Zekelman School of Business & Information Technology

Data Analytics for Business

Fantasy Premier League (FPL) Player Points Predictor

CAPSTONE PROJECT

Group 7

Team Members:

Ammar Ammar (0936710)

Bhumiben Sureshbhai Lad (0851209)

Karan Khadka (0836689)

Shimona Shrestha (0848947)

Yeshovardhan Singh (0848098)

Date: 26 April 2025

Table of Contents

Abstract

Fantasy Premier League (FPL) is a globally distributed fantasy football game with millions of managers selecting real players based on their actual performance. The project goal was to use machine learning methods to predict Expected Points for Next Gameweek (EP Next), with the aspiration of helping users with player transfers, maximizing points, and getting the transfers right.

We used a dataset of 684 players scraped from fplform.com, and following exploratory and statistical analyses (ANOVA, T-tests, Pearson Correlation), feature engineering, and regression modelling (Random Forest, Gradient Boosting, Neural Network), the model with the highest test R^2 score of 0.734 was produced, which was deemed to have relatively good predictive power without suffering from overfitting. The final output was published through a Streamlit web application hosted on Hugging Face Spaces.

---

1. Introduction

The Fantasy Premier League is a fantasy football manager game that requires users to manage a squad of Premier League players to try and earn points based on what the players do in real games. Accurately forecasting future player points is a vital part of being successful in the game, while also guiding users on transfer and captaincy decisions.

In this project, we hoped to produce a machine learning model that can predict a player's Expected Points for the Next Gameweek (EP Next). To create the model, we took a historical performance dataset from the Fantasy Premier League, conducted statistical analysis and explored the variables to find the most impactful influence.

---

2. Problem Statement

Faced with the challenging world of FPL, managers typically select players based on their own biased opinions or overall trends, leading to a lack of overall success. An objective, data-driven predictive model can supplement managers' subjective opinion-based decisions with reasonably accurate predictions of players' performances in future fixtures and identify which players to transfer out to improve net points in FPL.

---

3. Target Audience

- **FPL Managers:** To improve a team's selection process, therefore maximizing the total number of points in each gameweek.

- **Fantasy Football Analysts:** To leverage model insights to provide deeper evaluations of individual player performance.
- **Gaming App Developers:** To design recommendations for distant players, which can then be incorporated into the fantasy platform.
- **Data Science Practitioners:** To assess sports analytical modeling methodologies.

---

4. Data and Methods

a. Model Overview

We framed our project as a regression problem where our target was EP Next (Expected Points for Next Gameweek). Our baseline and advanced models included:

- **Baseline Model:** Linear Regression, Decision Trees
- **Final Models:** Random Forest Regressor, Gradient Boosting Regressor (GBR)

b. Data

- **Source:** Scraped from fplform.com
- **Rows:** 684 players
- **Columns:** 42 original features (later narrowed to 31)
- **Target Variable:** EP Next
- **Initial Preprocessing:**
    - Removed irrelevant features (like redundant IDs)
    - Standardized the numerical features (Z-score normalization)
    - One-hot encoded the categorical features ("Position" and "Team")
    - Handled the missing data (mean/mode impute)
    - Removed "Form" (which exhibited high multicollinearity with EP Next)

c. Feature Selection

A mixed-method approach was taken:

- **Filter Methods**: Correlation Matrix to remove highly correlated features (i.e., "Form").
- **Wrapper Methods:** Recursive Feature Elimination (RFE) using Random Forest.

