on home team win percentage
$P_{VAR} - P_{NoVAR} = 0$

H_A: VAR has an impact on home team win percentage
$P_{VAR} - P_{NoVAR} \neq 0$



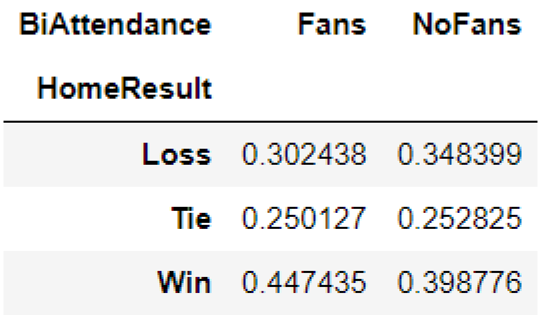
VAR	0	1
HomeResult		
Loss	0.286342	0.302438
Tie	0.236984	0.250127
Win	0.476673	0.447435



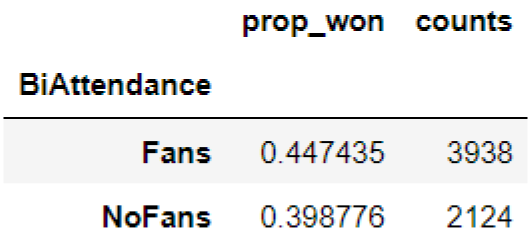
	prop_won	counts
VAR		
0	0.476673	2958
1	0.447435	3938

Based on the count of matches, we have the minimum 10 games required for testing, and all assumptions for testing are met.The resulting P-value for this test is 0.0159. With this P-value, at an 0.05 level of significance, we reject the null hypothesis that VAR has no impact on home team win percentage.

This next test was conducted on the subset of all VAR games:
All other variables held constant, controlling for the presence of VAR:
H_0: The presence of fans has no impact on home team win percentage
$P_{FANS} - P_{NoFANS} = 0$
H_A: The presence of fans has impact on home team win percentage
$P_{FANS} - P_{NoFANS} \neq 0$



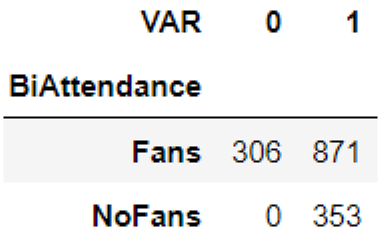
BiAttendance	Fans	NoFans
HomeResult		
Loss	0.302438	0.348399
Tie	0.250127	0.252825
Win	0.447435	0.398776



	prop_won	counts
BiAttendance		
Fans	0.447435	3938
NoFans	0.398776	2124

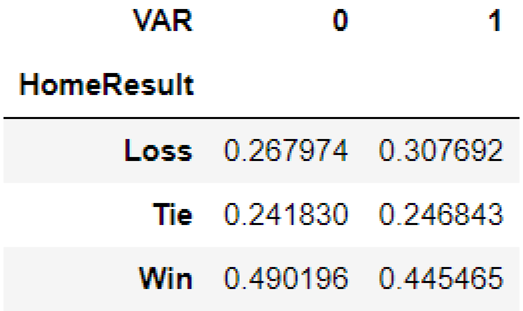
Based on the count of matches, we have the minimum 10 games required for testing, and all assumptions for testing are met.The resulting P-value for this test is 0.00026. With this P-value, at an 0.05 level of significance, we reject the null hypothesis that the presence of fans has no impact on home team win percentage.

Here are the results for the German Bundesliga.

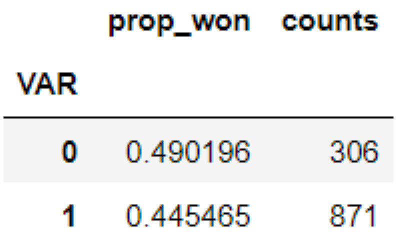


VAR	0	1
BiAttendance		
Fans	306	871
NoFans	0	353

Here are the null and alternative hypothesis for this test
H_0: VAR has no impact on home team win percentage in the Bundesliga
H_A: VAR has an impact on home team win percentage in the Bundesliga
$P_{BundesligaVAR} - P_{BundesligaNoVAR} = 0$
$P_{BundesligaVAR} - P_{BundesligaNoVAR} \neq 0$



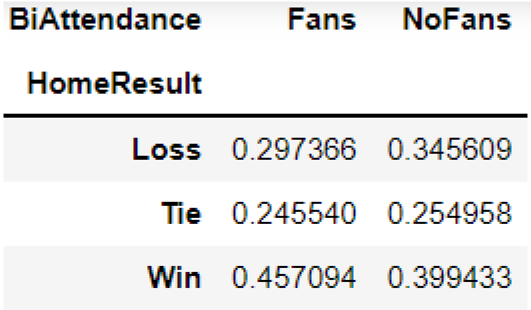
VAR	0	1
HomeResult		
Loss	0.267974	0.307692
Tie	0.241830	0.246843
Win	0.490196	0.445465



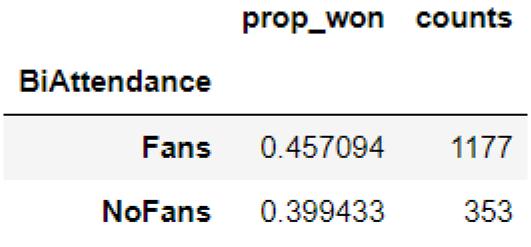
	prop_won	counts
VAR		
0	0.490196	306
1	0.445465	871

Based on the count of matches, we have the minimum 10 games required for testing, and all assumptions for testing are met. The resulting P-value for this test is 0.1766. With this P-value, at an 0.05 level of significance, we fail to reject the null hypothesis that VAR has no impact on home team win percentage in the Bundesliga.

This is the test we will conduct on the subset of ALL Bundesliga games:
All other variables held constant:
H_0: The presence of fans has no impact on home team win percentage in the Bundesliga
$P_{BundesligaFANS} - P_{BundesligaNoFANS} = 0$
H_A: The presence of fans has impact on home team win percentage in the Bundesliga
$P_{BundesligaFANS} - P_{BundesligaNoFANS} \neq 0$



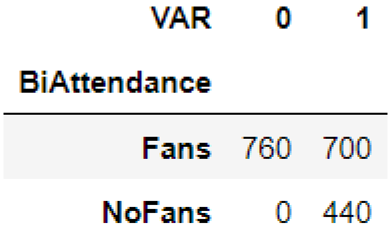
BiAttendance	Fans	NoFans
HomeResult		
Loss	0.297366	0.345609
Tie	0.245540	0.254958
Win	0.457094	0.399433



	prop_won	counts
BiAttendance		
Fans	0.457094	1177
NoFans	0.399433	353

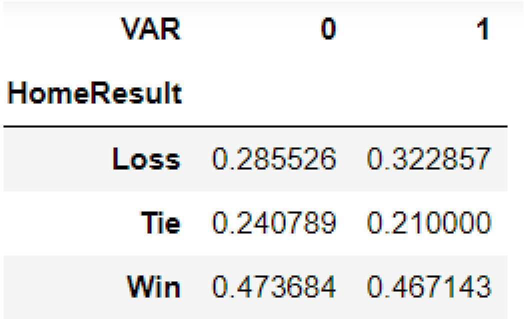
Based on the count of matches, we have the minimum 10 games required for testing, and all assumptions for testing are met. The resulting P-value for this test is 0.0558. With this P-value, at an 0.05 level of significance, we fail to reject the null hypothesis that The presence of fans has no impact on home team win percentage in the Bundesliga.

