Project Report

Project: Predicting the Winning Football Team

Dhiraj H. Gawhare
Roll Number: 22M0062
M. Tech (Aerodynamics)

IIT Bombay

Project Title: Predicting the Winning Football Team

Overview:

The objective of this project is to design a predictive model that can accurately predict whether the home team will win a football match. The project utilizes historical football match data and various features related to team performance, goals scored and conceded, team form, and other match statistics to build and evaluate predictive models.

Dataset:

The project uses a dataset containing historical football match data from various seasons of the English Premier League. The dataset includes information such as match date, home team, away team, full-time home and away team goals, full-time result, half-time goals, match statistics (shots, corners, fouls, etc.), and more. The dataset is organized by season, and data is collected from the 2000-01 season to the 2017-18 season.

Data Preprocessing and Feature Engineering:

1. Goals Scored and Conceded: The project calculates aggregated goals scored and goals conceded for each team based on match data. These are arranged by teams and matchweek.
2. Points Calculation: Points are calculated for each team based on match results (win, draw, or loss) and accumulated over matchweeks.
3. Team Form: The form of each team is calculated based on the last few matches, and this information is added to the dataset.
4. Additional Features: Goal difference, point difference, win/loss streaks, and scaled features are calculated to enrich the dataset.

Exploratory Data Analysis:

The project uses visualizations to explore the dataset's distribution and correlations among features. Scatter plots and heatmaps are used to visualize relationships between features and check for correlations among them.

Data Splitting:

The dataset is split into features (X) and the target variable (y), which is the full-time result (FTR) indicating whether the home team wins, loses, or the match is a draw. The data is further split into training and testing sets using a 70-30 split.

Model Building and Evaluation:

1. Logistic Regression: A logistic regression model is trained using the training set and evaluated using classification metrics such as confusion matrix and classification report.
2. Support Vector Machine (SVM): A support vector machine model with a radial basis function (RBF) kernel is trained and evaluated.
3. Random Forest: A random forest classifier is trained and evaluated.
4. XGBoost: An XGBoost classifier is trained and evaluated, yielding the highest F1 score and accuracy on the test set.

Model Tuning:

Parameter tuning is performed using GridSearchCV for the XGBoost model to find the optimal combination of hyperparameters that maximizes the F1 score.

Conclusion:

The project successfully builds predictive models to determine the outcome of football matches based on historical data and match statistics. The XGBoost model proves to be the most effective in predicting match outcomes, achieving the highest F1 score and accuracy on the test set. The accuracy of the predictive models demonstrates the potential of utilizing machine learning to predict football match results, which has implications for sports betting and match analysis.

Future Directions:

1. Feature Engineering: Explore more advanced features related to player performance, team dynamics, and match context.
2. Ensembling: Combine predictions from multiple models to improve overall accuracy.
3. Real-time Prediction: Develop a web application to predict match outcomes in real-time during ongoing matches.
4. Model Interpretability: Use techniques to interpret and visualize the model's decision-making process.
5. Incorporate External Data: Include external factors such as player injuries, weather conditions, and stadium locations to enhance predictive capabilities