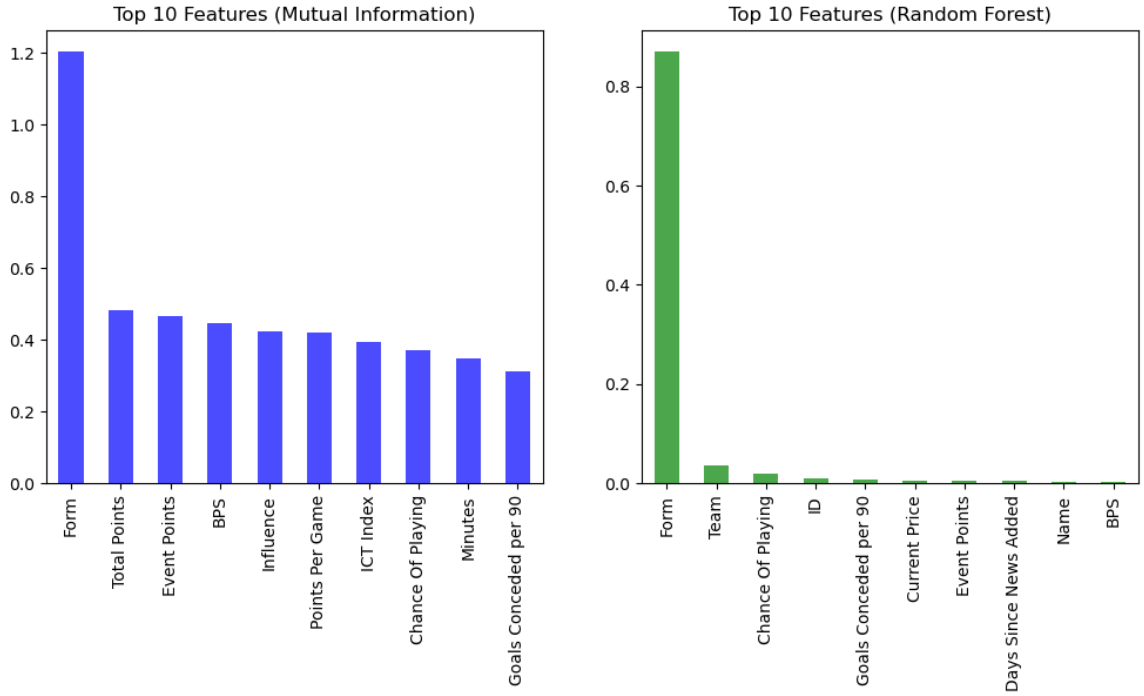
- **Embedded Methods:** Lasso Regression (L1 penalty) and Random Forest Feature Importances.

	with form(R-squared)		without form(R-Squared)	
feature selection	train	test	train	test
all columns	0.9607	0.9648	0.7611	6869
with mututal information	0.8586	0.9492	0.7619	0.7096
with random forest	0.9705	0.9679	0.7714	0.6748
wrapper method	0.9819	0.9694	0.7621	0.734
embedded method	0.9663	0.9608	0.7676	0.6791

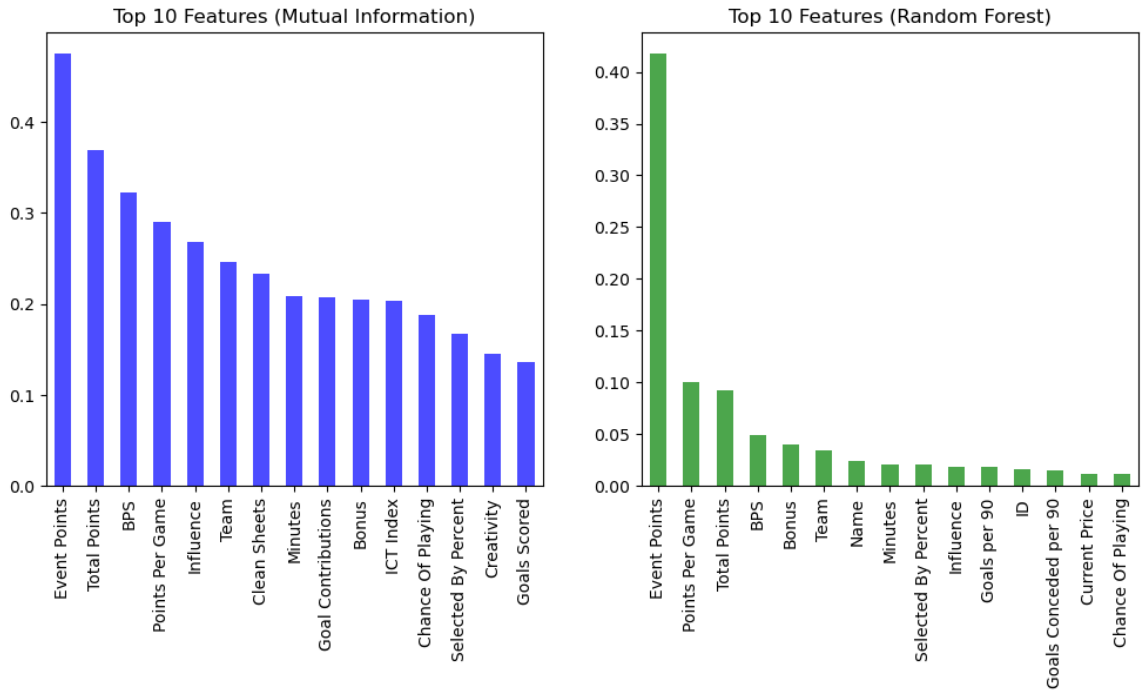
Modeling with and without 'Form':