Here are the results for the English Premier League (EPL):

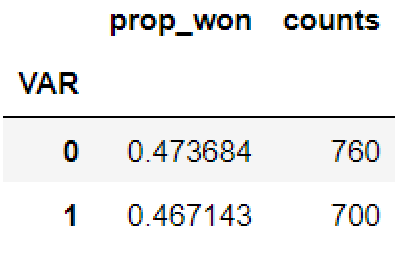


VAR	0	1
BiAttendance		
Fans	760	700
NoFans	0	440

Here are the null and alternative hypothesis for this test
H_0: VAR has no impact on home team win percentage in the EPL
$P_{EPLVAR} - P_{EPLNoVAR} = 0$
H_A: VAR has an impact on home team win percentage in the EPL
$P_{EPLVAR} - P_{EPLNoVAR} \neq 0$



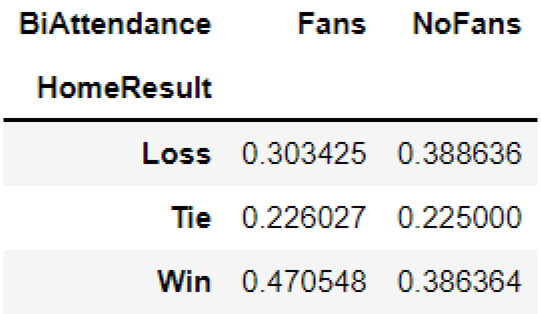
VAR	0	1
HomeResult		
Loss	0.285526	0.322857
Tie	0.240789	0.210000
Win	0.473684	0.467143



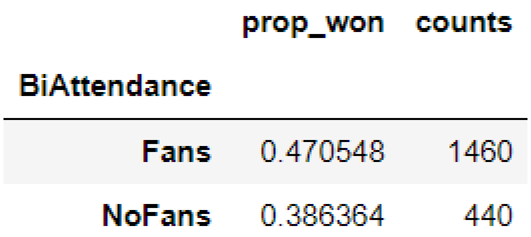
	prop_won	counts
VAR		
0	0.473684	760
1	0.467143	700

Based on the count of matches, we have the minimum 10 games required for testing, and all assumptions for testing are met.The resulting P-value for this test is 0.802. With this P-value, at an 0.05 level of significance, we fail to reject the null hypothesis that VAR has no impact on home team win percentage in the EPL.

This is the test we will conduct on the subset of ALL EPL games:
All other variables held constant:
H_0: The presence of fans has no impact on home team win percentage in the EPL
$P_{EPLFANS}$ -$P_{EPLNoFANS}$ = 0
H_A: The presence of fans has impact on home team win percentage in the EPL
$P_{EPLFANS}$ -$P_{EPLNoFANS}$ ≠ 0



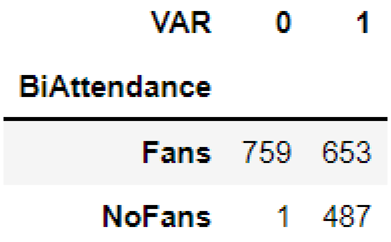
BiAttendance	Fans	NoFans
HomeResult		
Loss	0.303425	0.388636
Tie	0.226027	0.225000
Win	0.470548	0.386364



	prop_won	counts
BiAttendance		
Fans	0.470548	1460
NoFans	0.386364	440

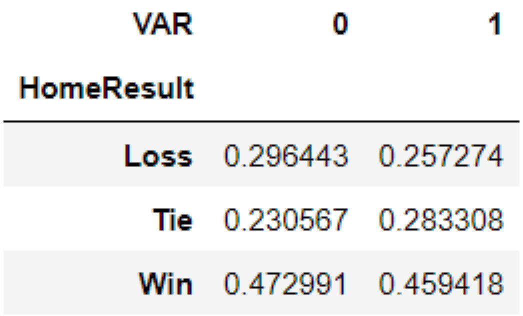
Based on the count of matches, we have the minimum 10 games required for testing, and all assumptions for testing are met.The resulting P-value for this test is 0.0019. With this P-value, at an 0.05 level of significance, we reject the null hypothesis that The presence of fans has no impact on home team win percentage in the EPL.

Here are the results for the Spanish La Liga:

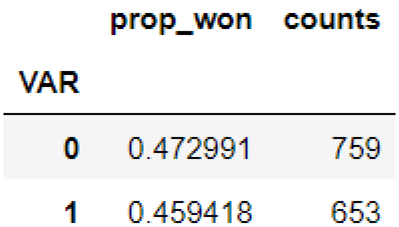


VAR	0	1
BiAttendance		
Fans	759	653
NoFans	1	487

Here are the null and alternative hypothesis for this test
H_0: VAR has no impact on home team win percentage in La Liga
H_A: VAR has an impact on home team win percentage in La Liga
$P_{LaLigaVAR}$ -$P_{LaLigaNoVAR}$ = 0
$P_{LaLigaVAR}$ -$P_{LaLigaNoVAR}$ ≠ 0



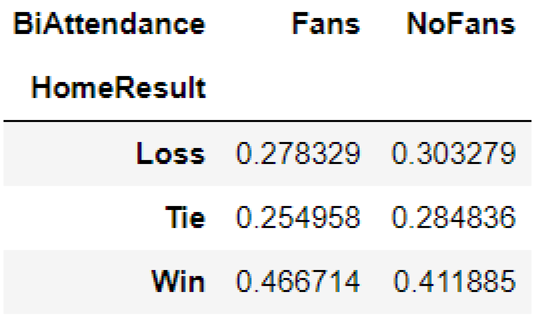
VAR	0	1
HomeResult		
Loss	0.296443	0.257274
Tie	0.230567	0.283308
Win	0.472991	0.459418



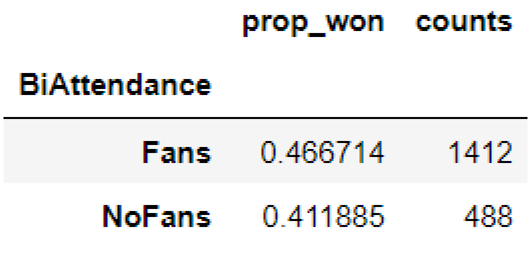
	prop_won	counts
VAR		
0	0.472991	759
1	0.459418	653

Based on the count of matches, we have the minimum 10 games required for testing, and all assumptions for testing are met.The resulting P-value for this test is 0.61. With this P-value, at an 0.05 level of significance, we fail to reject the null hypothesis that VAR has no impact on home team win percentage in La Liga.

This is the test we will conduct on the subset of ALL La Liga games:
All other variables held constant:
H_0: The presence of fans has no impact on home team win percentage in La Liga
$P_{LaLigaFANS} - P_{LaLigaNoFANS} = 0$
H_A: The presence of fans has impact on home team win percentage in La Liga
$P_{LaLigaFANS} - P_{LaLigaNoFANS} \neq 0$



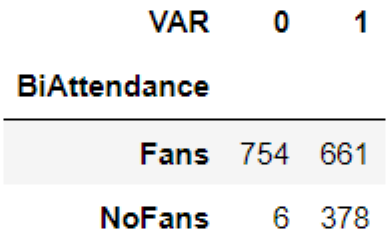
BiAttendance	Fans	NoFans
HomeResult		
Loss	0.278329	0.303279
Tie	0.254958	0.284836
Win	0.466714	0.411885



	prop_won	counts
BiAttendance		
Fans	0.466714	1412
NoFans	0.411885	488

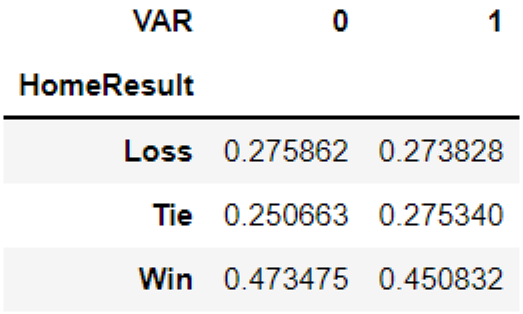
Based on the count of matches, we have the minimum 10 games required for testing, and all assumptions for testing are met.The resulting P-value for this test is 0.0359. With this P-value, at an 0.05 level of significance, we reject the null hypothesis that The presence of fans has no impact on home team win percentage in La Liga.

