**How do match stats help shape the final results?**

Match statistics are an important part of the data collected by football teams and the authorities. These statistics can be used to effectively establish relationship between various kind of outcomes. The match statistics that we are going to discuss about include Full Time Home Goals, Fulltime Away Goals, Full Time Results, Half Time Home Goals, Half Time Away Goals, Home Shots, Away Shots, Home Shots on Target, Away Shots on Target, Home Fouls, Away Fouls, Home Corners, Away Corners, Home Yellow cards, Away Yellow cards, Home Red cards, and Away Red cards.

Chart, calendar

Description automatically generated with medium confidence

All Statistics Correlation

As we can see from the correlation plot, a significant number of home and away statistics is related to the full-time result. There are no highly correlated variables throughout the dataset, and this is established through the above correlation plot. The method used for the correlation plot is Spearman’s because the data is not normal, and Spearman’s is a non-parametric correlation test.

**Do home statistics affect the match outcome?**

The match is being played either in home stadium or away stadium. But whenever a match is being played at home stadium, it is a general notion that the odds are in favor of the home team. The home fan’s termed as the ‘twelfth man’ as they add to the team performance. Teams perform better when they are playing in front of their home crowd.

Calendar

Description automatically generated

Home Statistics Correlation

For answering whether the home statistics turn the odds in favor of the home team, we first separate the home statistics variables from all statistics. This isolates the home statistics and gives us better idea of effects of home statistics on full time result. Then we check Variance Inflation Factor (VIF) to further check the multi collinearity. Ideally, the VIF should be less than 5 to ensure no multi collinearity but below 10 is acceptable as VIF between 5 and 10 means that there is mild multi collinearity. In our case all the variables have VIF less than 5 meaning there exists no multi collinearity between the variables.

A picture containing text

Description automatically generated

VIF of Home Statistics Variables

Since there is no multi collinearity, we can proceed with Linear Multiple Regression analysis for home statistics with target variable as full-time results. First, we get the coefficients for establishing the relation between feature variables and target variables, which in this case is home statistic variables and full-time results.

A picture containing chart

Description automatically generated

Coefficients of Home Statistics for Linear Regression Model

A general notion about coefficient is that positive means that it affects the target variable positively and vice-a-versa. In this case, it is true since FTHG (Full Time Home Goals) has the maximum coefficient which is true as more full-time goals tips the result in the favor of the home team. Whereas HR (Home Red cards) means a loss of player. This clearly is a disadvantage for the home team and hence has a negative effect on the end results as it tips the result in favor of the opponent or away team. Further, it is also interesting to prove with data that just taking shots towards goal does not pressurize the away team, rather taking shots on target does help the result to tip in the favor of home team. This analysis is very important as the results from this analysis can help the home team strategize their plan to win a game keeping the home statistics under control as home statistics are heavily influenced by the home team. Thus, if the home team score more goals, take more shots on target, and carefully play the match without getting a red card, the chances of home team winning look very promising.

Just from the coefficient we can establish very significant relationship. The model summary for home statistics dives deeper and explains more about the effect of home statistics on full time result.

A screenshot of a computer

Description automatically generated with low confidence

Model Summary for Linear Regression Model of Home Statistics

The most important metric that we see here is the multiple R-squared value which is 0.4118. This means about 41% of the outcome is explained by home statistics. The hypothesis that we assumed in the conclusion is effectively proved by this model as the FTHG (Full Time Home Goals), HST (Home Shots on Target), and HR (Home Red card) have relationship with our target variable i.e., FTR (Full Time Results). In conclusion, out of all home statistics, the above mentioned three statistics significantly affect the results.

Moving on to diagnostic plots for the Linear Regression Model. The diagnostic plots help us understand the Linear Regression Model in much deeper sense. The diagnostic plot talks about various metrics like overfitting, residual, normality etc., and can be easily interpreted by only seeing the graphs.

The first plot that we see is the Residual vs Fitted plot. The main goal of this plot is show if there is a linear relationship between our predictor variable and target variable. In this case, it is very clear from the first plot that there exists a linear relation to some extent but overall, there is no linear relationship.

Next, we look at the second plot which Normal Q-Q plot. This plot shows whether the residuals are normally distributed or not. In this case, the residuals are normally distributed with some exceptional cases that can be clearly seen near the origin of the plot.

The third plot is the Scale-Location plot. To put it in simple term, this plot shows how the residuals are spread or in statistical term, Homoscedasticity. Although, the residuals are not spread randomly, i.e., they do follow a pattern, but the line is not totally horizontal. This indicated the presence of random spread of residuals within our model.

The last plot is Residuals vs Leverage. This plot serves the purpose of identifying influential outliers within our total spread of the residual values. In this case, we see that all the outliers are within Cook’s distance and hence there is no presence of influential outliers.

Chart, scatter chart

Description automatically generated

Home Statistics Diagnostic Plot

Now we move on with our second research question.

**Do away statistics affect the match outcome?**

Here we look at the away statistics of the match being played at the opponent’s home stadium. We are trying to establish a relationship between the away statistics and match outcome in such a way as to give fair chance of opponent to win the match.

