Home Field Advantage in the NFL

An analysis of the Impact of Geographic and Environmental Factors on Game Outcomes
Using Machine Learning

Michel Robert
Computer Science
University of New Mexico
Albuquerque, NM
Mrobert12@unm.edu

Daniel Morales-Garcia
Computer Science
University of New Mexico
Albuquerque, NM
dmorales33@unm.eduu

ABSTRACT

Home field advantage has been long discussed in the sports community. While the comfort of the home team competing in a more familiar setting is acknowledged, our study looks deeper at a multitude of factors to determine their influence on determining game winners of NFL games.

We began with an extensive data collection process in which we found an existing data source and added to it from a few different API's. Using various big data techniques we were able to refine our dataset for comprehensive analysis. This process included extensive data cleaning where we ensured accuracy and consistency of our dataset. It also included normalization of our various features due to the difference in scale between them.

Following the cleaning of our data we focused on the creation of the specific features we would be using in our analysis. We were able to build a data set specifically customized with, what we believed, were the key elements influencing game outcomes. This important step laid the groundwork for all the analysis techniques we planned to employ.

We implemented various machine learning models to help in our analysis. These models included ordinary least squares regression(OLS), logistic regression, random forest analysis and a neural network model, using these models we aimed to predict game outcomes and determine which of our features were the most crucial in making these predictions.

Through our employment of our machine learning techniques we determined that record difference was the most impactful measure in predicting the winner of games. Of our constructed features, temperature-humidity interaction and temperature-elevation interaction were the most impactful in our predictions. Logistic regression achieved an 80% accuracy rate in its predictions. Due to the high mean square error in this model we developed a neural network model which achieved an 80% prediction accuracy.

Our study underscores the role played by the selected features representing home field advantage in predicting NFL game winners.

1 Data

We began data collection by looking for existing datasets that may contain relevant data to our research. We found an existing data set[1] containing NFL game data from 1966 to 2023 with data for 13,298 games. The data included the date of games, the home and away teams as well as the scores for both teams, the stadium where the game took place as well as some weather data for the game.

Alongside this dataset was one containing information about the stadiums, including the town where the stadium is located, stadium capacity, latitude longitude, and elevation data.

Using python's pandas library we began cleaning this data by filling in the large amount of missing data in the latitude longitude and the elevation data. There were a few towns missing and were filled in by hand. Using the tows and google cloud console's API we were able to obtain the missing latitude longitude data and through the latitude longitude get the elevation data and timezone data. These values were updated in our stadium dataset and had to be translated to our main dataset for every game. We did this by matching the teams who were playing in the individual games to the teams labeled in the.

While our initial dataset had some weather data about the home town we wanted to be able to compare weather between home and away towns. We used open-mateo's API from which we were able to obtain temperature and humidity data from the latitude longitude and date values.

1.1 Features

There were a number of features that we determined would be useful for predicting the winner of NFL games. We started with geographical differences including: distance between the two teams' towns, difference in elevation between the towns and number of timezones traveled by the away team. These were pretty straight forward to calculate distance between the two towns was obtained through google's distance API given the two towns latitude and longitude coordinates. The distance traveled can give us insight on jet lag and travel fatigue. We created a binary feature for win and loss by the home team 0 for loss 1 for win as well as relative score difference which also calculates the winner of the game just not in a binary manner.

Elevation, which was obtained previously, was calculated by subtracting away towns elevation from the home towns elevation. Elevation difference may not be as big a factor in most games as most stadiums are quite close to sea level but for stadiums like mile high stadium in denver this can play a big role as the stadium is located at over 5000 ft which can cause away teams extra challenge as they aren't accustomed to the extra strain the elevation can cause.

The four different time zones were given numerical values from 0 to 3. 0 was given to the pacific time zone, 1 to the mountain time zone,2 to the central time zone, and 3 to the eastern time zone. The number of timezones traveled was calculated by taking the absolute value of the difference of the two town's time zones. For example traveling from the pacific time zone to the easter time zone yields a 3 time zone difference. Time zone differences, like distance traveled, can cause jet lag and disrupt a player's circadian rhythm.

