Faculty of Media Engineering and Technology
German University in Cairo
Dr. Nourhan Ehab

Milestone 1: English Premier League Companion

Introduction

The Fantasy Premier League has become one of the most popular online strategy games, where millions of soccer fans worldwide manage virtual teams based on real-world football performances. Success in this game relies not only on intuition and football knowledge but also on data-driven decision-making, as player performance, team dynamics, and fixture difficulty all play critical roles. For managers (players), predictive insights offer a competitive edge in selecting lineups and making transfers, while for researchers, FPL provides a rich domain to explore sports analytics and machine learning.

This project implements a comprehensive machine learning pipeline to predict player points in upcoming game weeks. We analyze datasets combining historical player statistics and fantasy game records to uncover actionable insights and build predictive models for FPL performance. This analysis can provide fantasy managers with valuable insights, both when building a new squad during a wildcard and when making transfer adjustments throughout the season.

Dataset:

Fantasy Football

The Fantasy Football (*FPL*) dataset provides detailed player-level records from multiple Premier League seasons, aligned with the official Fantasy Premier League scoring system. Each row corresponds to a player's performance in a given **gameweek**, including both their fantasy outcomes and underlying football statistics.

Key identifiers include *player name*, *team, position,* and *gameweek*, which allow analysis across time, clubs, and roles. Outcome variables such as *total points, minutes played, goals scored, assists, clean sheets, saves, yellow/red cards*, and *penalties* capture the core events that contribute to a player's fantasy score. While columns such as *selected_by_percent* and *transfers_in/out* reflect player popularity and manager decision trends; however, they are out of scope of our research question in this project.

We want to do our research based only on the performance of the players, not their popularity across fantasy managers.
Dataset Source: Kaggle – Fantasy Football

Project Objectives

1. **Data Cleaning:**

   Remove unnecessary columns and handle null values/duplicates/inconsistent data, if any.

2. **Data-Engineering Questions**

   In order to capture the performance of the players, we need to add a new column, form, which, for this project, is simplified as the average total points over the past four gameweeks (if available), divided by 10.

   From the given dataset, you should analyze the given CSV file to answer and **visualize** the reasoning for the following data engineering questions:
   a. Across the seasons, which player positions (e.g., goalkeeper, defender, midfielder, etc.) score the largest sum of total points on average?
   b. Using the form feature, how did the performance of the top five players evolve across gameweeks during the 2022–23 FPL season? Are the top players in form the same top players with the highest total points?

3. **Predictive Modeling Task**
   Develop a statistical ML model or shallow Feed-Forward Neural Network (FFNN) to predict a new column called upcoming_total_points using:
   a. **Match-related features:** things that happen in or around a game (e.g., goals, minutes, assists, clean_sheets).
   b. **Player-related features**: things that describe the player's role or qualities (e.g., position, creativity, influence, value)
   The model's output will be the player's points for the following week.
   The intuition is that we want to be able to predict the player's future points from his performance in the current week we are in.

This new column should represent the value of points shifted **one week ahead**. In other words:

- For each game week & player, *upcoming_total_points* should contain the total points from the following week for said player.
- The last row will not have a value for *upcoming_total_points* (since it's the final week, there is no “upcoming” week), so you will need to drop the last week.

This is a Regression Problem. The developed models should be evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2).

4. **Model Explainability**

To gain deeper insights into the model’s behaviour, apply explainable AI techniques (SHAP, LIME) to interpret predictions, ensuring transparency about the most influential features.

Project Deliverables

1. A refined Jupyter Notebook documenting all steps and providing a reproducible workflow; Adding justifications and explanations to the design and results.
2. A cleaned dataset with the *form* column added, and modified columns after the feature engineering step (if any).
3. An analytical text report answering and visualizing the data engineering questions. The report should also include the features used in the predictive model and why each feature was selected.
4. Predictive model (statistical ML or shallow FFNN) for estimating *upcoming_total_points*.
5. XAI outputs (SHAP plots and LIME explanations) illustrate the contribution of different features to predictions.
6. An inference function that takes raw input and returns the model prediction.

Submission and Deadline

Please submit your GitHub repository, fulfilling the specified requirements, using the following form: https://forms.gle/GyFaidvUY2BoKCby6 by **October 22nd at 11:59 pm**

Ensure your repository remains private until the deadline. After the deadline, you will be required to make it public or add the course account (csen903w25-sys) as a collaborator for milestone grading.