The Impact of Home Climate on Soccer Performance
CIS4930: Data Mining

Mason Russo, Daniel Dang, Miguel Montesinos, Zachary De Aguiar, Merzan Aymaz
https://github.com/danieltdang/Impact-of-Home-Climate-on-Soccer-Performance

I. Introduction

The world of soccer is affected by many factors, such as the player's current form, head-to-head record, or even if they are playing at their home turf. The latter is most similar to our research question, and we believe it is a decisive factor in the outcome of a match: Does the weather and climate of a soccer player's home country influence their performance in a game played under different environmental conditions? Drived by our curiosity, we set out to understand how an athlete's familiarity with specific weather patterns - temperature, humidity, wind speed, and more - might shape their capabilities in a game played in contrasting climates.

This research stems from the observation that players often experience variable performance levels when competing away from their native climatic conditions. Our goal is to explore whether there is a tangible connection between these fluctuations and the athletes' climatic acclimation.

In professional sports, success is based on a multitude of factors ranging from individual players to team strategy. One underlying aspect however that may significantly shape soccer performance is environmental conditions. Throughout this paper, we seek to uncover the potential implications of a soccer team's performance based on the home team’s climate.

Soccer on the surface is not merely a competition of physical ability, but an amalgamation of variables that contribute to the outcome of a game. One of the variables in particular is the concept of home-field advantage. This allows players to feel more relaxed due to familiarity of their surroundings and not having to experience any travel fatigue (opposed to the away team). However, the impact of the climate within a team’s home stadium remains unstudied and rarely noticed.

II. Literature Survey

The study from Périard, Racinais, and Sawka explores the topic of heat acclimation and its implications on competitive athletes and sports performance. They traverse various adaptations and mechanisms that cause heat acclimation while also explaining the physiological processes. The paper thoroughly analyzes the applications of heat acclimation on athletes, how it can enhance performance, and reduce the risk of heat-like illnesses. The authors offer a broad synopsis of the field; detailing benefits for heat acclimation, and offer practical considerations for athletes performing sports or physical activities in demanding environmental conditions. From the study, the following were found as conclusions, “Exercise heat acclimation induces physiological adaptations that improve thermoregulation, attenuate physiological strain, reduce the risk of serious heat illness, and improve aerobic performance in warm-hot environments.” (Périard et al. 20-38)

In the research by Peñas and Ballesteros, they examined many performance factors to see how the location of a game will affect a team’s performance. They confirm that ”more successful technical and tactical indicators would be performed at home compared to away” (Peñas and Ballesteros 468–69). These indicators would include: goals, shots, passes, crosses, etc. This would mean that many characteristics that determine how well a player did in a game had increased rates playing at home compared to away. This was further confirmed when they found out that in “ the 2008-2009 season of the Spanish League, 61.95% of the games were victories for the home teams” (Peñas and Ballesteros 467). As the probability of winning in home territory is over 60%, it is fair to suggest that there is an advantage of some sort playing in the home stadium. While the research paper examines the increased statistics and probabilities at home, they do not examine what could possibly lead to this, especially nothing related to weather and climate. Thus, we use this paper to compare the probability globally and also to use the same type of statistics for comparisons between home and away.

McSharry's paper reveals that physiological performance in international football games is significantly impacted by altitude, affecting both aerobic and anaerobic activities. Higher altitudes are correlated with decreased performance levels, marked by alterations in strategy, pacing, and overall physical exertion. This can be analyzed by pulling out an insert from the paper, it states, "The surprising result is that the high altitude teams also had an advantage when playing at low altitude, benefiting from a significant advantage over their low altitude opponents at all locations." (NLOM). We can expand on the idea that it is feasible to calculate an advantage between two teams based on historical data. Although the paper highlights how altitude influences performance, a detailed exploration of how varying altitude levels impact teams originating from different base altitudes could offer nuanced insights. Furthermore, intertwining the effects of altitude with weather patterns and their collective impact on performance would expand the research framework, correlating it with a multifactorial approach toward understanding performance in different environmental contexts.

III. Methodology

In our scenario, the objective is to analyze and predict the performance of Home and Away teams in a match. To accomplish this, we reframed from using the traditional classification problem, instead utilized a regressor model. This shift allowed us to derive meaningful data that would allow us to visualize performance instead of predicting a binary outcome of a win or loss.