Here are the results for the French Ligue 1:

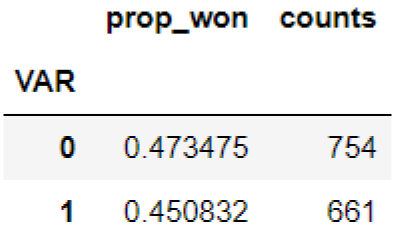


VAR	0	1
BiAttendance		
Fans	754	661
NoFans	6	378

Here is the null and alternative hypothesis for this test
H_0: VAR has no impact on home team win percentage in Ligue 1
H_A: VAR has an impact on home team win percentage in Ligue 1
$P_{Ligue1VAR} - P_{Ligue1NoVAR} = 0$
$P_{Ligue1VAR} - P_{Ligue1NoVAR} \neq 0$



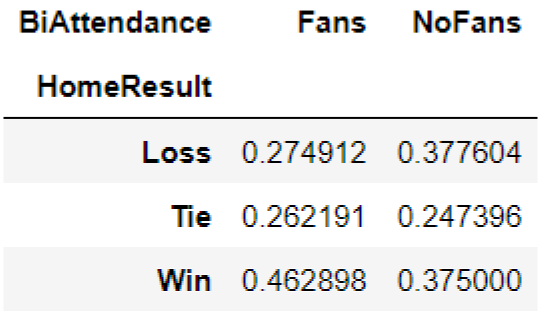
VAR	0	1
HomeResult		
Loss	0.275862	0.273828
Tie	0.250663	0.275340
Win	0.473475	0.450832



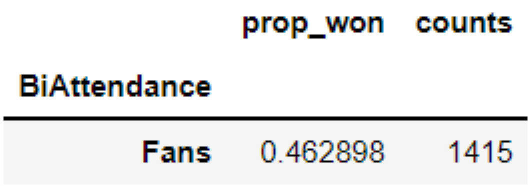
	prop_won	counts
VAR		
0	0.473475	754
1	0.450832	661

Based on the count of matches, we have the minimum 10 games required for testing, and all assumptions for testing are met.The resulting P-value for this test is 0.394. With this P-value, at an 0.05 level of significance, we fail to reject the null hypothesis that VAR has no impact on home team win percentage in Ligue 1.

This is the test we will conduct on the subset of ALL Ligue 1 games:
All other variables held constant:
H_0: The presence of fans has no impact on home team win percentage in Ligue 1
$P_{Ligue1FANS}$ -$P_{Ligue1NoFANS}$ = 0
H_A: The presence of fans has impact on home team win percentage in Ligue 1
$P_{Ligue1FANS}$ -$P_{Ligue1NoFANS}$ ≠ 0



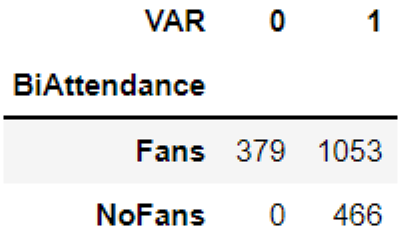
BiAttendance	Fans	NoFans
HomeResult		
Loss	0.274912	0.377604
Tie	0.262191	0.247396
Win	0.462898	0.375000



	prop_won	counts
BiAttendance		
Fans	0.462898	1415
NoFans	0.375000	384

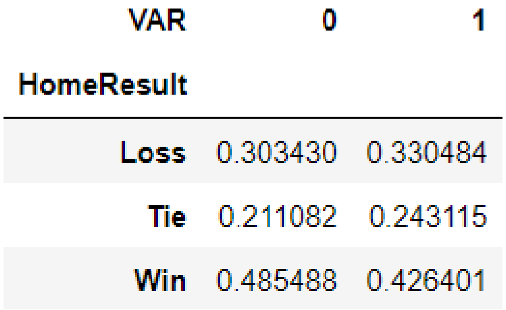
Based on the count of matches, we have the minimum 10 games required for testing, and all assumptions for testing are met.The resulting P-value for this test is 0.002. With this P-value, at an 0.05 level of significance, we reject the null hypothesis that The presence of fans has no impact on home team win percentage in Ligue 1.

Last, here are the results for the Italian Serie A:

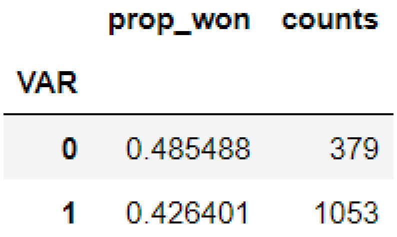


VAR	0	1
BiAttendance		
Fans	379	1053
NoFans	0	466

Here is the null and alternative hypothesis for this test
H_0: VAR has no impact on home team win percentage in the Serie A
$P_{SerieAVAR}$ -$P_{SerieANoVAR}$ = 0
H_A: VAR has an impact on home team win percentage in the Serie A
$P_{SerieAVAR}$ -$P_{SerieANoVAR}$ ≠ 0



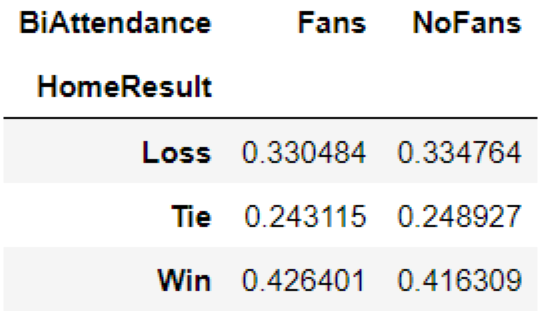
VAR	0	1
HomeResult		
Loss	0.303430	0.330484
Tie	0.211082	0.243115
Win	0.485488	0.426401



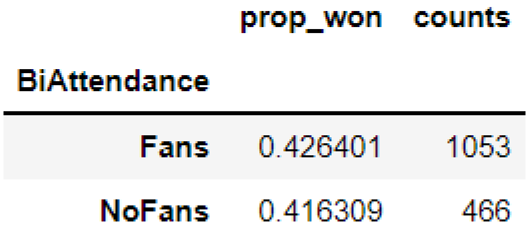
	prop_won	counts
VAR		
0	0.485488	379
1	0.426401	1053

Based on the count of matches, we have the minimum 10 games required for testing, and all assumptions for testing are met.The resulting P-value for this test is 0.047. With this P-value, at an 0.05 level of significance, we reject the null hypothesis that VAR has no impact on home team win percentage in the Serie A.

This is the test we will conduct on the subset of all Serie A games with VAR:
All other variables held constant:
H_0: The presence of fans has no impact on home team win percentage in the Serie A
$P_{SerieAFANS}$ -$P_{SerieANoFANS}$ = 0
H_A: The presence of fans has impact on home team win percentage in the Serie A
$P_{SerieAFANS}$ -$P_{SerieANoFANS}$ ≠ 0



BiAttendance	Fans	NoFans
HomeResult		
Loss	0.330484	0.334764
Tie	0.243115	0.248927
Win	0.426401	0.416309



	prop_won	counts
BiAttendance		
Fans	0.426401	1053
NoFans	0.416309	466

Based on the count of matches, we have the minimum 10 games required for testing and all assumptions for testing are met.The resulting P-value for this test is 0.714. With this P-value, at an 0.05 level of significance, we fail to reject the null hypothesis that the presence of fans has no impact on home team win percentage in the Serie A.

A summary table of all the testing results can be found in the summary of findings section. After completing our hypothesis testing and ensuring our final dataset includes the independent variables of interest, we can begin the modeling process.

Given that our sport management customer would ultimately care about the sports organization's performance, our response variable of the model will be win percentage.

As a part of the data transformation process for modeling, win rates were calculated for each individual stadium in our dataset that had at least 30 games played in it. Our initial dataset of 9,027 matches corresponded to 143 different home stadiums. Of those, 571 matches corresponded with 32 stadiums that did not have at least 30 games played in them. This resulted in a dataset of 8,456 matches across 111 unique stadiums. Win rates were calculated on a per season, per stadium basis: at most a single stadium would have 5 unique seasonal-stadium win rates. The result was 445 observations used for modeling.