During our modeling phase, we recognized that 'Form' was going to be hugely correlated with the target variable EP Next. To help support robust and generalizable modeling, we built two versions of our regression models. One version included 'Form' and one version excluded 'Form'.

- With 'Form' in our models the R^2 scores were marginally higher, but with high risk of overfitting from multicollinearity.



- Without 'Form' the R^2 scores were slightly reduced, but model interpretability was improved and the risk of overfitting was reduced.



The final deployed model was built without 'Form', finding the best balance of predictive accuracy and better generalization to unknown data.

Final Selected Features:

- Influence, Points Per Game, Bonus, Minutes Played, Assists, Selected By Percent, Chance of Playing, Current Price, Clean Sheets, Goals Conceded, BPS, Event Points, Threat per 90, 2 Goals Conceded Adder, Days Since News Added.

d. Architecture

- **Random Forest Regressor:**
  Captures feature interactions while quickly avoiding overfitting.
- **Gradient Boosting Regressor (GBR):**
  An ensemble model with good performance in structured data.
- **Streamlit App:**
  Designed simple interactive interface to deploy model.

e. Training, Validation, and Tuning

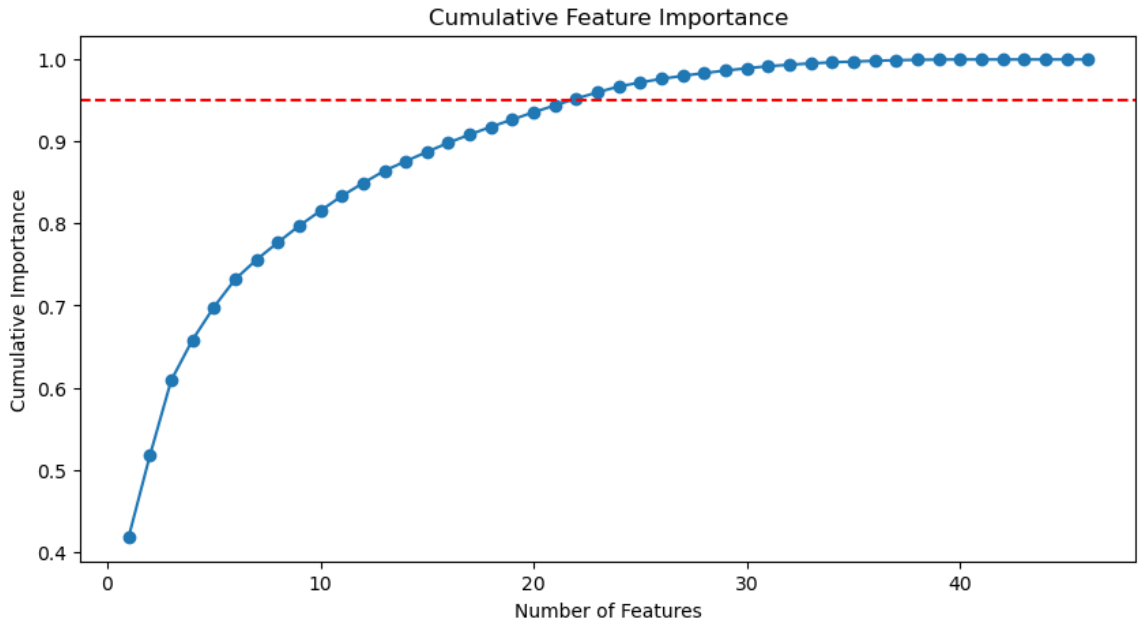
- **Train-Test Split:** 80/20 split
- **Cross-Validation:** 10-fold CV during hyperparameter tuning
- **Hyperparameter Tuning:** Used PyCaret's tune_model() functionality

