Football Match Predictions

Aksel Kaasik, Tomi Lahe

Link to the repository: https://github.com/AkselK1/FootballPrediction/tree/main

Task 2

Background - Football is one of the most popular sports globally, and predicting match outcomes has long been a topic of interest for fans, analysts, and stakeholders in the industry. Accurate match predictions can provide valuable insights for various purposes, such as enhancing fan engagement, improving team strategies, or even informing betting markets. Modern data science techniques, combined with the availability of comprehensive football data, allow for the development of predictive models that can provide deeper insights into the factors influencing match outcomes. This project aims to leverage these advancements to develop predictive models for football matches.

Business goals - The primary goal of this project is to predict key aspects of European football matches, including the match result (win, draw, or loss) and the number of goals scored. The project aims to uncover factors that influence these outcomes, such as team rating, location of the tournament (home or away), historical data, player statistics (maybe), team that started the penalty shootout etc. By achieving these goals, the project seeks to provide actionable insights for football clubs, fans, and stakeholders who rely on data-driven decisions.

Business success criteria - We want to achieve a prediction accuracy of atleast 65% for the match result prediction and atleast 50% accuracy for the number of goals scored.

Inventory of resources - The project utilizes three datasets that cover various aspects of football matches. Results.csv provides detailed information on match scores, venues, and dates, enabling us to identify and filter European matches. Shootouts.csv offers insights into penalty shootouts, a frequent occurrence in European tournaments. Lastly, Goalscorers.csv allows for the analysis of individual and team scoring patterns, including penalties and own goals. The team is equipped with tools like Python and libraries such as Pandas for data handling, Scikit-learn for machine learning, and Matplotlib for visualization. The focus on European matches ensures the analysis remains concentrated and relevant, with the expertise and resources available to handle this narrowed scope effectively.

Requirements, assumptions, constraints - The primary requirement for this project is to preprocess the datasets to filter out European matches, ensuring that the analysis is region-specific. Predictive models will then be developed to estimate outcomes such as match results and goals scored, with evaluations based on relevant performance metrics. The project assumes that the datasets include sufficient European matches to support robust analysis and that historical performance is a significant indicator of future outcomes. Constraints include the limited timeframe of the project and the computational challenges of modeling complex scenarios, such as penalty shootouts or high-scoring matches.

Risks and contingencies - We have a risk of overfitting due to the smaller data size after filtering out European matches. We will try to mitigate this risk with cross-validation and regularization.

Terminology - To ensure clarity, key terms are defined here. A neutral venue refers to a match location not designated as the home ground for either team. An own goal is a goal inadvertently scored by a player against their own team. A penalty shootout is a tiebreaker mechanism used in European knockout tournaments to determine the winner when a match ends in a draw.

Costs and benefits - The primary cost of this project lies in the time and effort required for data preprocessing, particularly filtering for European matches, and developing predictive models. Computational expenses for training and testing models on refined datasets are another consideration.

Data-mining goals - The primary data-mining goal for this project is to develop predictive models capable of accurately forecasting key outcomes of European football matches. These outcomes include:

1. Match Result - Predicting whether the match will result in a win, draw, or loss for the home or away team.
2. Goals Scored - Estimating the number of goals scored by each team during the match.

Data-mining success criteria - Success in data mining will be measured by the performance of the predictive models and the insights generated. Specific criteria include:

1. Accuracy: Achieving high accuracy in predicting match results, with metrics such as precision and recall score.
2. Explainability: Identifying and understanding the key factors influencing predictions, ensuring the models provide actionable insights rather than just outputs.
3. Generalization: Ensuring the models perform well on unseen data through proper validation techniques like cross-validation.
4. Usability: Delivering results and visualizations that are interpretable and actionable.

Task 3

1. Gathering Data

For this project, three datasets were provided to predict the outcomes (win, draw, or loss) of European football matches:

1. **Results Dataset**
    - Contains key match information such as the date, home and away teams, goals scored by each team, match venue, tournament type, and whether the match was played on a neutral ground.
2. **Goalscorers Dataset**
    - Provides details on individual goals, including the scorer, time of the goal, and whether it was a penalty or own goal.
3. **Shootouts Dataset**
    - Details outcomes of penalty shootouts in knockout matches, including the winner and the team that took the first shot.

These datasets form the basis for understanding match results and team performance during European matches. There may be a need to create more attributes based on the existing data when we get to Feature Engineering.

1.2. Outline Data Requirements

The primary objective is to predict match outcomes (win, draw, or loss) for European matches. For this, the following data requirements were established:

- **Match Features**:
    - Date of the match
    - Match venue: Home/away/neutral.
    - Tournament type: Friendly, qualifier, or tournament stage. This is important as it affects the performance of both teams.
- **Team Performance (Needs to be added to the dataset)**:
    - Recent form: Goals scored, conceded, and results from the last 5 matches.
    - Opponent strength: Indicator like FIFA ranking
- **Goalscoring Features**:
    - Timing of goals, penalties, and own goals, which may significantly impact match outcomes.