The next group of features was weather related. We had two weather related features: temperature and humidity. These values were a bit more difficult to compare as the comparison is not as straightforward as geographical data. We would have liked to take the weekly average of temperature data but due to constraints of the open-mateo API we had to limit our number of queries. What we ended up doing was comparing the weather from the game day in the home and away cities. For both the temperature and the humidity values we subtracted away team values from home team values.

The last feature that we used that we thought was directly related to home field advantage was difference in stadium capacity, this metric can have an impact on pressure of games and fan influence on the game.

Some features we included after initially running our models to help get a better idea on how much influence each feature had on the model were team record differences and their point differences. These made the models significantly better at predicting the outcome of games as it gives an idea of who the better team is before the game is even played.

We also looked at the interaction between multiple of our crafted features. There were a total of nine of these interaction features: time zone and elevation, time zone and temperature, temperature and humidity, temperature and distance, temperature and elevation, elevation and distance, stadium capacity and distance, stadium capacity and elevation and finally timezone and distance. We created these to improve our models performance, and to highlight combined effects of our features.

Since many of our features are on different scales (distance measured in miles, elevation in meters and our output being binary etc.) we normalized all of our features, using min-max normalization, to be able to more easily compare the data and to hopefully reduce overfitting.

2 Methodology

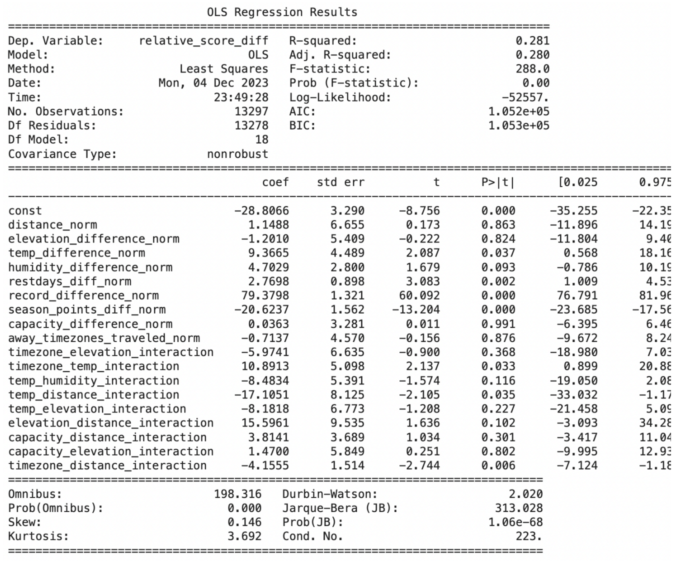
This research aimed to dissect the multifaceted nature of the home field advantage in NFL games, analyzing various factors such as travel, weather, and elevation differences impact team performance. The methodology encompassed several statistical and machine learning techniques, employing a comprehensive dataset of every NFL game from 1966 to 2023.

2.1 **Initial Statistical Analysis**

In our initial exploratory analysis, we focused on quantifying the home field advantage in the context of NFL games. This involved calculating the overall win percentages under various conditions. Firstly, we determined the win percentage of home teams across all games in our dataset. The analysis revealed that home teams won 57.11% of the time, suggesting a notable advantage. Furthermore, We analyzed the win percentage of the favorite team, irrespective of playing at home or away. The favorite team emerged victorious 65.94% of the time. We delved deeper into the scenarios where the home team was also the favorite to win. In such cases, the favorite home team's win percentage increased to 67.53%, reinforcing the concept of a home field advantage. These initial statistics set the stage for more complex analysis, providing a foundational understanding of the home team advantage in NFL games.

