Under the guidance of

Prof. Dr. Mahmoud Daneshmand

MIS 637: Data Analytics & Machine Learning

|  |  |
| --- | --- |
| Premier League  Premier League Moneyball  By- Apoorv Akhouri | **Abstract**  Premier League is the most competitive league out there in European football. Each season it is almost unpredictable as to which team will win the league until at least the second half of the season. Due to this, there is very high expenditure done each transfer season by the clubs of Premier league. What we aim to do here is build a platform where clubs can make informed decisions of which player at what position should they bid for.  **Apoorv Akhouri** |

**Problem Statement**

A picture containing company name

Description automatically generated

In this project, we will extract season data from 2015 to 2020 & analyze player stats of the winning clubs by the season, to formulate one Key Performance Indicator (KPI) that would measure the efficiency of these winning teams. This KPI would be a formulated using a set of attributes which we will decide upon in the first half of this project. After formulating this KPI, an analyst can place its own club’s attributes in the formula & see:

1. Where does their club’s score stand on the efficiency scale this season?
2. Looking at the current rumors/facts of the upcoming transfer window, which player would be a better fit to maximize this club’s KPI.

**Sample Data**

Data has been extracted from the following link:

<https://fbref.com/en/comps/9/400/2009-2010-Premier-League-Stats>

Below is a link of the datasets sorted by the season year, that we would be using for the scope of this project:

<https://drive.google.com/drive/folders/1tkoRBISsCfdppR8nMsd3csZ5riAk8MAR?usp=sharing>

In this section, we will see standard player stats & the playing time stats of the 2015 Premier League winning team, Chelsea.

2015 Standard Stats Sample:

Text, application

Description automatically generated

*G+A-PK/90: Goals Plus Assists Minus Penalty Kicks / number of minutes played per 90 mins*

2015 Playing Time Stats:

Table

Description automatically generated with medium confidence

*On-Off: Net goals the team scored per 90 minutes when the player was on the pitch minus Net Goals The team scored per 90 minutes when the player was off the pitch.*

**Relevant Methodologies/algorithm(s):**

To solve the above problem, it would be required to get a good understanding of each column & how it is calculated. We would have to then run EDA, data cleaning & correlation on the datasets to get our relevant attributes. Our y variables would be ‘G+A-PK/90’ for the standard stats data & ‘On-Off’ for the playing time data (description of these 2 attributes are mentioned in the previous page). Attributes relevant to these 2 y variables could be decided upon though decision trees or linear regression. We could also use k-NN classification & get insights on what are our options.

We would split our player data to 3 datasets:

1. fw (Forward)
2. md (Mid)
3. def (Defense)

The fw dataset players will have target variable **‘G+A-PK/90’**.

The md dataset players will have target variable **‘On-Off’**.

The def data players will have target variable **‘0.5\*(G+A-PK/90) + 0.5(On-Off)’**.

We will then use k-NN Classification to train on the data from 2015-2020 & predict on the 2021 data.

Once modelled correctly, the sum of averages of our target variables from fw, md & def from one team, is expected to be the team’s efficiency score.

We will then pick 2 teams at random, find their efficiency scores & then make suggestions on where the team is lacking in order to be Premier League Champions in the future.

**Software Packages to be used:**

For the scope of this project, we would be using Python for data manipulation, making new datasets, analysis & modeling.

We would be using Tableau for visualizations & insights.