The idea of not turning this into a classification still allowed us to articulate a way to score the data based on binary outcomes but also visualize the expected performance. We were able to still tell who won the game by simply choosing the higher player average from the prediction. This method gave us a way to score the predictions against the real results.

In order to get a good idea of how this is possible, we want to explore how the Random Forest Regressor works. It's essentially built on the principles of ensemble learning, which utilizes the predictions from multiple decision trees to produce an accurate prediction. Each tree is built with a random subset of data and features in order to reduce overfitting which helps in preventing the model from becoming tunnel visioned on only the training data.

Another model that was implemented to predict the Away Average Rating was the multiple linear regression model. This model would use the entire dataset, with a 75/25 split between training and testing data size. The model would then be evaluated by using the Mean Absolute Error, Mean Square Error, and Root Mean Square Error, to determine whether the model is predicting the Away Average Rating accurately. The multiple linear regression model outputs an intercept and coefficients, which is used to understand the relative importance of each feature in predicting the Away Average Rating.

After using these two models, we also decided that analyzing the data retrieved with a statistical methodology would further help to indicate what features influenced the performance ratings of each player the most. The analysis of variance (ANOVA) was our choice of analysis, as it is used to assess the significance of the relationship between the three features that we had against the target variable, the Away Average Rating. To be more specific, the ANOVA's main statistics to look at are the F value and the P value. If the F value is high and the P value is lower than a significance level of 95%, we can safely reject the null hypothesis and suggest that that feature does have a significant relationship with the Away Average Rating.

The dataset was also visualized using box plots, scatter plots, a distribution plot, and a heatmap. This was done to visualize data distribution, skewage, outliers and how the Away Average Rating was related to the features.

IV. Implementation

In order to start the project, it was important to search for meaningful data that could be used to help us support the case that players are actually impacted by climate. We encountered a problem of getting individual reports of players' performance to average together and get the total team's performance. In this case, it was important to find the most accurate ratings without any bias. We could either create our own algorithm taking into account tons of attributes from a player's report from a specific game, or outsource the player ratings from a reputable company called FotMob.

Unfortunately, FotMob's API did not have public documentation, but it was easy to figure out using Chrome Tools Network Activity tab to gather the correct endpoints needed to gather all related information. In Python, we implemented a basic web scraper utilizing the request library in order to handle all 'GET' requests. It was also required to gather the exact elevation using a public API called open-elevation.com to send the longitude and latitude as parameters, to get the elevation as a response. From the longitude and latitude, we were also able to retrieve the average ambient Temperature (c), specific Humidity (g/kg), and the Koppen Geiger Climate classifier using a Python library called pvcz, which stands for Photovoltaic Climate Zones.

With all the data gathered, we simply imported it straight into a CSV file with all attribute columns correctly formatted to start training the model. It was important to have the correct format to be able to open the files using Pandas to manipulate the data further.

Random Forest Regressor Model

To successfully begin classifying the data, we started by utilizing the Random Forest Regressor Model to predict average player ratings. As previously stated, it was important for the model to output predicted average player ratings instead of binary 0s and 1s. This was important because it still allowed us to turn the output into wins/losses by simply picking the higher player rating average as the winner.