Win percentage is modeled using the following independent variables:

1. Stadium Capacity
2. Attendance
3. Stadium Utilization= Attendance/Stadium Capacity
4. Home Team Fouls
5. Home Team Tackles
6. Home Team Interceptions
7. Home Team Yellow
8. Home Team Red
9. Away Team Fouls
10. Away Team Tackles
11. Away Team Interceptions
12. Away Team Yellow
13. Away Team Red
14. Goals Scored= Home Team Goal + Away Team Goal
15. Home Pass Accuracy = Home Team Passes Completed/Home Team Passes Attempts
16. Home Shot Accuracy= Home Team Shots / Home Team Shots on Target
17. Away Pass Accuracy= Away Team Passes Completed/Away Team Passes Attempts
18. Away Shot Accuracy = Away Team Shots on Target / Away Team Shots
19. League (Indicator for Bundesliga, EPL, LaLiga, Ligue1, SerieA)

For each of the above independent variables listed, the seasonal-stadium average metric was calculated and joined up with their corresponding observation. The result was a dataset with the above 19 independent variables, and our dependent variable the seasonal-stadium win rate, across all 445 observations.

As we begin modeling, we want to keep our customer's use case in mind and make sure the final model is interpretable for our sports management audience. As a result, and to avoid loss of interpretability, variables are not transformed unless absolutely necessary. For instance, having management interpret coefficients as the impact of increasing the log(stadium capacity) by 1 unit is a step better left as unrequired. In addition, the model will retain its simplicity to facilitate a user's ability to understand and interpret its insight, regardless of statistical knowledge. To do this, we explore a multivariate linear regression model rather than starting with a more complex machine learning model. With a linear regression, we can see the fitted formula and isolate the impacts of each variable. With a machine learning model, it can feel like a "black box" and hard to interpret.

To evaluate the performance of each model, we fit multiple linear regressions using the statistical metric RMSE (root mean square error). The dataset is split into a train (80%) and test (20%) datasets. Using the train dataset, the model will be fit and evaluated on the test dataset using its RMSE. This metric will measure how much error exists when a model is used to predict on a new dataset. A low RMSE indicates the model's greater performance. The following table represents a full comparison of the fitted linear models.

Model	Independent Variables	RMSE on Test Dataset
1	Attendance + Stadium Utilization + Stadium Capacity + Home Fouls + Home Tackles + Home Interceptions + Home Yellow + Home Red + Away Fouls + Away Tackles + Away Interceptions + Away Yellow + Away Red +Goals Scored + Home Pass Accuracy + Home Shot Accuracy + Away Pass Accuracy + Away Shot Accuracy + dummy Bundesliga + dummy EPL + dummy LaLiga + dummy Ligue1	0.178
2	Stadium Utilization + Stadium Capacity + Away Fouls + Away Tackles + Away Interceptions + Away Yellow +Goals Scored + Home Pass Accuracy + Home Shot Accuracy + Away Pass Accuracy + Away Shot Accuracy + dummy Bundesliga + dummy EPL + dummy LaLiga + dummy Ligue1	0.199
3	Stadium Utilization + Stadium Capacity + Away Fouls + Away Tackles + Away Yellow +Goals Scored + Home Pass Accuracy + Home Shot Accuracy + Away Pass Accuracy + Away Shot Accuracy + dummy Bundesliga + dummy EPL + dummy LaLiga + dummy Ligue1	0.190
4	Stadium Utilization + Stadium Capacity + Away Fouls + Away Tackles + Away Yellow + Goals Scored + Home Pass Accuracy + Home Shot Accuracy + Away Pass Accuracy + Away Shot Accuracy	0.182
5	**Stadium Utilization + Stadium Capacity + Away Fouls + Away Tackles + Away Yellow + Home Pass Accuracy + Home Shot Accuracy + Away Pass Accuracy + Away Shot Accuracy**	**0.176**

6	Stadium Utilization + Away Fouls + Away Tackles + Away Yellow + Home Pass Accuracy + Home Shot Accuracy + Away Pass Accuracy + Away Shot Accuracy	0.182
7	Stadium Utilization + Away Yellow + Home Pass Accuracy + Home Shot Accuracy + Away Pass Accuracy + Away Shot Accuracy	0.177

Our final model is the following:

Win Percentage = 0.093 + 0.099*Stadium Utilization + 0.0000012*Stadium Capacity - 0.0145*Away Team Fouls - 0.007*Away Team Tackles + 0.054*Away Team Yellow + 1.486*Home Team Pass Accuracy + 1.356*Home Team Shot Accuracy - 1.110*Away Team Pass Accuracy - 0.889*Away Team Shot Accuracy

With an R-squared of 62.5% and a low RMSE on the test dataset of 0.176, we are happy with the performance of our final model given it is interpretable for the customer.

Final Visualization

As part of the final deliverable, Analytical Consulting provides an interactive dashboard that allows management to interact with our final dataset as well as our final linear regression model. Keeping in mind that not all users have a statistical background, it is important that our UI (user interface) design remains easy to understand. With respect to visualization design, using Deloitte's tips for effective dashboard design as referenced by Vanderlaken, the following main objectives were taken into account (Vanderlaken):

1. Well-documented
2. Clearly defined sections
3. Well-labeled graphs
4. Ease of use/interaction
5. Interpretability: customer can immediatley understand the insights
6. Strategic use of color coding

The platform of choice is R Shiny. The following link is a live beta version of our tool:

https://team-analytica.shinyapps.io/Final_Report/

In the top right-hand corner the user will find important information about the data update and the raw data source. The first section of the dashboard, "Exploring the Game Data"

(Figure 8), allows the customer to dynamically explore the relationship between multiple independent variables and the result of the soccer game. The user can change the x-axis variable and then the boxplot visual will update automatically. An additional slicing that the user can add to their analysis is comparing the distributions with and without fans and/or video assistant referees (VAR).

Depending on the league management a user represents, they might only be interested in certain soccer leagues which can easily be filtered using the dropdown menu. Recent trends can be viewed by selecting only the most recent season(s).

The “Future Capacity Planning & Predictions” section of the dashboard (Figure 9) allows the customer to dynamically change the inputs of the linear regression model (via the sliders) to see how their assumptions impact the predicted win percentage. This will help management think through future capacity planning. If management is considering investing in a major stadium capacity upgrade, they can adjust the stadium capacity slider and see the impact on their win percentage. The predicted win percentage value, located in the superior part of the graph section, will change automatically, as well as the blue dot representing the future state within the scatterplot. The customer will get a feel for the resulting change in magnitude and direction. Similar to the “Exploring the Game Data” section of the dashboard, the user can change the x-axis variable on the scatterplot.

In addition to the visualization, and to provide background on the methods and data, we will provide our customer with the code in GitLab. The following is the link to the GitHub:

https://github.com/Fan-Attendance-Project/Fan-Attendance-Repo

Once the customer approves the beta version of the dashboard, we can fully productionize the tool. This will include automatic data updates to include recent soccer match data. Additionally, after each professional soccer season, the model can be refitted and updated.

Summary of Findings

Throughout the project, the team set out to complete a series of objectives in order to inform soccer executives on how changing stadium capacity can impact their team's ability to win. The key to this is, the ability to measure using rigorous statistical methodology and data to determine the impact of the fans presence in a stadium.