1.3. Verify Data Availability

All required data is partially available within the provided datasets.

- **Results Dataset** contains match outcomes and needed match context, covering 9449 European matches after filtering out non-European matches.
- **Goalscorers Dataset** provides detailed goal events but requires integration with the Results Dataset for additional features.
- **Shootouts Dataset** is limited to a subset of matches and will primarily be used for contextual insights rather than model development.

Some features, such as team rankings, are not provided and may need to be sourced externally to strengthen predictive performance.

1.4. Define Selection Criteria

- **Match Scope**: Filtered matches where both home and away teams are from UEFA member nations and Euroepean countries.
- **Tournament Scope**: Include all match types (e.g., friendlies, qualifiers, tournaments)
- **Outcome Scope**: Classify results into three categories: home win, draw, and away win. Shootouts will not initially be included, as they are only used after regular match outcomes are decided.

2. Describing Data

The three datasets were summarized as follows:

1. **Results Dataset**:
   - Total Rows: 9449 (filtered for European matches).
   - Key Attributes: date, home_team, away_team, home_score, away_score, tournament, city, country, neutral
2. **Goalscorers Dataset**:
   - Total Rows: Match-dependent, variable.
   - Key Attributes: Scorer name, goal timing, penalties, own goals.
3. **Shootouts Dataset**:
   - Total Rows: Limited to knockout matches.
   - Key Attributes: Winning team, first shooter.

The datasets provide good coverage of match outcomes and events for European football. There will be a need to generate more attributes based on the three datasets as discussed earlier.

3. Exploring Data

- **Outcome Distribution**:

    - Home wins are the most frequent outcome, consistent with the "home advantage" in football.
    - Draws are relatively less frequent but significant
- **Tournament Impact**:
    - Friendly matches tend to have more goals than other tournaments.

4. Verifying Data Quality

Ensuring data quality is critical to the project's success. Initial checks revealed the following:

- **Completeness**:
    - The Results Dataset is comprehensive for match outcomes, but alignment across datasets (e.g., matching dates and teams in Goalscorers and Shootouts) must be verified.
- **Accuracy**:
    - No anomalies were noted in any datasets
- **Consistency**:
    - Team names and tournament labels vary (e.g., abbreviations, spellings) and require standardization.
- **Outliers**:
    - Matches with unusually high scores or extreme durations (e.g., extended extra time) should be reviewed.
- **Duplicates**:
    - No duplicate entries were found in the Results Dataset, but alignment with other datasets will require further validation.

Conclusion

The data understanding phase highlights a strong foundation for building a predictive model, with datasets offering rich insights into European football. However, some challenges, such as missing external features (e.g., team rankings) and the need for dataset alignment, need to be addressed in the preprocessing phase.

Next, we proceed to data preparation, including feature engineering, handling missing data, and standardizing datasets for modeling.

Task 4

1. Data Preprocessing and Cleaning (3 hours, the dataset is already quite clean)

- **Tasks**:
  - Format dates and detect/remove outliers.
- **Tools & Methods**:
  - Pandas, NumPy, Jupyter Notebooks.
- **Workload: Tomi, Aksel - 3 hours**

2. Exploratory Data Analysis (EDA) (20 hours)

- **Tasks**:
  - Visualize distributions of match outcomes and goals.
  - Analyze relationships between home/away status, match outcomes, and goal counts.
  - Explore trends over time and correlations between features.
- **Tools & Methods**:
  - Matplotlib, Seaborn, Pandas, Jupyter Notebooks.
- **Workload: Tomi - 10 hours, Aksel - 10 hours**

3. Feature Engineering (15 hours)

- **Tasks**:
  - Convert match outcomes into numerical values.
  - Create features like goal difference, home/away indicator, uefa ranking, form, etc.
- **Tools & Methods**:
  - Pandas, Scikit-learn, Python.
- **Workload: Tomi - 10 hours, Aksel - 5 hours**

4. Model Training (15 hours)

- **Tasks**:
  - Select and train models (Logistic Regression, Random Forest, XGBoost).

    - Use cross-validation to tune hyperparameters.
    - Evaluate performance with metrics like accuracy,recall and precision.
- **Tools & Methods**:
    - Scikit-learn
- **Workload: Tomi - 5 hours, Aksel - 10 hours**

5. Model Evaluation and Validation (7 hours)

- **Tasks**:
    - Generate confusion matrices and calculate performance metrics.
    - Validate the model on unseen data.
- **Tools & Methods**:
    - Scikit-learn, Matplotlib, Seaborn.
- **Workload: Both 3.5 hours**