Multiple Linear Regression Model

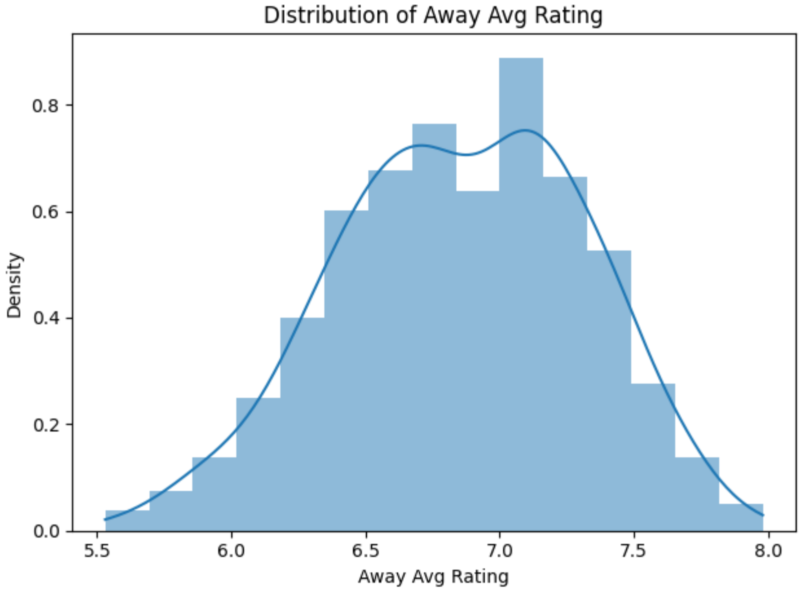
After the Random Forest model, we moved onto implementing visualization statistics along with Multiple Linear Regression. Using the Seaborn and Pandas libraries, we were able to split the data into the features and the target variable (Away Average Rating) for use in analyzing the data. The box, distribution, scatter, and heatmap plots were all created using the seaborn library that plots out everything when the features and target variable were inputted and then saved using the matplotlib library, specifically the pyplot module to create subplots whenever there were multiple plots for that type of plot. From there, the sklearn library was used to implement multiple linear regression, where the data was split into 75% training data and 25% testing data. Then, the LinearRegression method was called to fit the training data, where the intercept and coefficients for each feature were calculated and printed out to the user. The multiple linear regression model was then evaluated using the metrics module that was a part of sklearn, by producing the Mean Absolute Error, Mean Square Error, and Root Mean Square Error. This was important to understand the accuracy of the model and also to see the relationship between the features and Away Average Rating.

ANOVA Test

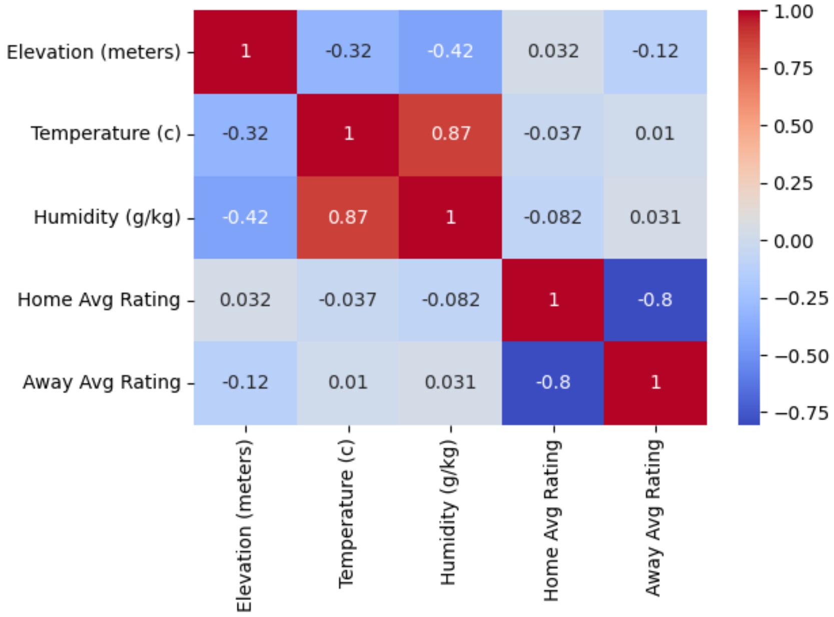
In the ANOVA implementation, the use of a new library called statsmodels was used. This library is used for statistical modeling, and we imported the Ordinary Least Squares (OLS) function to base ANOVA on. The data was imported as a DataFrame using pandas, and then standardized so that there were no special characters in the features and target variable name, due to a limitation with ANOVA where special characters will be recognized as a new line for a new feature. Using OLS, we were able to fit the model to take in the data and evaluate it based on feature and Away Average Rating, to which this model would then be inputted into statsmodels using the anova_lm function, the linear model, it would return the complete ANOVA test based on the features and Residual along with their sums squared, F value, and P value.