The result of our hypothesis testing can be seen below:

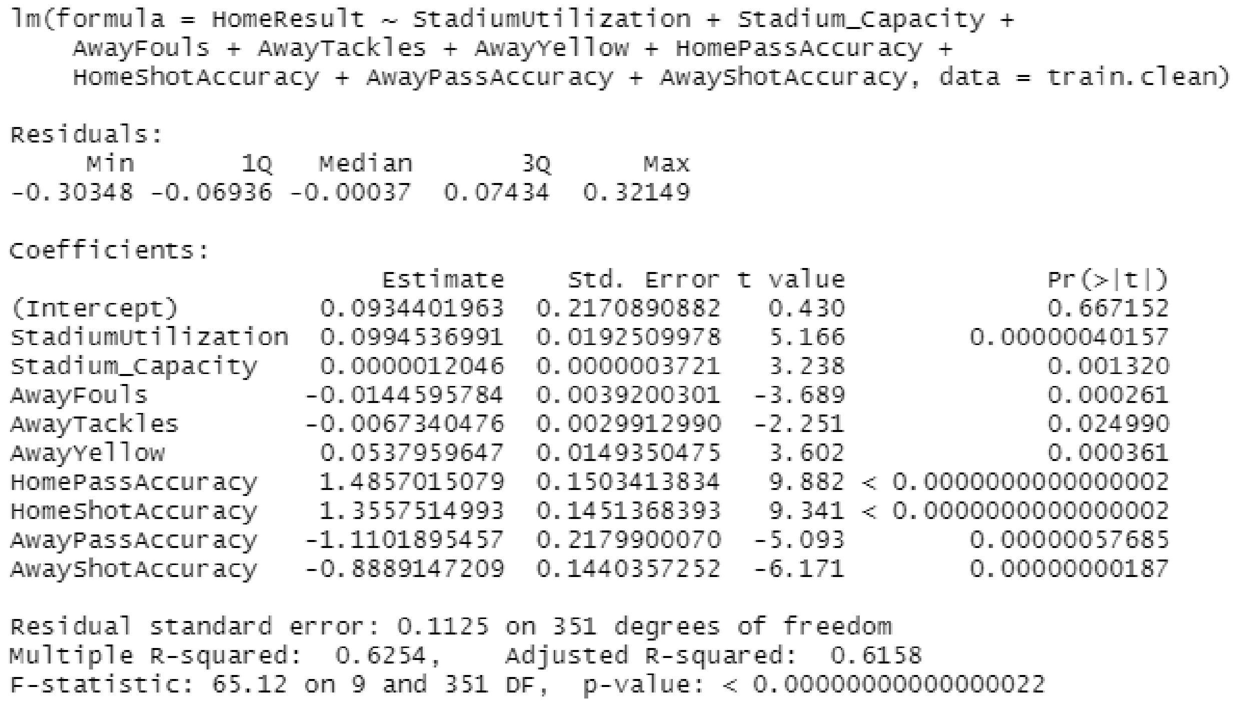
League	VAR significant?	Δ Win rate	Fans Significant?	Δ Win rate
ALL	Yes (P-value is 0.0159)	-0.029	Yes (P-value is 0.00026)	-0.049
Bundesliga	No (P-value is 0.1766)	-0.045	No (P-value is 0.0558)	-0.058
EPL	No (P-value is 0.802)	-0.007	Yes (P-value is 0.0019)	-0.084
La Liga	No (P-value is 0.61)	-0.01	Yes (P-value is 0.0359)	-0.055
Ligue 1	No (P-value is 0.394)	-0.02	Yes (P-value is 0.002)	-0.0879
Serie A	Yes (P-value is 0.047)	-0.059	No (P-value is 0.714)	-0.01

The results of testing are mixed: While generally, the effects of the Video Assistant Referee did have a statistically significant effect on home team win rate, we can see that the majority of this is due to the outsized effect that the deployment of VAR had on the Serie A. For all other leagues VAR showed no statistically significant effect on home team win rate. Conversely, generally and more crucially, fan attendance does have a significant effect on the home team win rate. On average, holding all other variables constant, not having fan attendance can drop a home teams' win percentage by almost 5%. That being said, Fans were shown to have a statistically significant effect on home team win rate in the EPL, La Liga, and Ligue 1, while Serie A and the Bundesliga show no statistically significant effect.

In all cases, regardless of significance, the home team win rate went down when fan presence was removed. This data validates the long standing belief that fans do make a difference to home team performance. That being said, The data does suggest that fans from different leagues also have differing effects on their home team's win rates.

Moving onto our modeling results, the objective here was to provide executives with a tool to be able to simulate expected seasonal home team win rates by adjusting various in-game performance statistics, as well as the stadium utilization and stadium capacity variables.

Below is the R output for our final model. For more visuals on the model performance please see the appendix (Figure 10):



```
lm(formula = HomeResult ~ StadiumUtilization + Stadium_Capacity +
    AwayFouls + AwayTackles + AwayYellow + HomePassAccuracy +
    HomeShotAccuracy + AwayPassAccuracy + AwayShotAccuracy, data = train.clean)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30348 -0.06936 -0.00037  0.07434  0.32149

Coefficients:
                      Estimate    Std. Error t value             Pr(>|t|)
(Intercept)        0.0934401963 0.2170890882   0.430             0.667152
StadiumUtilization 0.0994536991 0.0192509978   5.166         0.0000040157
Stadium_Capacity   0.0000012046 0.0000003721   3.238             0.001320
AwayFouls         -0.0144595784 0.0039200301  -3.689             0.000261
AwayTackles       -0.0067340476 0.0029912990  -2.251             0.024990
AwayYellow         0.0537959647 0.0149350475   3.602             0.000361
HomePassAccuracy   1.4857015079 0.1503413834   9.882 < 0.0000000000000002
HomeShotAccuracy   1.3557514993 0.1451368393   9.341 < 0.0000000000000002
AwayPassAccuracy  -1.1101895457 0.2179900070  -5.093         0.0000057685
AwayShotAccuracy  -0.8889147209 0.1440357252  -6.171         0.0000000187

Residual standard error: 0.1125 on 351 degrees of freedom
Multiple R-squared:  0.6254,    Adjusted R-squared:  0.6158
F-statistic: 65.12 on 9 and 351 DF,  p-value: < 0.00000000000000022
```

With an Adjusted-R^2 value of 0.6158, our model is moderately effective at predicting win rate, while simultaneously having highly interpretable results as stated in our Data Analysis and Modeling section.

We can see the effect of the independent variables on our model in the form of P-values. While both Stadium Capacity and Stadium Utilization are both statistically significant and have impact on win rates, the effect of stadium utilization is much stronger. In fact, only in game passing, and shooting accuracy metrics were better indicators of seasonal home team win rates. This tells us that while it is still important to have a large number of supporters in the stands, it is more important to ensure that the stadium is as fully utilized as possible, and there is a diminishing return on adding extra stadium capacity if you are unable to fill the seats.

For example, a stadium with a maximum capacity of approximately 35,700 seats but has a stadium utilization of 100% will win, on average all other variables constant, approximately 48% of its games. A stadium double that size, 71,400 with a stadium utilization of 50%, and all other variables held constant would win on average 48% of its games, resulting in no improvement in win rate. However, if the stadium capacity grew to approximately 48,000, while maintaining a stadium utilization of 100%, the home steam seasonal win rate will, on average all other variables held constant, will improve to 50%.

These insights empower executives to make better decisions around how much larger to make stadiums as well as make projections for expected seasonal win rates under varying stadium and team conditions.

Recommendations

We recommend that our dashboard be used by professional soccer league management to understand the variables that contribute to a soccer team's likelihood of winning matches. Specifically, leagues and individual teams should review the stadium capacity variable to determine the level of investment that would minimize cost and maximize wins. They can also use the stadium utilization variable to understand the effect of fan presence in relation to stadium capacity and determine how much to invest in marketing and sales. In addition to thinking about stadium capacity, leagues could also consider using the dashboard to understand how gameplay impacts teams performance. For example, the dashboard can help teams focus on minimizing fouls or improving pass accuracy by examining how those factors impact win percentage.

Future Research

While the work performed in this study is useful, further research can yield more insights on the impact of fans on in game professional sport performance.

For example, instead of testing to see if the win rate differs when fans are not present, we can test to see if the win rate is lower when fans are not present using a one sided two sample proportion test as opposed to a two sided one. Additional testing can be done on the loss rate or tie rate of home teams, to test if lack of fans means teams are more likely to either lose or tie game under certain conditions.

Going a step further, we can also test in-game performance metrics to see if, for example, more goals are scored when fans are in attendance, or if more fouls, yellow, or red cards are awarded when VAR is being used. Since VAR is a new technology, and changes to rules are also likely, this approach can also be used to see how rule changes on how VAR is used affects match performance.