- **Best Performance:**
    - Random Forest Test $R^2 = 0.734$
    - GBR Train $R^2 = 0.96$

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	0.4616	0.7511	0.8486	0.7621	0.2626	0.6279	0.0750
et	Extra Trees Regressor	0.4584	0.7516	0.8515	0.7579	0.2521	0.5974	0.0750
gbr	Gradient Boosting Regressor	0.4866	0.7603	0.8591	0.7540	0.2612	0.5940	0.0330
lightgbm	Light Gradient Boosting Machine	0.5074	0.7684	0.8653	0.7527	0.2893	0.6590	0.0960
huber	Huber Regressor	0.5291	0.8049	0.8871	0.7344	0.3100	0.5651	0.0190
br	Bayesian Ridge	0.5733	0.8263	0.9008	0.7245	0.3294	0.6348	0.0190
ridge	Ridge Regression	0.5705	0.8286	0.9020	0.7226	0.3291	0.6331	0.0200
lar	Least Angle Regression	0.5703	0.8290	0.9022	0.7224	0.3291	0.6330	0.0200
lr	Linear Regression	0.5726	0.8343	0.9053	0.7215	0.3313	0.6389	0.0200
ada	AdaBoost Regressor	0.7022	0.9428	0.9609	0.6963	0.3765	0.8591	0.0530
omp	Orthogonal Matching Pursuit	0.5837	0.9588	0.9693	0.6864	0.3330	0.8294	0.0200
dt	Decision Tree Regressor	0.5633	1.2305	1.0879	0.6091	0.3204	0.7670	0.0200
knn	K Neighbors Regressor	0.6744	1.3856	1.1645	0.5568	0.3739	0.6323	0.0240
par	Passive Aggressive Regressor	0.8013	1.5218	1.2045	0.5219	0.4390	0.8395	0.0190
en	Elastic Net	0.8627	1.5236	1.2193	0.5202	0.4159	0.8238	0.0200
lasso	Lasso Regression	1.0726	2.2478	1.4817	0.2949	0.5165	0.9307	0.0190
llar	Lasso Least Angle Regression	1.0726	2.2478	1.4818	0.2949	0.5165	0.9307	0.0180
dummy	Dummy Regressor	1.3213	3.1935	1.7727	-0.0154	0.6445	1.0778	0.0180

f. Explainability

- **Feature Importance Analysis:** Random Forest and PyCaret plots.
- **Residual Analysis:** Checked error distributions.
- **Cumulative Feature Importance:** Established the minimum features needed to acheive max performance.



g. Deployment

- **Platform:** Hugging Face Spaces
- **App Features:**

  The FPL Player Points Predictor application is a Streamlit application provided on Hugging Face Spaces. It should be simple, quick, and useful to Fantasy Premier League players.

  Major features:

  - **Pick Players by Budget and Position:** Users filter players by how much money they have and what position they need (Goalkeeper-, Defender-, Midfielder-, or Forward- level).
  - **Best Player Suggestions:** After applying filters, the application shows the best players with highest predicted points for the next gameweek.
  - **Follow the FPL Rules:** The application directly enforces FPL rules like limiting a player to a maximum of 3 players from any one real team (just like in the actual game).
  - **Easy Warnings:** Should they disobey FPL rules (choose 4 players from a single real team for example), the application identifies warnings that provide indications of wrong choices.
- **Interface:** Streamlit app linked to backend prediction engine.

