English Premier League Predictions

CT-21013: Data Science Laboratory

Full Name	Enrollment No.
Shyam Aradhye	112003158
Snehal Shinde	142103013
Prashant Sonawane	112003140

Project Guide : Dr. Y.V.Haribhakta

Department of Computer Engineering
Coep Technological University Pune
Shivaji Nagar, Pune, India

May 7, 2023

Contents

1 | Introduction

In this project, we will be using data science techniques to analyze historical data from the English Premier League in order to make predictions about upcoming matches. The Premier League is one of the most popular and competitive football leagues in the world, with millions of fans tuning in to watch matches every week.
Our goal in this project is to build a predictive model that can accurately predict the outcome of Premier League matches. We will be using a variety of data sources, including team and player statistics, historical match data, and betting odds, to train our model. We will then test the accuracy of our model by making predictions about upcoming matches and comparing our predictions to the actual outcomes.
This project will require us to use a range of data science tools and techniques, including data cleaning and preprocessing, feature engineering, and machine learning algorithms. By the end of the project, we hope to have developed a powerful predictive model that can help football fans and bettors make more informed decisions about upcoming Premier League matches.

2 | Objectives

The main objective of the Premier League Prediction Project are :

- Predict the match results
- Predict the Table at the end of season

3 | Requirements

The project is build in Python3. There are several Packages used in this project so they are required to run the project. Flask, gunicorn, pandas, joblib, numpy, selenium, IPython, sklearn, webdriver-manager, xgboost, scipy, requests.

4 | Methodology

Below is the step-by-step methodology. And methodology flowchart (Figure 4.1).

- Data Collection: Collect the necessary data sources from reliable and comprehensive sources, including team and player statistics, historical match data, and betting odds.
- Data Cleaning and Preprocessing: Clean and preprocess the collected data by removing duplicates, handling missing values, and formatting the data to ensure that it is consistent and ready for analysis.
- Feature Engineering: Extract and select relevant features from the cleaned and preprocessed data, including team and player statistics, historical match data, and betting odds. Use domain knowledge to create new features that may improve the predictive power of the model.
- Data Exploration and Visualization: Visualize and explore the data to gain insights into the underlying patterns and relationships in the data.
- Model Selection: Choose an appropriate machine learning algorithm for the prediction task. Consider different types of models, including linear regression, logistic regression, decision trees, random forests, and neural networks.
- Model Training: Train the selected model on a training dataset using the features selected during the feature engineering stage.
- Model Evaluation: Evaluate the performance of the trained model on a testing dataset. Use appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score, to assess the model's performance.
- Model Tuning: Fine-tune the model parameters and hyperparameters to improve its performance.

- Prediction: Use the trained and tuned model to make predictions about upcoming Premier League matches.
- Results and Conclusion: Analyze the prediction results and draw conclusions about the effectiveness of the model. Provide recommendations for future improvements to the model and its application.

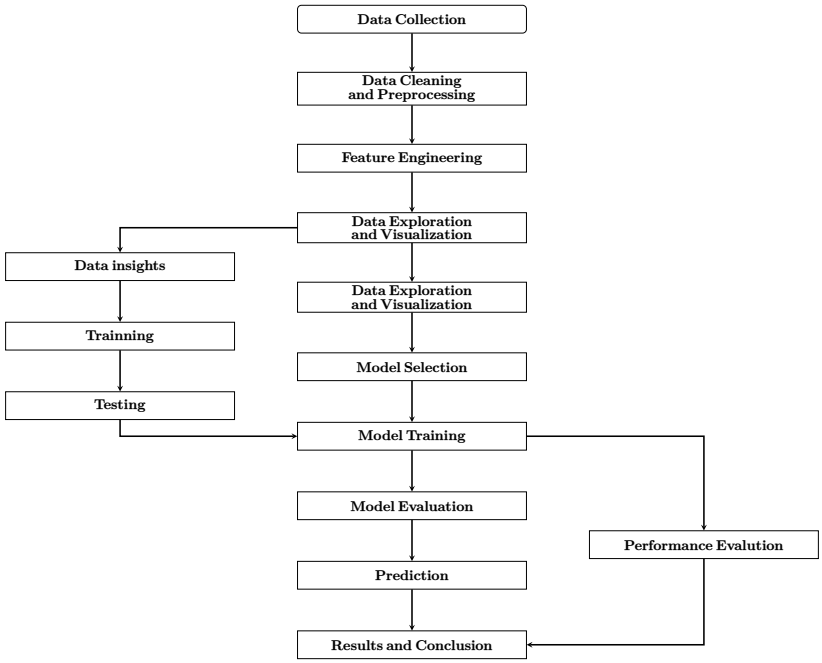
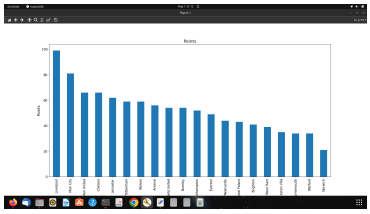


Figure 4.1: Methodology.