This can also extend to other sports which were forced to play professional matches under similar conditions (i.e. in the same stadium simply without fans in attendance) assuming there are enough matches to meet the requirements of hypothesis testing.

From a modeling perspective, since our initial model focused on interpretability, some predictive power was sacrificed. For example, data transformations, like scaling or log transformations, were deliberately avoided. If applied to future models, it can possibly result in increased predictive power. Additional forms of regression or classification such as logistic regression or decision trees can be used to predict whether a match is won, lost or drawn given certain conditions.

From a data perspective, increased match data could be brought into the analysis, including individual player data which will allow us to study if players in certain positions or roles within the team are more likely to be affected by fan absence.

The most powerful benefit our project provides is its reproducibility and scalability. By storing all of our code, data, and material on a publicly accessible Github repository, we are open sourcing our project for use of this analysis across a wider variety and volumes of matches played around the world. This facilitates further exploration of the power of fan impact on match results.

Conclusion

We believe the model and visualization we built provide a useful tool that soccer leagues can employ in determining where to invest their money and energy in order to maximize wins. In conjunction with this tool, soccer leagues may want to consider incorporating financial data in future research. Incorporating financial data would help leagues anticipate how wins impact their bottom line and the overall success of their business. Additionally, we believe that other sports leagues can extrapolate on this research and incorporate our methods to understand how the same variables impact their win percentage.

References

Albert, Jim, et al. *Handbook of Statistical Methods and Analyses in Sports*. Boca Raton: CRC Press, Taylor & Francis Group, 2017, p. 407.

FBREF. *Sports Reference*, https://fbref.com/en/. Accessed 24 April 2022.

Martin, Christopher B. “EPL History, Part 1: Scraping FBref.” *chmartin.github,* [Scholarly Project], https://chmartin.github.io/2019/02/18/EPL-History-Scraping.html. Accessed 17 March 2021.

Pattnaik, Satya. “Testing a Difference in Population Proportions in Python” *Medium,* https://medium.com/analytics-vidhya/testing-a-difference-in-population-proportions-in-python-89d57a06254. Accessed 4 June 2022.

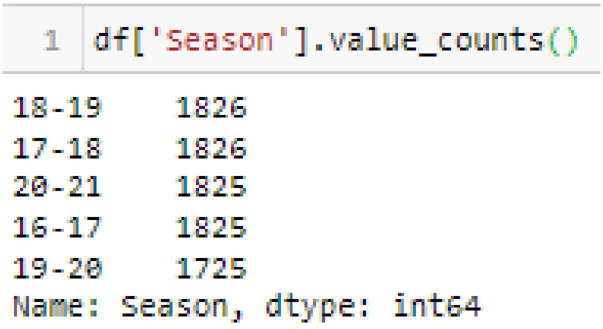
Vanderlaken, Paul. “10 Tips for Effective Dashboard Design by Deloitte.” *Disentangling Data Science*, ttps://paulvanderlaken.com/2020/09/29/10-tips-effective-dashboard-design -deloitte/. Accessed 2 June 2022.

“Video Assistant Referee (VAR) Protocol.” *The International Football Association Board,* https: //www.theifab.com/laws/latest/video-assistant-referee-var-protocol/. Accessed 4 June 2022.

Appendix

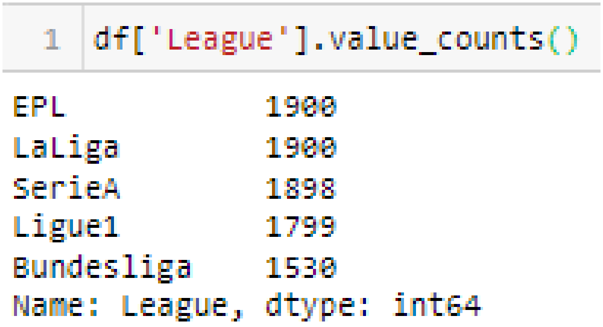
Figure 1

Number of Matches by Season



```
df['Season'].value_counts()
```

```
18-19    1826
17-18    1826
20-21    1825
16-17    1825
19-20    1725
Name: Season, dtype: int64
```

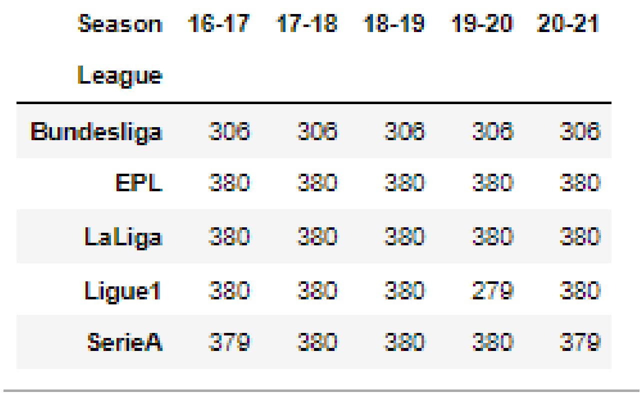


```
df['League'].value_counts()
```

```
EPL           1900
LaLiga        1900
SerieA        1898
Ligue1        1799
Bundesliga    1530
Name: League, dtype: int64
```

Figure 2

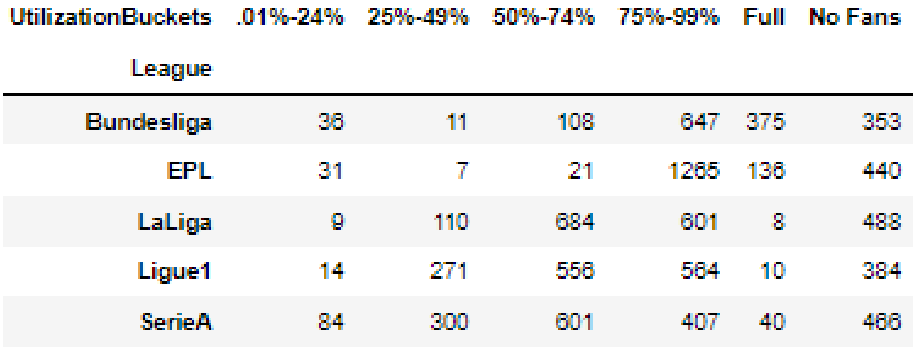
Number of Matches by Season and League



Season	16-17	17-18	18-19	19-20	20-21
League					
Bundesliga	306	306	306	306	306
EPL	380	380	380	380	380
LaLiga	380	380	380	380	380
Ligue1	380	380	380	279	380
SerieA	379	380	380	380	379

Figure 3

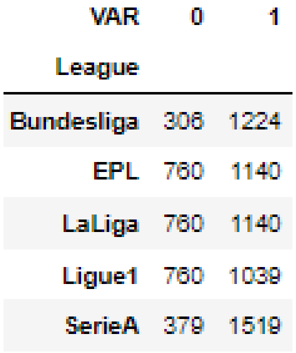
Stadium Utilization by league



UtilizationBuckets	.01%-24%	25%-49%	50%-74%	75%-99%	Full	No Fans
League						
Bundesliga	36	11	108	647	375	353
EPL	31	7	21	1265	136	440
LaLiga	9	110	684	601	8	488
Ligue1	14	271	556	564	10	384
SerieA	84	300	601	407	40	466

Figure 4

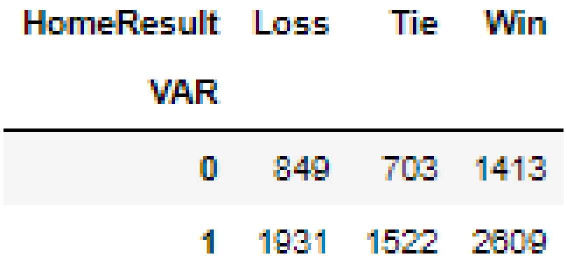
Presence of Video Assistant Referee in Matches:



VAR	0	1
League		
Bundesliga	306	1224
EPL	760	1140
LaLiga	760	1140
Ligue1	760	1039
SerieA	379	1519

