Team Hyperborea : Sports | Project Proposal

When it comes to sports, specifically soccer, home and away advantage can be a big factor on many outcomes and statistics during a match. For example, a home team might have home advantage and the referee might be in favor of them and that could lead to less fouls and cards called against them and more for the away team. Many athletes, like ourselves at some point, have stated that they rather play at home due to the atmosphere and not having to worry about traveling and getting exhausted of going to an away field.

"Home field advantage" is a concept in sports that has ties to emotions of the fan base and anecdotal support. I want to be able to report whether data supports this effect, but also which statistics are predictive of this difference to point a casual fan's eye towards a data-based "why." This also could help advise folks on when to tune into a team they might only casually support. If this difference is pronounced enough, a fan could be advised to avoid watching matches in which Liverpool is playing away - since results would be negative, and thus less enjoyable for that fan.

This project will investigate whether or not there is an advantage to playing at home or away for the European soccer team Liverpool. Specifically I hope to answer an experimental question with a null hypothesis that models return home statistics when Liverpool are playing home that are equal to away statistics when Liverpool is away, with an alternative hypothesis that my models return home statistics that are better than away statistics for a variety of response variables.

I will be using a combination of datasets provided by Footy Stats that has collected stats from soccer matches dating all the way back to the 2007/2008 season to the end of the 2020/2021 season. These data came in several separate .csv files that have been combined and gone through a processing stage, leaving us with 36 variables, 32 of which are numeric, 4 of which are categorical or descriptive (i.e. "team name") and 532 rows. Those rows are individual

observations of Liverpool matches only. The unprocessed data (which some may return to) has 65 variables and 5320 rows.

Using the EPL Matches data, I hope to answer several questions regarding the relationship between home field advantage or away field disadvantage and goals scored per match. I also want to look at other variables affecting the total goal count for example: attendance and total goal count, Liverpool corner count home and away, Liverpool fouls home and away, Liverpool possession home and away. Using the Liverpool Matches data, I hope to answer the following questions:

- What kind of impact do referees have for Liverpool when the team is at home and away?
    - Fouls afford and against them
    - Total yellow and red cards given
- How did the PPG change from playing home and away?
- What was their goal count at home? Away goal count?
- What was the possession at home vs away?

PPG is a variable that tracks a team's run of form on a moving average basis. Teams earn three points a match for a win, one point for a draw, and no points for a loss. For each team, their run of form is tracked from the beginning of my data set up until the most current match played. Teams with high PPG have a strong correlation to winning games. I will build models that focus on highlighting this correlation. In particular he will focus on adding weight to other top sides in the league. In games where there is a top five team going against a lower table team, the home/away advantage may not mean much because the talent of the better side is so much better. Because Liverpool is a top side, they usually go into games as a favorite with the exception of games against other top sides. Binning teams at the lower end of the table, where Liverpool have a significantly higher PPG in the matchup, and focusing on teams that Liverpool is more equally matched with has the potential to alleviate this and provide valuable insights as to how a close PPG affects the team when they are playing at home versus away.

I will work focusing on evaluating models heavily concerned with the categorical variable “referee” to see if a dummy variable analysis will reveal trends that differ in home and away stats. Specifically with an eye on the response variables: home_team_fouls and away_team_fouls, but also running a model of home_ppg and away_ppg, home_team_goals and away_team_goals, etc. If time, he will zoom out to the full data set (including non-Liverpool matches) to try to model various home and away response variables, using techniques of multiple regression analysis. Following any of this model building, those models will be used on the Liverpool stats isolating rows “when they are the home team” for the home stats vs. rows “when they are the away team” for the away stats to see if the model results differ in favor of one direction or another. This process may extend in future work from the full group on expanding to the full data and other teams besides Liverpool.

Attendance is a major benefactor for the home team in any sport. Teams with stronger home field advantage tend to win more games at home. I will build models with the prime directive to explore this aspect of the dataset. The response variable will be home_team_goal_count and away_team_goal_count while also keeping in mind other variables such as home_team_posession, away_team_posession, and total_goal_count. The model building will focus largely on Liverpool attendance at home and the correlation to the home team goal count and see if playing at home has an advantage or if there is no difference between playing at home and away.

Possession means that a team or player is in control of the ball. A player (and his team) are in possession while he has the ball, or is passing to or receiving from another member of his team. If he loses the ball, he loses possession. One of the best things about having good possession numbers is how much it can frustrate the opposing team. Not only do teams get frustrated with the fact that they can’t get the ball, they will often make multiple mistakes in an attempt to regain possession. The response variables will be home_team_red_cards, away_team_red_cards and home_team_yellow_cards, away_team_yellow_cards. I will build models that focus on possession correlation of home and away.

xG is basically expected goals measures the quality of shots on several variables such as

assist type, shot angle and distance from goal, whether it was a headed shot and whether it was defined as a big chance. Here there are two teams, home team and the away team, the xG which is the expected goals scored by each team, the team with higher expected goals makes the strong correlation. To explore this part of the dataset, I will create models displaying that correlation. The focal point of the model is that it is the best single metric for understanding performance of a soccer team and predicting it's future. The response variables will be home_team_shots_on_target and away_team_shots_on_target and also total_goals_at_half_time. The model will concentrate mainly on Liverpool expected goals.

Home field advantage is an oft-cited “fact” of sports. This study will evaluate whether or not it is present in the data and report back findings of the facts via statistical analysis. The regression analysis will tell us information about how well home and away statistics are being predicted by my features. These models will then be used by the group to gain insights on the presence (or lack thereof) of an edge. The authors leave it to those that might read the study on what decisions might be affected by this analysis, but hopefully this exploration deepens the enjoyment of the sport for those that read it. Sports is a much discussed topic that can open conversation among diverse sets of people, so I hope a statistical backed exploration of the actual items that might - or might not - contribute to home-field advantage contributes to conversations and insight to those looking for an interesting insight to offer a group of football fans - or perhaps give someone a unique, statistically supported foothold to claim a reason Liverpool’s recent success is only due to an “advantage”.