2.2 **Linear Regression Analysis**

Building upon the initial insights, we employed Ordinary Least Squares (OLS) regression to examine the relationships between various normalized factors and relative score difference in NFL games. The predictors included factors such as the normalized differences in distance traveled, elevation, temperature, and humidity between the competing teams locations, along with other game-specific variables like rest days, team record and season points differences. The OLS model's R-squared value of 0.281 indicated that approximately 28.1% of the variance in the relative score difference could be explained by our model.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     relative_score_diff   R-squared:                       0.281
Model:                             OLS   Adj. R-squared:                  0.280
Method:                  Least Squares   F-statistic:                     288.0
Date:                 Mon, 04 Dec 2023   Prob (F-statistic):               0.00
Time:                         23:49:28   Log-Likelihood:                -52557.
No. Observations:                13297   AIC:                         1.052e+05
Df Residuals:                    13278   BIC:                         1.053e+05
Df Model:                           18
Covariance Type:             nonrobust
=====================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975
-----------------------------------------------------------------------------------------------------
const                             -28.8066      3.290     -8.756      0.000     -35.255     -22.35
distance_norm                       1.1488      6.655      0.173      0.863     -11.896      14.19
elevation_difference_norm          -1.2010      5.409     -0.222      0.824     -11.804       9.40
temp_difference_norm                9.3665      4.489      2.087      0.037       0.568      18.16
humidity_difference_norm            4.7029      2.800      1.679      0.093      -0.786      10.19
restdays_diff_norm                  2.7698      0.898      3.083      0.002       1.009       4.53
record_difference_norm             79.3798      1.321     60.092      0.000      76.791      81.96
season_points_diff_norm           -20.6237      1.562    -13.204      0.000     -23.685     -17.56
capacity_difference_norm            0.0363      3.281      0.011      0.991      -6.395       6.46
away_timezones_traveled_norm       -0.7137      4.570     -0.156      0.876      -9.672       8.24
timezone_elevation_interaction     -5.9741      6.635     -0.900      0.368     -18.980       7.03
timezone_temp_interaction          10.8913      5.098      2.137      0.033       0.899      20.88
temp_humidity_interaction          -8.4834      5.391     -1.574      0.116     -19.050       2.08
temp_distance_interaction         -17.1051      8.125     -2.105      0.035     -33.032      -1.17
temp_elevation_interaction         -8.1818      6.773     -1.208      0.227     -21.458       5.09
elevation_distance_interaction     15.5961      9.535      1.636      0.102      -3.093      34.28
capacity_distance_interaction       3.8141      3.689      1.034      0.301      -3.417      11.04