Figure 5

Home Team Record Given Presence of VAR



HomeResult	Loss	Tie	Win
VAR			
0	849	703	1413
1	1931	1522	2609

Figure 6

Distribution of Independent Variables by League

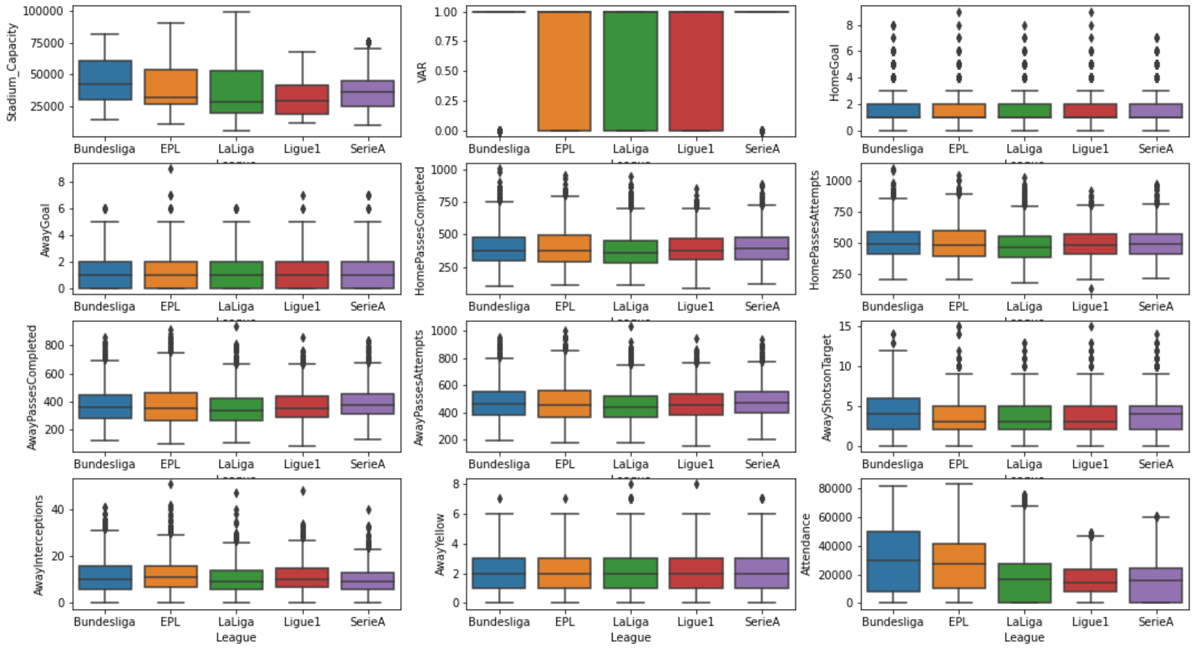


Figure 7

Correlogram

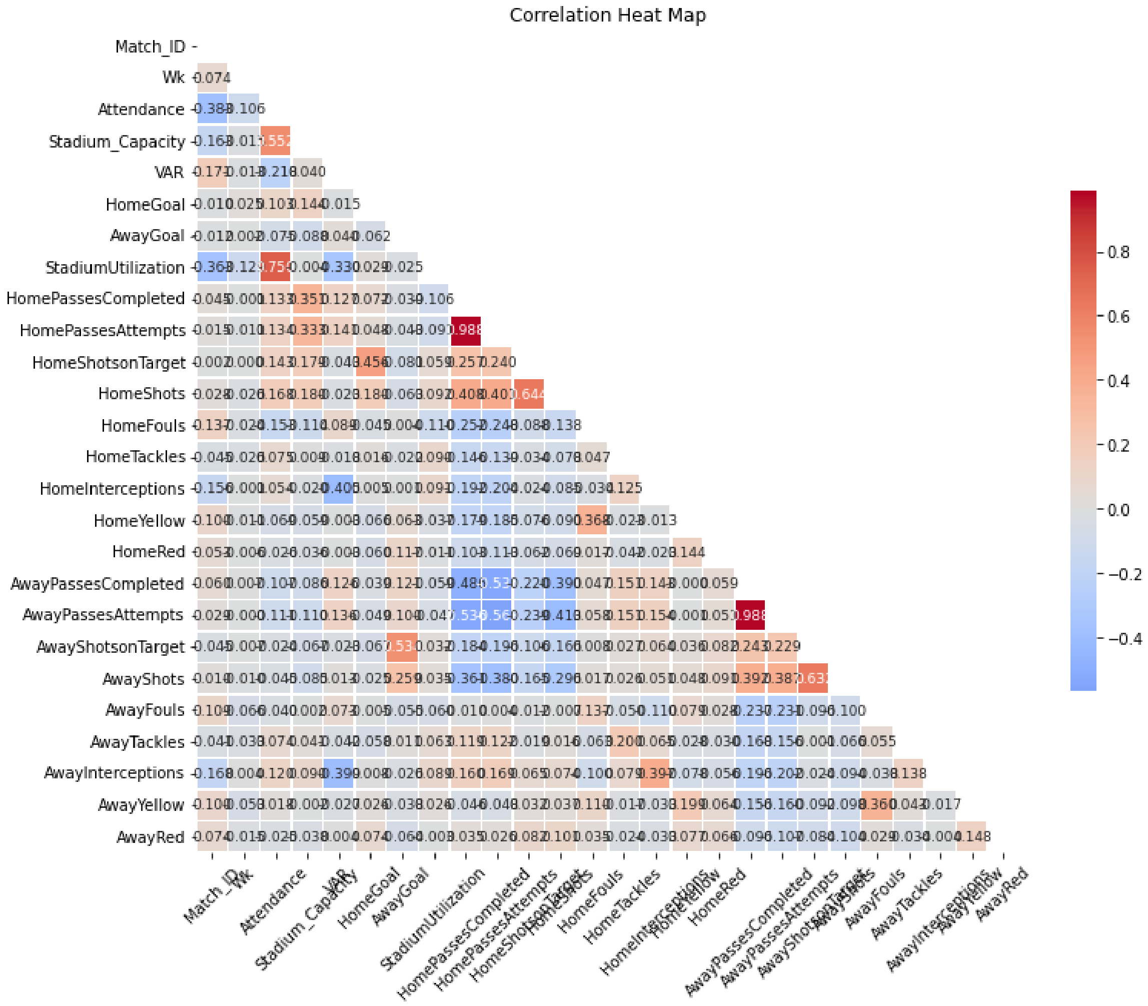
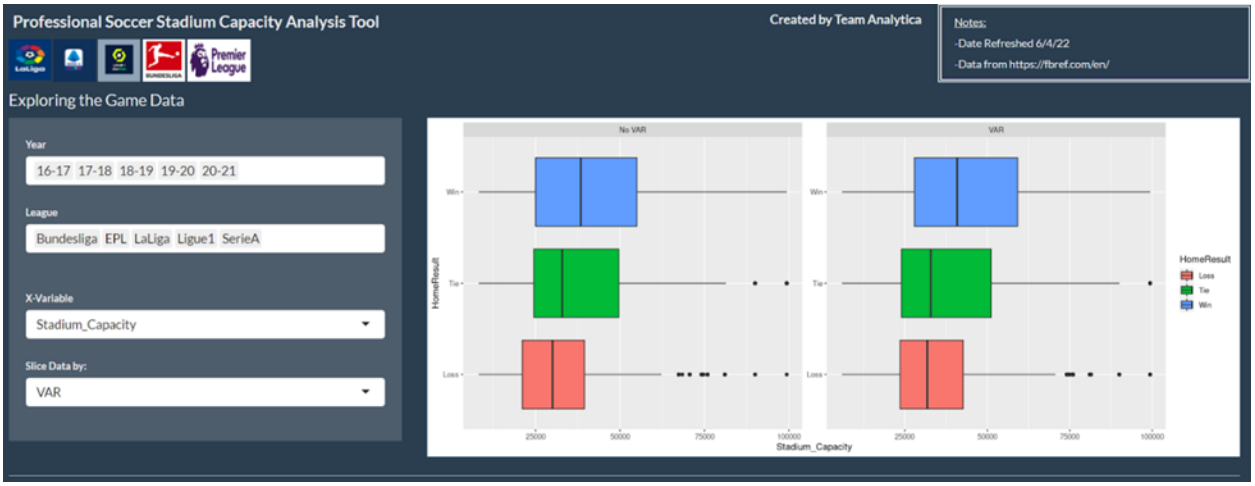