V. Data Visualization

Distribution Plot

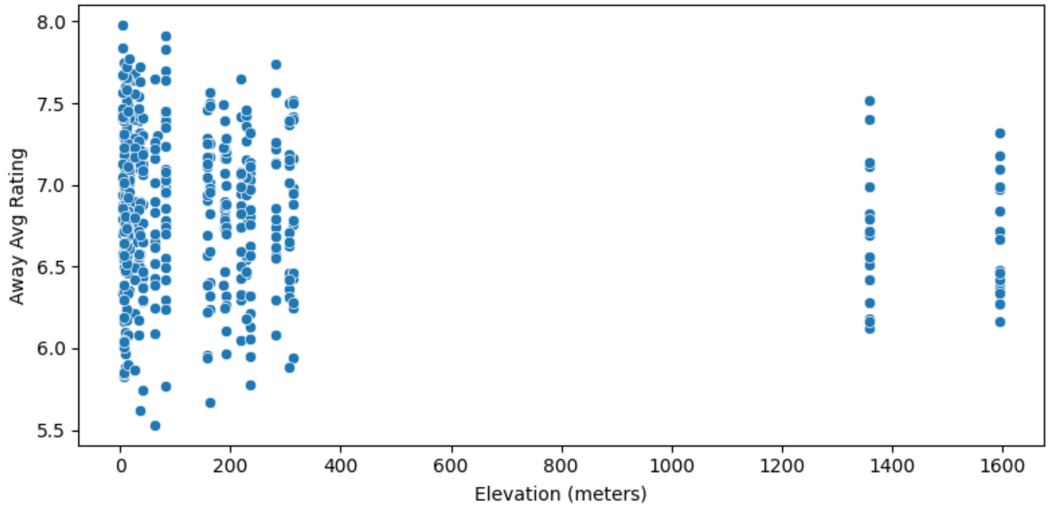


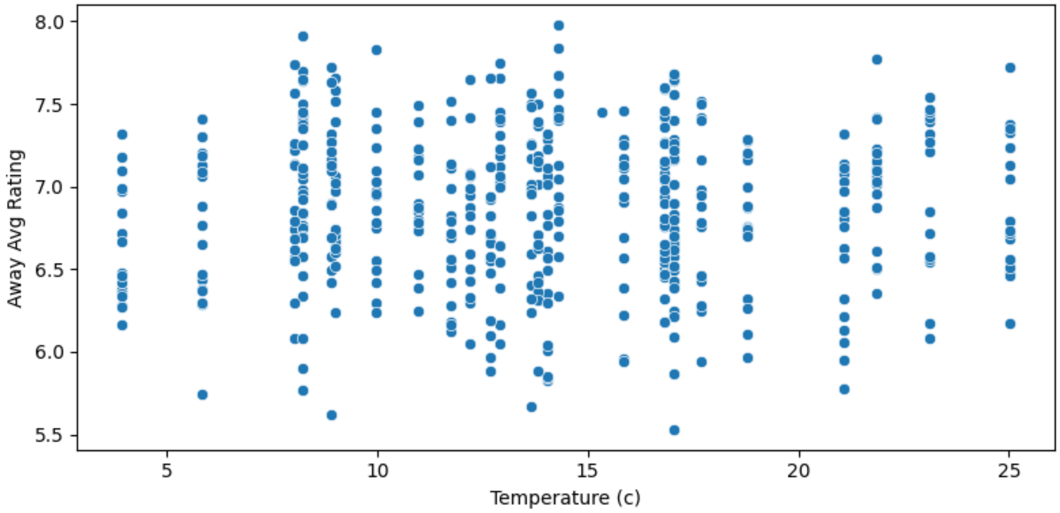
Heatmap

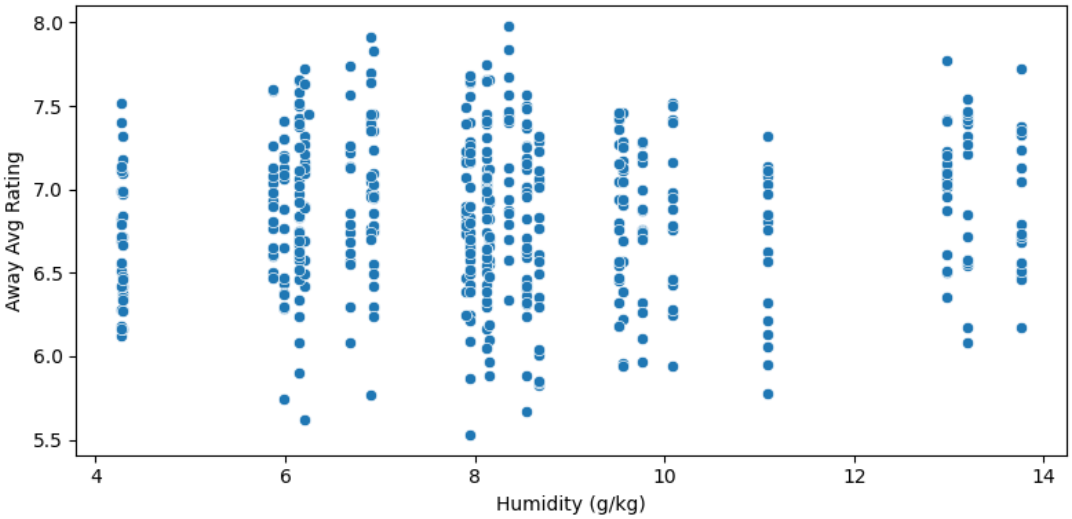


	Elevation (meters)	Temperature (c)	Humidity (g/kg)	Home Avg Rating	Away Avg Rating
Elevation (meters)	1	-0.32	-0.42	0.032	-0.12
Temperature (c)	-0.32	1	0.87	-0.037	0.01
Humidity (g/kg)	-0.42	0.87	1	-0.082	0.031
Home Avg Rating	0.032	-0.037	-0.082	1	-0.8
Away Avg Rating	-0.12	0.01	0.031	-0.8	1