capacity_elevation_interaction      1.4700      5.849      0.251      0.802      -9.995      12.93
timezone_distance_interaction      -4.1555      1.514     -2.744      0.006      -7.124      -1.18
==============================================================================
Omnibus:                      198.316   Durbin-Watson:                   2.020
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              313.028
Skew:                           0.146   Prob(JB):                     1.06e-68
Kurtosis:                       3.692   Cond. No.                         223.
==============================================================================
```

Results of OLS Regression

The regression coefficients provided valuable insights into how each factor influenced game outcomes. For instance, record difference norm, a measure of the teams' performance disparity, showed a substantial impact with a coefficient of approximately 79.38, suggesting that teams with better season records had a significant advantage. In contrast, factors like distance and elevation difference had minimal influence, as indicated by their low coefficients and high p-values.

Residual plots were generated to evaluate the model's fit, displaying the differences between observed and predicted values. The spread of residuals helped in identifying any potential patterns or outliers, ensuring the robustness of our model. The linear regression analysis was crucial in quantifying the extent to which various environmental and situational factors influenced the outcomes of NFL games, providing a more nuanced understanding of the home field advantage phenomenon.

2.3 Hyperparameter Tuning

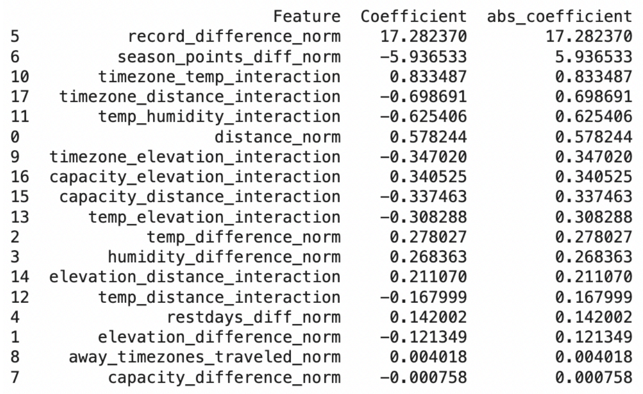
To refine the predictive performance of our logistic regression model, we employed GridSearchCV, a systematic approach for tuning hyperparameters. This technique is crucial in machine learning models to find the most effective parameters, which can significantly impact the model's accuracy and efficiency. In our analysis, the hyperparameter 'C', which determines the strength of regularization in logistic regression, was varied across a range of values [0.001, 0.01, 0.1, 1, 10, 100].

GridSearchCV evaluated the model's performance for each value of 'C' using cross-validation, ensuring a comprehensive search over the parameter space. The process aimed to identify the optimal balance between bias and variance, thereby enhancing the model's ability to generalize to new data. The best parameter obtained was 'C'=100, with an improved score of 55%, indicating that a lower regularization strength led to better performance in our model. This fine-tuning of hyperparameters was a vital step in optimizing our logistic regression model, ensuring its robustness and reliability for predicting home team wins in NFL games.

2.4 Feature Importance Analysis

Following the optimization of our logistic regression model, we conducted a feature importance analysis to discern which predictors had the most significant impact on forecasting the outcome of home team wins. This analysis provides crucial insights into the dynamics influencing game results, highlighting the most influential factors.