Calendar

Description automatically generated

Away Statistics Correlation

It is very interesting to note that the away statistics affect the result in negative way. Which is true as better the away team plays, less are the chances of home team winning. This relationship can be clearly established from the above correlation diagram. To add to this, FTAG (Full Time Away Goals), HTAG (Half Time Away Goals), AS (Away Shots), and AST (Away Shots on Target) are the factors that heavily influence the match outcome. These variables have the ability to alter the match results and can easily tip the match outcome in away team’s favor. Further the away red card has a positive correlation here which is very similar to the home statistic correlation. This is because if any of the team, be it home or away loses an on-field player due to red card, then the other team will heavily benefit from this and may have a chance to win the match.

The correlation plot also shows that there are no variables that are highly correlated. Hence, there is no need to eliminate any variable. We now take a look at VIF (Variance Inflation Factor).

Text

Description automatically generated with medium confidence

Away Statistics VIF

The VIF is below 5 and well below 10. This shows that there is no multi collinearity within the dataset and gives a green signal that the variables are good to be used as predictor variables for our Linear Regression Model. Hence, we move on with the available variables and model a Linear Regression with target variable as FTR (Full Time Result).

Letter

Description automatically generated

Away Statistics Coefficients

The coefficients just reiterate on the conclusions that we drew from the correlation plot. The FTAG (Full Time Away Goals) and AS (Away Shots) work against the home team as they negatively affect the outcome and push it in away team’s favor. The AR (Away Red cards) work in favor of the home team and hence have a positive coefficient.

Looking at the summary of Linear Regression Model with Away Statistics as the predictor variable and FTR (Full Time Result), we see very similar summary statistics. The Multiple R-squared is roughly 42%. That means about 42% of the outcome is explained by our model. The residual error is also less. But here we see three significant variables. When we look closely, we can see that significance codes are being marked by ‘\*’ (asterisk) beside the predictor variables. The p-value can also be used to find out the significant variables. In case of away statistics, FTAG (Full Time Away Goals), AS (Away Shots), and (Away Corners) are the three variables that effect the match outcome. Although the numbers of this model are very similar to the numbers shown by home statistic model, the explaining variables are different. In case of away statistics, away corners play an important role. This single outcome can be very much influential in turning around the results in the favor of away team. The away team can strategize their game plan as to take more shots and get maximum number of corners to pressurize the home team. This is again very true in real life soccer scenarios.

Table

Description automatically generated

Model Summary for Linear Regression Model of Away Statistics

To explain the model better, we now look at the diagnostic plot of the regression model. This will specifically talk about the model and not about the research questions.

Chart

Description automatically generated

Away Statistics Diagnostic Plot

From the first plot, i.e., the Residuals vs Fitted plot, we conclude that there is a partial linear relationship between predictor variables and target variables. But overall, the model does not show a linear relationship.

From the second plot, i.e., Normal Q-Q plot, we can see that the residuals are fairly normal with a handful of exceptional outliers that are in the top left corner of the plot. Overall, the residuals are normally distributed.

From the third plot, i.e., Scale-Location plot, we see that the spread follows a pattern, but it is still random. This proves the existence of Homoscedasticity within our model, as the spread is random and is difficult to determine.

The fourth plot, i.e., Residuals vs Leverage shows that there are no influential outliers. We need to check whether an outlier is influential or not before we eliminate it. In this case, there are couple of outliers but since they lie within the Cook’s distance, they are not that influential. Hence, no influential outliers exist in our residuals.

**What conclusions can we finally draw from the analysis?**

The conclusions are very sensible when we talk practically about soccer as a sport. Hypothetically, if home statistics heavily influence the match outcome, then almost every home match should have been won just by focusing on home statistics which is not the case. Similarly, if the match results are heavily influenced by away statistics, then every away team can avoid the face of defeat by focusing on the statistics that heavily affect the match outcome. But again, that is not what we see in real life scenario. A match is fair and square when we talk only about home statistics and away statistics. A model that consists of both the home and away statistics is much more efficient in determining the match outcome.

Text

Description automatically generated with low confidence

All Statistics Model Summary

Here, we can see that combined statistics of both home and away teams are able to explain about 74% of the match outcome. This is a significant jump from 42%. This is evidence that no match is single handedly affected by home or away statistics. Rather home and away statistics both affect the match outcome. The home team must perform better, play high intensity, and cautious football to win the match. If the away team succumbs to this pressure, eventually the away statistics will take a hit as they will take less shots and eventually score less goals. This will ensure the home team victory. But as we concluded above, there are some factors that can help the either of the team perform better. Like for home team, they can push their team to take more shots on target rather than just wasting opportunities and playing carefully without committing too many fouls. On the other hand, the away team can focus more on pressurizing the home team by taking more shots and winning more corners. This will slightly improve their odds of winning but in no one can ensure that either of the team wins.

Hence, we conclude that there are handful of home and away statistics that affect the Full Time Result. But in reality, both the set of factors have to be considered simultaneously to better calculate the influence of the match statistics on the Full Time Result.