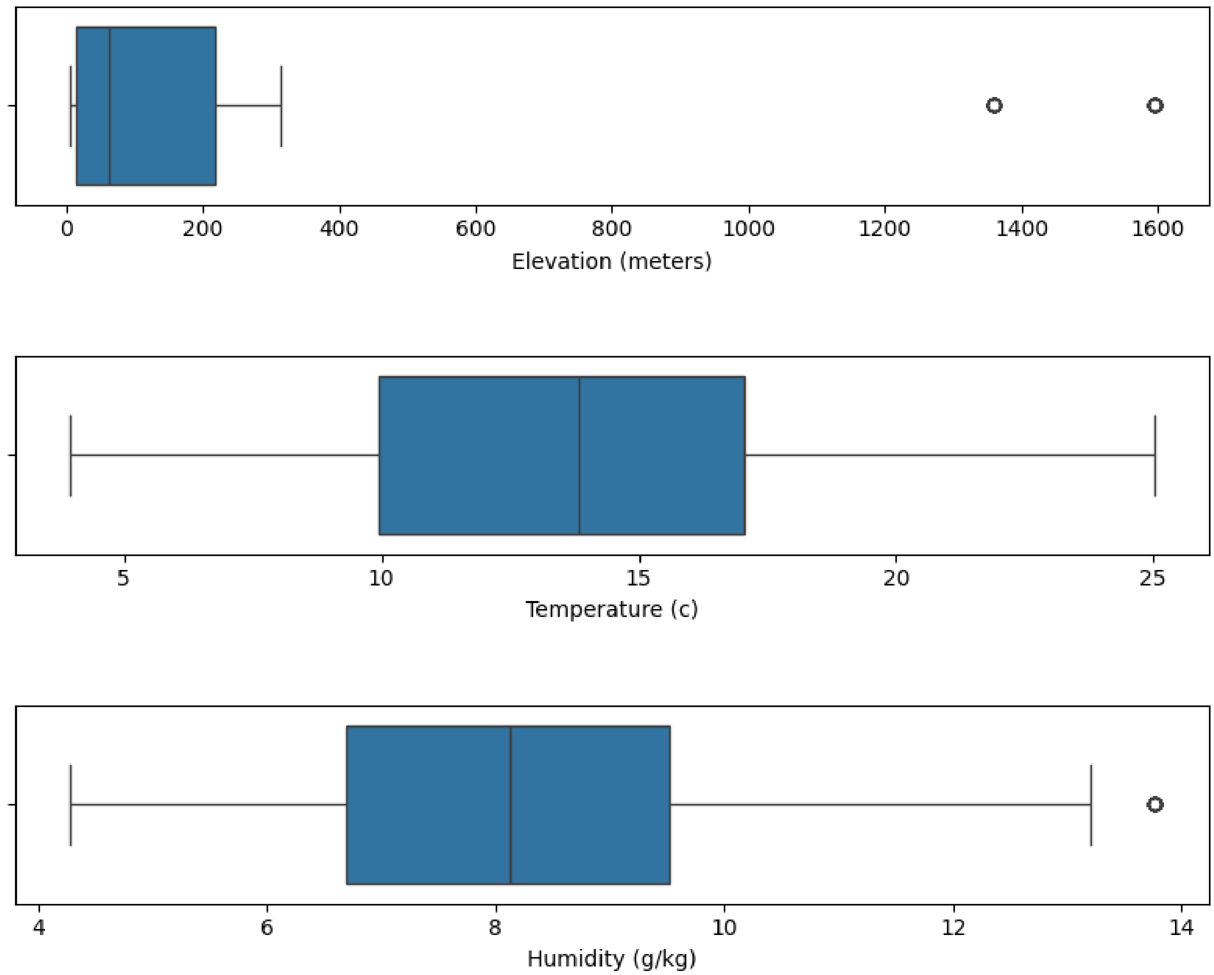
Scatter Plots







Box Plots



VI. Visualization Interpretation

With a quick glance, the Box Plot shows the density of the average player ratings during the 2022 MLS Season.

Utilizing the scatter plots, we can easily visualize the large variance of player ratings when put up against attributes. With the Elevation being on the x-axis, it can be observed that there's some interesting discrepancies. Not one team was predicted to be over 7.5+ average player rating and this could mean a lot of things. We can infer that away teams are affected by elevation or the quality of the team playing as the home team.

Looking at the Temperature and Humidity on the x-axis, it can be inferred that teams playing in temperatures less than 10 Celsius or greater than 20, could be affected as well.

Based on the extremes of the heatmap, we can see relationships that would be a given, like how humidity and temperature have a really high positive correlation to each other. However, we see that the Home Average Rating and the Away Average Rating have a high negative correlation with each other, with -0.8. This indicates that when one team does good, the other team drastically does bad, implying that most games result in ratings on both teams that aren't usually close to each other.

VII. Results

In order to properly gauge if climate had an impact on a specific team, we were forced to only train and test one season at a time. This negated any outside factors like roster changes or long-term injuries.

Random Forest Model Evaluation

True Positives (TP)	27
False Positives (FP)	21
True Negatives (TN)	69
False Negatives (FN)	30
TP/(FP+FN)	0.529412