The logistic regression model's coefficients were scrutinized to understand the relative importance of each feature. Features with higher absolute coefficient values were deemed more influential in predicting the home team's victory. The analysis revealed that 'record_difference_norm', representing the disparity in teams' performance records, was the most influential feature with the highest coefficient. This finding suggests that the historical performance of the teams was a critical determinant in predicting game outcomes.



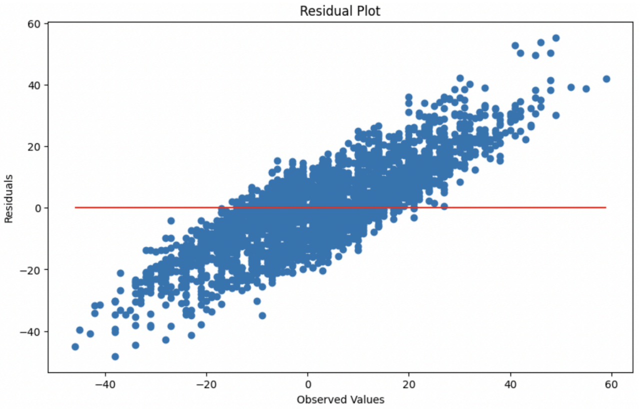
```
                          Feature  Coefficient  abs_coefficient
5          record_difference_norm    17.282370        17.282370
6           season_points_diff_norm   -5.936533         5.936533
10       timezone_temp_interaction    0.833487         0.833487
17   timezone_distance_interaction   -0.698691         0.698691
11       temp_humidity_interaction   -0.625406         0.625406
0                    distance_norm    0.578244         0.578244
9   timezone_elevation_interaction   -0.347020         0.347020
16  capacity_elevation_interaction    0.340525         0.340525
15   capacity_distance_interaction   -0.337463         0.337463
13      temp_elevation_interaction   -0.308288         0.308288
2             temp_difference_norm    0.278027         0.278027
3         humidity_difference_norm    0.268363         0.268363
14  elevation_distance_interaction    0.211070         0.211070
12       temp_distance_interaction   -0.167999         0.167999
4              restdays_diff_norm    0.142002         0.142002
1        elevation_difference_norm   -0.121349         0.121349
8      away_timezones_traveled_norm    0.004018         0.004018
7         capacity_difference_norm   -0.000758         0.000758
```

Feature Importance Results

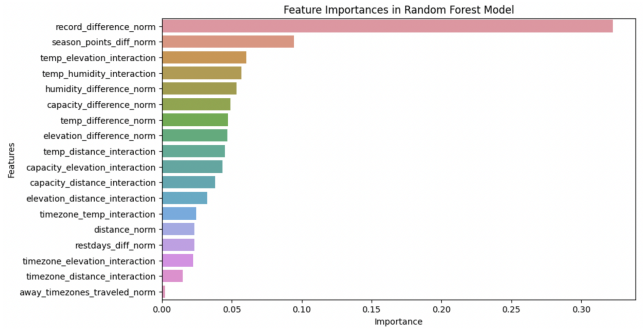
Other notable features included 'season_points_diff_norm' and 'timezone_temp_interaction', illustrating the importance of a team's seasonal performance and the interaction between time zones traveled and temperature differences. In contrast, features like 'distance_norm' and 'elevation_difference_norm' held less significance in the model, as indicated by their lower coefficients. This feature importance analysis was instrumental in understanding the relative impact of various factors on home team victories, providing strategic insights that could be leveraged by teams and coaches in game planning and preparation

2.5 **Random Forest Regression**

In our pursuit to enhance the predictive accuracy of our model, we employed a Random Forest Regressor, a robust machine learning technique known for its ability to handle complex datasets with multiple variables. This method was particularly suited for our NFL game data, which involved numerous interacting factors. The Random Forest model was trained to predict the relative score difference based on our set of predictors, including normalized differences in distance, elevation, and weather conditions. The effectiveness of this model was evaluated using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), which are standard metrics for assessing regression models. The MSE obtained was 155.22015992481204, and the corresponding RMSE was 12.458738295863352, providing a quantitative measure of the model's prediction accuracy.



