Michael Moreno

CS 677 – Spring 2

4/25/24

**Final Project**

**Introduction**

For my final project in CS 677 Data Science with Python, I decided to use the knowledge I gained in class to perform a data analysis on soccer data. I am a big fan of soccer, so I thought this would be a great way to show off my skills in data science. I will be using Python, NumPy, Pandas, SciKit-Learn, Seaborn, and Matplotlib in this project.

The objective of this project is to determine whether “home-field advantage” played a part in teams earning a positive result in the 21st century. In other words, I want to see how teams tend to perform at home to see if they tend to get a better result at home. This project defines a positive result as a win or a draw and a negative result as a loss. The dataset I used is from Kaggle and it is called “International Football Results from 1872 to 2024”, created by Mart Jürisoo. The initial dataset contains three .csv files, but for this project I decided to use the “results.csv” file because it contains the results of matches and relevant information about the home and away team for each match.

To achieve my objective, I will use these three machine learning models I learned in class: logistic regression, decision trees, and random forests. Before applying the models, I created a Pandas data frame containing information from the “results.csv” file. I also added a new column to this data frame that stores the class labels the models will be using to train and predict values. I defined a positive result(win/draw) as class label = 1 and a negative result(loss) as class label = 0.

**Data Specifications**

* The initial “results.csv” file contains results of matched played over 150 years, so I decided to trim the data and conduct my analysis on matched that were played within the 21st century.
* At the time I downloaded the data, the last matches that were included in the dataset were played on March 26, 2024. This will be the last day matches are included in the dataset.
* The results are from matches played by Men’s International Football Teams that represent their country.
* I also decided to exclude matches that were played at a “neutral” venue. A neutral venue is a match played outside the nations of both teams playing. I did this to make sure the models train the data only where there was an explicit home and away side.
* After trimming the data, the number of rows is: 16,461 rows.

**Logistic Regression**

Confusion Matrix:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | Accuracy | TPR | TNR |
| 2133 | 0 | 6098 | 0 | 100% | 100% | 100% |

Due to the 100% accuracy, high TPR, and high TNR rates, the logistic regression model did an excellent job at predicting the values from the dataset. This shows the logistic regression model suggests that “home-field advantage” could play a big role in a team’s likelihood of getting a positive result.

**Decision Trees (max\_depth = 6)**

Confusion Matrix:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | Accuracy | TPR | TNR |
| 2133 | 0 | 6098 | 0 | 100% | 100% | 100% |

Before I tested my data, I used the decision trees classifier to plot the best values of max\_depth and accuracy to determine the optimal value of the hyperparameter max\_depth. The optimal value was found to be max\_depth = 6. I then trained my data using the decision tree model with a max\_depth hyperparameter set to 6. The results are the same as the logistic regression model, with 100% accuracy, TPR, and TNR. Again, this suggests that “home-field advantage” could play a significant role in a team earning a positive result.

**Random Forests (N = 3, d = 5)**

Confusion Matrix:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | Accuracy | TPR | TNR |
| 507 | 12 | 6086 | 1626 | 80% | 23.77% | 99.80% |

Like the decision trees model, I used the random forests classifier to find the optimal values of the hyperparameters N (number of subtrees) and d (max\_depth). The optimal values were N = 3 and d = 5 with an error rate of 20%. Looking at the accuracy, TPR, and TNR rates, we can see that this classifier did not do as well as the others. The 80% accuracy rate is well below the 100% trend we saw. The low TPR rate suggests that the model struggled predicting positive results, but the high TNR rate suggests the model did well predicting negative results.

**Conclusion**

Looking at all the models I used, we can see that the logistic regression and decision trees were the best at predicting the values from the test data. They both had a 100% accuracy rate, while the random forests model accuracy rate was 80%. Personally, I would choose the logistic regression model because it is the best at providing a binary classification (either a positive result or a negative result). Seeing that two out of the three models I used had a perfect accuracy score, I can conclude that, based on data from the 21st century, “home-field advantage” does play a significant role in whether an international soccer team gets a win/draw.

.