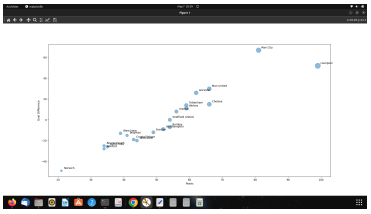
5 | Procedure

- Data Collection: Identify the data sources needed for the project, such as team and player statistics, historical match data, and betting odds. Collect the data from reliable and comprehensive sources, such as the Premier League's official website and sports betting websites. Store the collected data in a structured format, such as a relational database or a spreadsheet.
- Data Cleaning and Preprocessing: Remove duplicates and handle missing values in the collected data. Convert the data to a consistent format to ensure that it is ready for analysis. Normalize or scale the data to ensure that all features are on the same scale.
- Feature Engineering: Extract and select relevant features from the cleaned and preprocessed data. Create new features that may improve the predictive power of the model, such as goal difference, home or away advantage, and form. Use domain knowledge to determine which features are most important for the prediction task.
- Data Exploration and Visualization: Visualize and explore the data to gain insights into the underlying patterns and relationships in the data. Use statistical analysis techniques to uncover correlations and dependencies between features. Use visualization tools such as scatterplots, heatmaps, and histograms to represent the data visually.
- Model Selection: Choose an appropriate machine learning algorithm for the prediction task, such as linear regression, logistic regression, decision trees, random forests, or neural networks. Consider the strengths and weaknesses of each algorithm and select the one that is best suited to the data and the prediction task.
- Model Training: Split the data into training and testing datasets. Train the selected model on the training dataset using the features selected during the feature engineering stage. Use appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score, to assess the performance of the model.

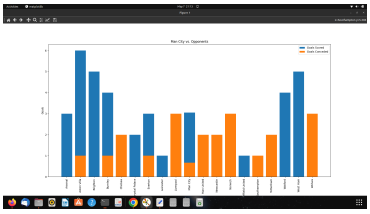
- Model Tuning: Fine-tune the model parameters and hyperparameters to improve its performance. Use cross-validation techniques to avoid overfitting and improve the generalizability of the model.
- Prediction: Use the trained and tuned model to make predictions about upcoming Premier League matches. Compare the predictions to the actual outcomes of the matches.
- Results and Conclusion: Analyze the prediction results and draw conclusions about the effectiveness of the model. Provide recommendations for future improvements to the model and its application.



(a) Bar Chart



(b) Scatter plot



(c) Stacked plot

Figure 5.1: Data Analysis

6 | Results

Prediction Results

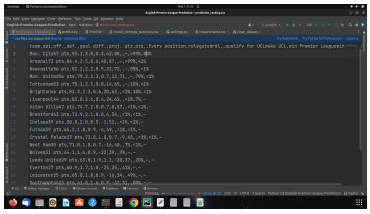


Figure 6.1

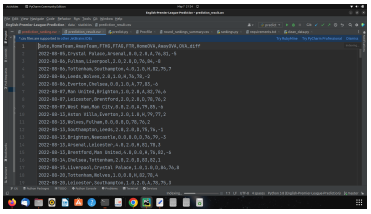


Figure 6.2

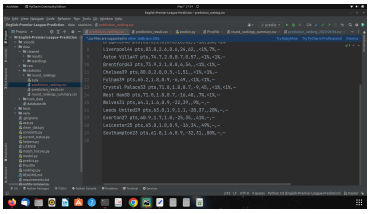


Figure 6.3

7 | Limitations

Here are the Project limitations listed below.

1. Sample Size: The size of the dataset used for the project may be limited, which can impact the accuracy of the predictions. A small sample size may not be representative of the entire population, and this can lead to incorrect predictions.
2. Dependency
   [a] Dependent on Sofifa Website
   [b] Requires ChromeDriver
   [c] Python3 Module dependent
3. Changing Environment: The Premier League is a dynamic and ever-changing environment. Factors such as player injuries, team form, and transfer activities can significantly impact the performance of teams. It may be challenging to keep up with all of these changes and adjust the model accordingly.

8 | Conclusion

In conclusion, the Premier League Prediction Project is an exciting and challenging data science project that can provide valuable insights into the world of football. By following a structured and systematic approach, it is possible to develop a powerful predictive model that can accurately forecast the outcomes of Premier League matches.
However, it is important to keep in mind the limitations of the project, such as data quality, sample size, changing environments, overfitting, unforeseen events, and the lack of causality. By acknowledging these limitations, it is possible to use the model's predictions in conjunction with other sources of information to make informed decisions.
Overall, the Premier League Prediction Project is a great opportunity to apply data science techniques and gain insights into one of the world's most popular football leagues. With the right approach and a good understanding of the data and the domain, it is possible to develop a highly accurate and effective predictive model that can help predict the outcomes of Premier League matches.