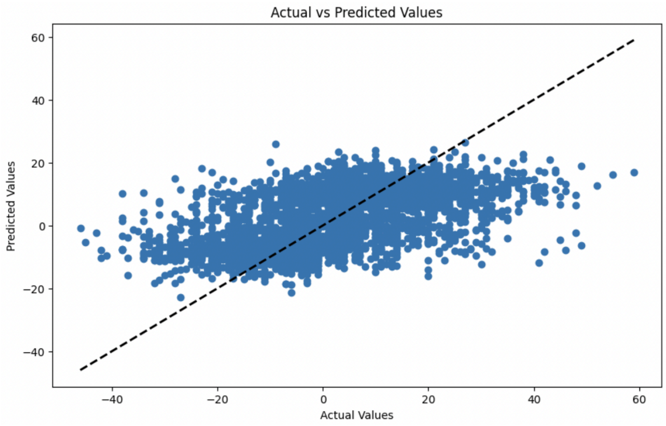
Additionally, we analyzed the feature importances derived from the Random Forest model to identify the most significant predictors. This analysis revealed that 'record_difference_norm' was the most influential factor, consistent with our previous findings.



Feature Importance in Random Forest Model

We found that certain interactions performed better than their individual counterparts, such as ‘temp_elevation_interaction’ (interaction between temp difference and elevation difference) performed better than temperature and elevation difference on their own. Although a low feature importance of these environmental features, they do play a role in game outcome prediction.

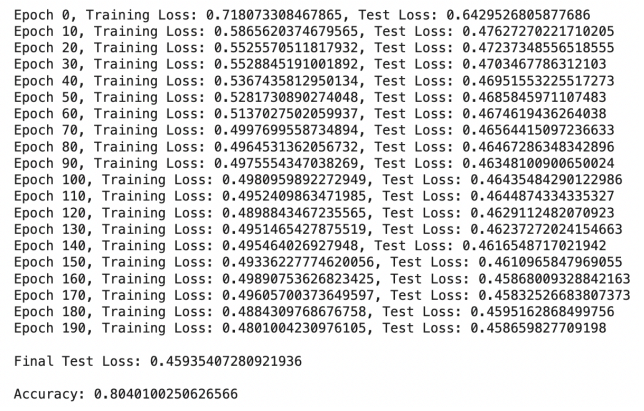
We also observed that the variance of the predicted values was significantly lower than that of the actual values, suggesting that the model might be underfitting the data.



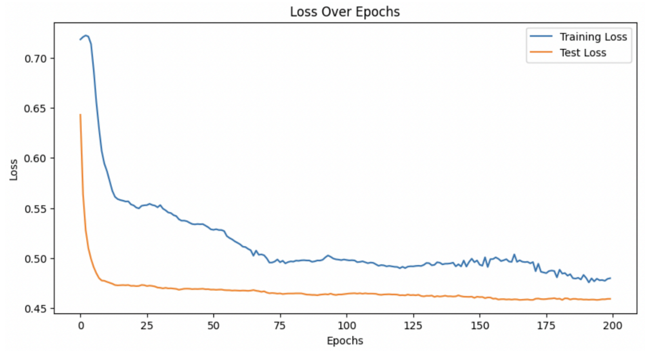
Actual vs Predicted Values in model

2.6 Neural Network implementation

To address the limitations observed in the Random Forest model and to harness the potential of more complex nonlinear relationships in the data, we implemented a neural network model. This model consisted of two hidden layers, utilizing the ReLU (Rectified Linear Unit) activation function known for its efficiency and effectiveness in deep learning models.



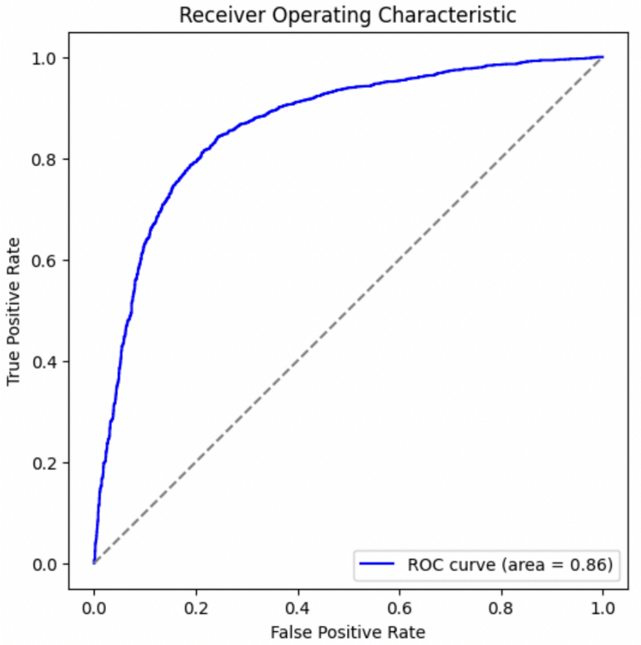
```
Epoch 0, Training Loss: 0.718073308467865, Test Loss: 0.6429526805877686
Epoch 10, Training Loss: 0.5865620374679565, Test Loss: 0.47627270221710205
Epoch 20, Training Loss: 0.5525570511817932, Test Loss: 0.47237348556518555
Epoch 30, Training Loss: 0.5528845191001892, Test Loss: 0.4703467786312103
Epoch 40, Training Loss: 0.5367435812950134, Test Loss: 0.46951553225517273
Epoch 50, Training Loss: 0.5281730890274048, Test Loss: 0.4685845971107483
Epoch 60, Training Loss: 0.5137027502059937, Test Loss: 0.4674619436264038
Epoch 70, Training Loss: 0.4997699558734894, Test Loss: 0.46564415097236633
Epoch 80, Training Loss: 0.4964531362056732, Test Loss: 0.46467286348342896
Epoch 90, Training Loss: 0.4975554347038269, Test Loss: 0.46348100900650024
Epoch 100, Training Loss: 0.4980959892272949, Test Loss: 0.46435484290122986
Epoch 110, Training Loss: 0.4952409863471985, Test Loss: 0.4644874334335327
Epoch 120, Training Loss: 0.4898843467235565, Test Loss: 0.4629112482070923
Epoch 130, Training Loss: 0.4951465427875519, Test Loss: 0.46237272024154663
Epoch 140, Training Loss: 0.495464026927948, Test Loss: 0.4616548717021942
Epoch 150, Training Loss: 0.49336227774620056, Test Loss: 0.4610965847969055
Epoch 160, Training Loss: 0.49890753626823425, Test Loss: 0.45868009328842163
Epoch 170, Training Loss: 0.49605700373649597, Test Loss: 0.45832526683807373
Epoch 180, Training Loss: 0.4884309768676758, Test Loss: 0.4595162868499756
Epoch 190, Training Loss: 0.4801004230976105, Test Loss: 0.458659827709198