Figure 8

Dashboard (Exploring Game Data)



Professional Soccer Stadium Capacity Analysis Tool

Created by Team Analytica

Notes:
-Date Refreshed 6/4/22
-Data from https://fbref.com/en/

Exploring the Game Data

Year: 16-17 17-18 18-19 19-20 20-21

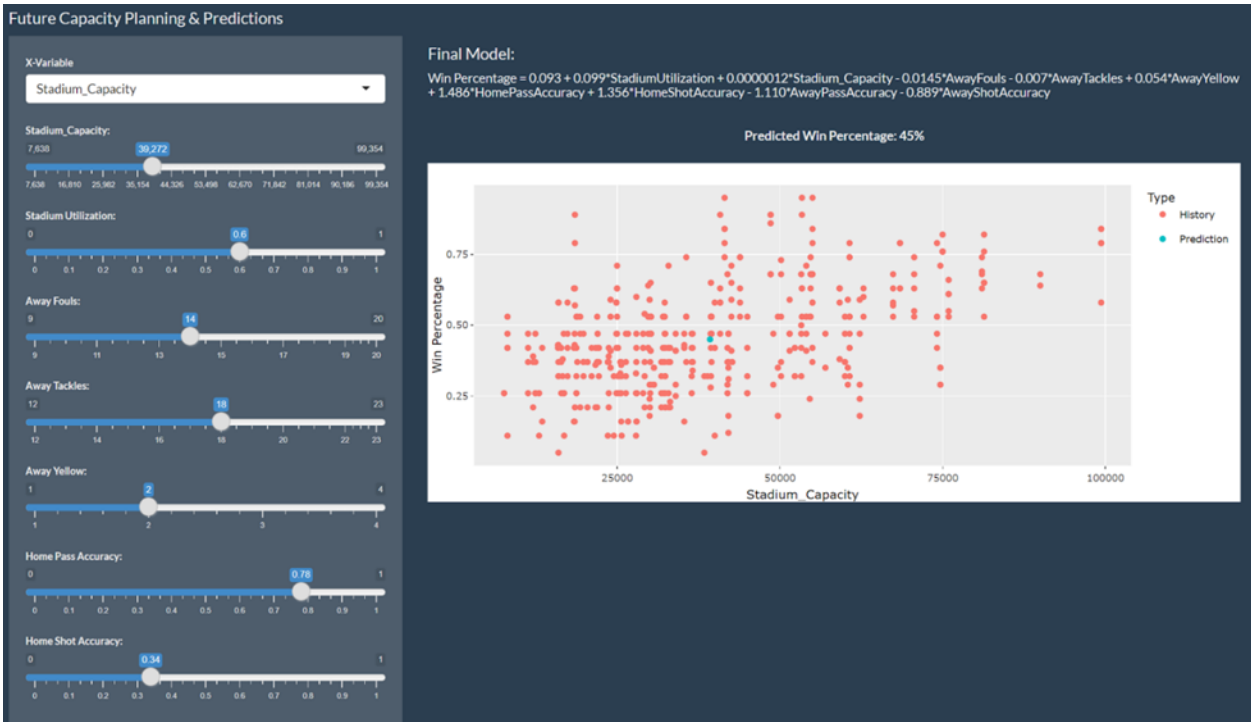
League: Bundesliga EPL LaLiga Ligue1 SerieA

X-Variable: Stadium_Capacity

Slice Data by: VAR

Figure 9

Dashboard (Future Capacity Planning)



Future Capacity Planning & Predictions

X-Variable: Stadium_Capacity

Stadium_Capacity: 39,272

Stadium Utilization: 0.6

Away Fouls: 14

Away Tackles: 18

Away Yellow: 2

Home Pass Accuracy: 0.78

Home Shot Accuracy: 0.34

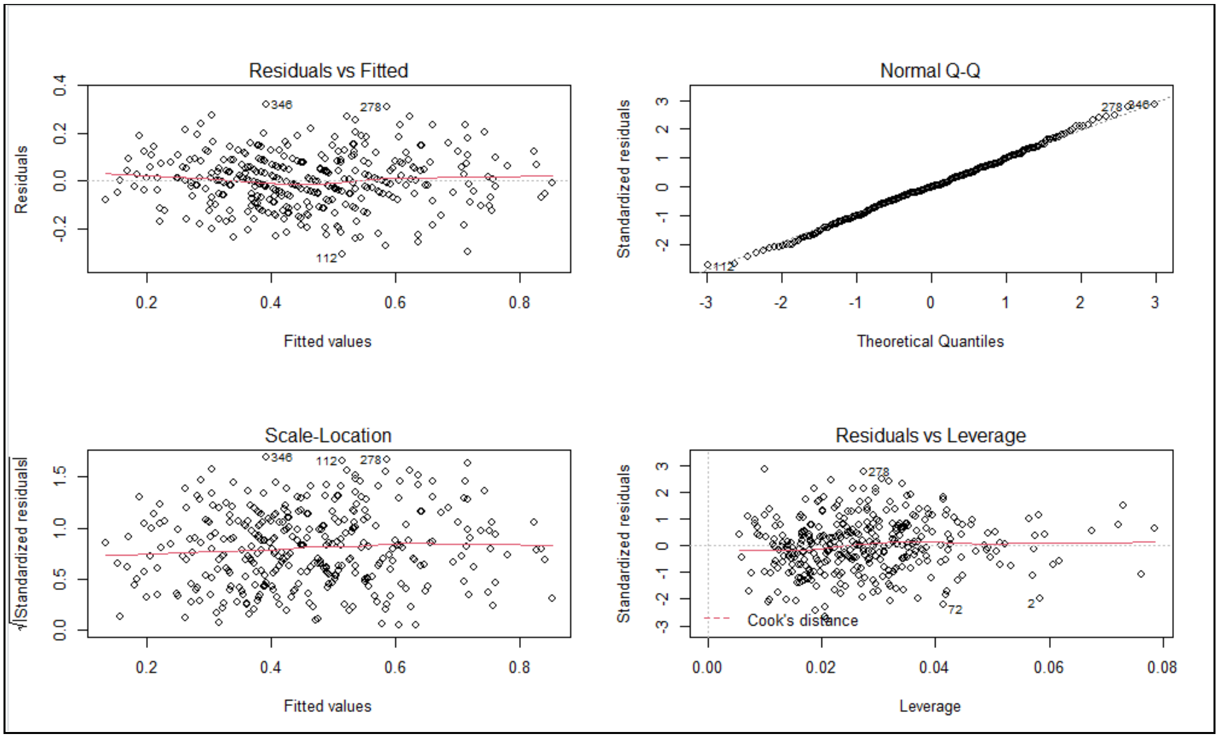
Final Model:

Win Percentage = 0.093 + 0.099*StadiumUtilization + 0.0000012*Stadium_Capacity - 0.0145*AwayFouls - 0.007*AwayTackles + 0.054*AwayYellow + 1.486*HomePassAccuracy + 1.356*HomeShotAccuracy - 1.110*AwayPassAccuracy - 0.889*AwayShotAccuracy

Predicted Win Percentage: 45%

Figure 10

Checking Linear Assumptions


Residuals vs Fitted
Residuals
Fitted values
346
278
112
Normal Q-Q
Standardized residuals
Theoretical Quantiles
278
346
112
Scale-Location
√|Standardized residuals|
Fitted values
346
112
278
Residuals vs Leverage
Standardized residuals
Leverage
278
72
2
Cook's distance