- **Public Link:** Visit our live app
- **Github repository:** https://github.com/bhumilad/FPL

---

5. Results

Model	Train R^2	Test R^2	MAE	RMSE
Random Forest	0.7621	0.7340	0.2528	0.4215
Gradient Boosting (GBR)	0.9607	0.7340	0.2051	0.3946

- **Random Forest** stood out as the superior performance model in terms of generalization and minimal overfitting.
- Predictions were close to actual player scores.
- App recommended top-performing players for FPL users.

6. Discussion

- Basic models (e.g., Linear Regression) faced challenges due to multicollinearity and nonlinearity.
- Ensemble methods such as Random Forest and GBR greatly improved predictability.
- The potential of models to predict players points performance could be improved further by using dynamic live-data feeds (e.g., injuries, team news).
- Real-world delivery as a web app makes reality highly feasible.

---

7. Conclusion

- We have successfully constructed, assessed, and deployed a Fantasy Premier League Player Points Predictor based on statistical methods and machine learning. Our solution showcases strong prediction accuracy and real-world applicability.
- Directions for Future Work:
    - Automate weekly data updates through APIs.
    - Build Captain and Transfer Recommendation modules.
    - Bundle the model within mobile applications.

---

8. Contributions

Member	**Key Contributions**
Shimona Shrestha	• Led outlier detection, trend and correlation analysis efforts • Developed and implemented Pearson/Spearman tests and T tests on player form • Performed VIF and filter-based feature removal • Established PyCaret pipeline (preprocessing, encoding)

Member	Key Contributions
	• Co-developed Streamlit UI and rule validation • Documented workflows & presentations
Bhumiben (Bhumi)	• Conducted trend identification and mutual information analysis • Ran ANOVA tests on categorical variables • Applied Lasso embedded feature selection • Used PyCaret's compare_models() to shortlist top regressors • Developed Streamlit elements for data filtering and user interactions
Karan Khadka	• Wrote the initial data scraping scripts • Performed chi square and other ANOVA tests • Created and tested baseline models (Linear, k NN, Decision Tree) • Plotted feature importance • Contributed to live app functional requirements and formation constraints definition
Ammar Ammar	• Wrangled data and imputed missing values • Conducted Recursive Feature Elimination (RFE) • Built Random Forest and neural network regression models • Hyperparameter tuning and 10-fold CV for GBR • Proposed team selection constraints logic for the app
Yeshovardhan (Yash)	• Engineered key features (Goal Contributions, Threat per 90) • Experimented multicollinearity and assisted method feature pruning • Validated final GBR model with sample player predictions

Member	Key Contributions
	• Saved and loaded trained model for deployment • Assisted UI brainstorming and testing

---

9. References

- Fantasy Premier League Data: fplform.com
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer.
- PyCaret Documentation
- Scikit-learn Documentation
- Streamlit and Hugging Face Spaces Documentation

---

10. Appendices

- **Appendix A: Feature Dictionary**

Feature Name	What it Means
Influence	How much impact a player has on the game (based on actions).
Points Per Game	Average points scored by the player in each game.
Bonus	Extra points given based on overall match performance.
Minutes Played	Total minutes the player was on the field.
Assists	Number of times the player helped another player score.
Selected By Percent	How many managers (percent) have picked this player.

Feature Name	What it Means
Chance of Playing	Likelihood (in %) that the player will play the next match.
Current Price	The current value of the player in the FPL game (in millions).
Clean Sheets	Number of games where the player's team didn't concede goals.
Goals Conceded	Number of goals the player's team allowed while they played.
BPS (Bonus Points System)	A score based on player's full match contribution.
Event Points	Points scored by the player in the most recent gameweek.
Threat per 90	How dangerous the player is (e.g., chances to score) per match.
2 Goals Conceded Adder	A penalty for defenders and goalkeepers if they concede 2+ goals.
Days Since News Added	How many days ago the player's last injury/news update was.