Final Test Loss: 0.45935407280921936

Accuracy: 0.8040100250626566
```



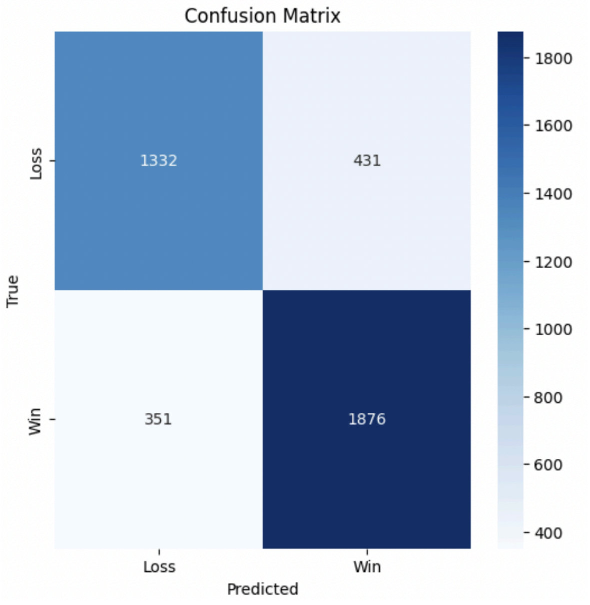
Neural Network Loss Over Epochs

The neural network was trained over a series of epochs, with the number of epochs and batch size carefully chosen to optimize the training process. During the training, we experimented with feeding Principal Component Analysis (PCA) transformed data into the model. However, this approach led to overfitting, indicated by a significant discrepancy between training and validation accuracy. Consequently, we reverted to using the original feature set to maintain the model's generalizability.

Model evaluation was thorough, including the calculation of the final test loss and accuracy. The test loss provided insight into the model's prediction error on unseen data, while the accuracy metric gave a direct measure of the model's performance. Additionally, a confusion matrix and a Receiver Operating Characteristic (ROC) curve were generated. These tools offered a comprehensive view of the model's classification ability, illustrating its true positive and false positive rates and providing an overall AUC (Area Under the Curve) score. The ROC curve, in particular, was instrumental in assessing the model's discriminative capacity, further validating its efficacy in predicting NFL game outcomes.



ROC Curve



Confusion Matrix

3 Discussion

Our comprehensive study, aimed at unraveling the factors contributing to the home field advantage in NFL games, particularly focused on environmental aspects like weather, elevation, stadium capacity difference, and travel distance. The analysis journey, from initial statistical explorations to advanced machine learning models, revealed insightful findings about these environmental influences.

Initially, our statistical analysis indicated a tangible home field advantage, with home teams winning 57.11% of the games. This finding set the premise for further investigation into more specific environmental factors. Our linear regression analysis, which examined the relationship between these factors and the relative score difference, shed light on the influence of variables such as temperature difference and elevation. Interestingly, temperature differences between competing teams' locations emerged as a significant factor, emphasizing the impact of climatic conditions on player performance.

However, when we delved deeper using logistic regression, the predictive accuracy of 80% highlighted that while these factors are influential, they might not be the sole determinants of game outcomes. This was further confirmed by the feature importance analysis, which showed that although geographical and weather-related features were relevant, their impact was relatively moderate compared to other game-specific variables. The Random Forest Regressor, while providing a more nuanced view of feature importances, revealed a crucial limitation in our predictive models. The variance in the predicted values was notably lower than that of the actual values, suggesting a potential underfitting of the model to the complex nature of our dataset. This led us to explore more sophisticated modeling techniques.

In our quest to capture the complexity of the dataset, we implemented a neural network model. The neural network, with its two hidden layers and ReLU activation function, offered a more flexible framework to model the intricate interactions between environmental factors and game outcomes. However, our initial approach of feeding Principal Component Analysis (PCA) transformed data into the model led to overfitting, indicating that the reduced dimensions did not adequately capture the necessary variability for prediction. Thus, we reverted to using the original feature set, achieving an accuracy of 80.4%, comparable to the logistic regression model, but with a more nuanced understanding of the data.