If an away team is predicted to win, which matches up with the actual result, we consider that a true positive. The same goes with a home team, we consider it a true negative. We decided to hone in on the away team being a critical measure, hence why the roles are flipped in our evaluation.

Using the 2022 MLS Season, the calculated accuracy of the model correctly predicted 65.31% of the 147 games that were extracted from that season. Something to note, it seems as though there's potentially a bias towards the home team. This is because the model predicted that the home team would win 67.35% of the time, which is significant given that the odds should only be a 50/50 split.

Multiple Linear Regression Model Evaluation

Intercept	6.86875
Coefficients (Elevation (meters), Temperature (c), Humidity (g/kg))	-0.00016047 -0.0027959 0.0059171
Mean Absolute Error (MAE)	0.36544
Mean Square Error (MSE)	0.19727
Root MSE (RMSE)	0.44416

The intercept of the multiple linear regression model indicates the baseline prediction when all the features are absent. The coefficients for each feature represent how much the Away Average Rating is influenced by the feature when the feature is increased by 1 unit. This means that when the elevation, temperature, and humidity increases by their respective unit and no changes to the other features, the Away Average Rating would then change by -0.00016047, -0.0027959, and 0.0059171 respectively. From these coefficients, it is suggesting that humidity affects the Away Average Rating the most, as the range of humidity is 4.27 to 13.766, which is a much smaller range than elevation and temperature, having much higher ranges with a much lower coefficient.