Throughout our analysis, it became evident that while environmental factors like temperature, humidity, and elevation differences do play a role in the home field advantage, their influence is complex and intertwined with various other game-specific factors. Our study underscores the multifaceted nature of home field advantage in NFL games, where geographical and weather factors are just pieces of a larger puzzle. As such, this research contributes to the broader understanding of sports analytics, providing valuable insights for teams, coaches, and sports scientists in strategizing and planning for games.

4 Conclusion

This study embarked on a detailed exploration of the home field advantage in NFL games, with a specific focus on environmental and geographical factors such as weather conditions, elevation, stadium capacity differences, and travel distances. The comprehensive analysis, utilizing various statistical and machine learning methods, aimed to unravel the complex dynamics influencing NFL game outcomes and to quantify the extent to which these external factors impact team performance.

Our findings revealed that while there is a discernible home field advantage, as indicated by the higher win percentages for home teams, the influence of the examined environmental factors is more nuanced. Temperature differences, humidity variations, and elevation disparities did show some impact on game outcomes, yet their roles were not as dominant or straightforward as initially hypothesized. The logistic regression model, with an accuracy rate of approximately 80%, underscored the multifaceted nature of predicting game outcomes, where environmental factors contribute alongside other game-specific elements.

The Random Forest Regression analysis provided further insights into the relative importance of various features, but it also highlighted the challenge of capturing the full spectrum of influences in such a complex dataset. The subsequent implementation of a neural network model addressed some of these challenges, offering a more flexible and nuanced approach to understanding the data. However, the journey through different models, from linear regression to neural networks, emphasized the complexity inherent in sports analytics and the difficulty of isolating the impact of environmental factors from other interplaying variables.

In conclusion, our study contributes to the ongoing discourse in sports analytics by highlighting the intricate and sometimes subtle ways through which environmental factors influence the home field advantage in NFL games. It underscores the need for a holistic approach in sports strategy and planning, where understanding the multifaceted nature of game dynamics can provide a competitive edge. Future research could build upon our findings by integrating more granular data or exploring the psychological and physiological impacts of these environmental factors on players, further enriching the understanding of home field advantage in sports.

REFERENCES

[1] [Jason Zivkovic, Spreadspoke(tobycrabtree),Ty Walters]. ([2017]). [NFL scores and betting data], [Version 42t]. Retrieved [10/22] from [https://www.kaggle.com/datasets/tobycrabtree/nfl-scores-and-betting-data/].

[2] Open-Meteo. (2023). Open-Meteo Weather API. Retrieved March 15, 2023, from https://open-meteo.com/.

[3] Google Cloud API [2023]. Google Cloud Console APIs and Services. Accessed October 2023, from https://console.cloud.google.com/google/maps-apis/

.