To assess the multiple linear regression model's accuracy in its predictions, the MAE, MSE, and RMSE were calculated and provided the errors of 0.36544, 0.19727, and 0.44416 respectively. As all three of these error evaluations were low, it can be said that the models were very accurate.

ANOVA Test Evaluation

	sum_sq	F	PR(>F)
Elevation	1.417594	6.697849	0.009943
Temperature	0.033545	0.158493	0.690723
Humidity	0.001415	0.006684	0.934873
Residual	102.438202	NaN	NaN

From the ANOVA results, there's a major dissimilarity between the Elevation and other attributes. The PR(>F) value for elevation is significantly low, which suggests that it has a substantial effect on the match outcome. Typically, an attribute value under 0.05 indicates that the factor has a statistically significant effect. To drive it home, the F-Statistic value is considerably higher than those for Temperature and Humidity.

VIII. Conclusion

Originally, our thoughts on the subject of performance in different climates would be that it were going to be more favorable for the home team, given that the climate for the home team is the climate they reside the most in, whether it is through playing games, practicing, or living in said climate. The subject of this experimentation would be Major League Soccer (MLS), specifically the 2022 season. This made the most sense in terms of leagues to test due to the MLS being based in the U.S., and the U.S. being a much larger country and thus having larger variance in climate for traveling teams than in other smaller countries in Europe and around the world.

Upon our research, the data gathered implies that there is in fact an advantage for the performance of the team playing on their home soil as opposed to the performance of the visiting team. Our models, the Random Forest Regressor and the Away Average Rating, as well as utilizing the analysis of variance (ANOVA) as our analysis of choice were the pillars of our experimentation, with player ratings from FotMob as our barometer for player performance.

With these pillars in place, the data it retrieved from the sampled 147 games revealed a 65.31% correct prediction of the games, with humidity seemingly having the biggest effect on Away Average Rating compared to other factors such as elevation and temperature. With regards to future research directions, there is something to be said about the advantages of playing at home outside of just the climate. Our model did present potential bias towards the home team, as it predicted victory for the home team 67.35% of the time, and this could imply that there are larger, or at least more factors in play for the performance of the teams other than whether or not they are acclimated to the climate they are playing in. Future research into this topic should stem away from the climate as a sole measure of performance and instead focus on the other aspects of playing in an away stadium such as crowd atmosphere and even travel distance, which are not directly related to the climate. To ensure that climate is of little relevance in terms of its effect on performance in this research, data could only be pulled from games that are played in subjectively “good” weather (little altitude etc.) or closed-roof stadiums between two teams that do not play in largely differing climates. From there, other factors such as crowd atmosphere and travel distance can be best measured in terms of player performance, and can even be taken a step above with higher intensity games such as playoff games or cup games.

References

Aughey, Robert J., et al. "Soccer activity profile of altitude versus sea-level natives during acclimatisation to 3600 m (ISA3600)." *British Journal of Sports Medicine*, vol. 47, no. Suppl 1, Nov. 2013, pp. i107-i113.

Illmer, Sarah, and Frank Daumann. "The effects of weather factors and altitude on physical and technical performance in professional soccer: A systematic review." *JSAMS Plus*, vol. 1, Oct. 2022, p. 100002.

Lago-Peñas, Carlos, and Joaquin Lago-Ballesteros. "Game Location and Team Quality Effects on Performance Profiles in Professional Soccer." *Journal of Sports Science and Medicine*, vol. 10, no. 3, Feb. 2011, pp. 465-471.

McSharry, Patrick E. "Effect of altitude on physiological performance: a statistical analysis using results of international football games." *BMJ*, vol. 335, no. 7633, Dec. 2007, pp. 1278-1281.

Peel, M. C., et al. "Updated world map of the Köppen-Geiger climate classification." *Hydrology and Earth System Sciences*, vol. 11, no. 5, Oct. 2007, pp. 1633-1644.

Périard, Julien D., et al. "Adaptations and mechanisms of human heat acclimation: Applications for competitive athletes and sports." *Scandinavian Journal of Medicine & Science in Sports*, vol. 25, May 2015, pp. 20-38.