PROJECT REPORT

A PREPRINT

Vamsi Bushan
School of Informatics, Computing and Engineering
Indiana University
Bloomington, IN 47408
vbhushan@iu.edu

Asim Azmi
School of Informatics and Computer Engineering
Indiana University
Bloomington, IN 47408
asimazmi@iu.edu

Ayush Bhatia
School of Informatics and Computer Engineering
Indiana University
Bloomington, IN 47408
aybhatia@iu.edu

December 2, 2019

ABSTRACT

In the world of sports, it is always difficult to predict the outcome of a match. This is more evident in an unpredictable sport like soccer where its nearly impossible to predict the outcome of a match even after the available data like team configuration, player attributes and other such features.Being soccer enthusiasts, and upstarts in Machine Learning, we would like leverage the learning methods of Machine Learning for an in-depth analysis of European Soccer data-set.

***K*eywords** Data-set · Research Questions · Evaluation

1 Introduction

The goal of this project is to make predictions and classifications on a European soccer dataset that contains the data of different matches from various european leagues like EPL, Bundesliga, La Liga, etc. based on different attributes of players and teams. We have presented three research questions for our analysis. First question deals with the prediction of the outcome of the match(Win, Draw or Lost) for the home team.We have used classification algorithms like Logistic Regression, Random Forest and Naive Bayes Classifier for this task. Second question deals with predicting a similar player with a given set of player attributes using k-means clustering. Third question is about predicting the overall rating of a given player with a given set of player attributes. This task is accomplished using machine learning algorithms like linear regression, random forest and decision trees

For all the given tasks, we had to complete some data preprocessing and data cleaning in order to make the data appropriate for the machine learning model.Also, the dataset was available in a SQLite format, so we had to implement SQL queries in the pandas library of python in order to convert the tables in the database to the python dataframes.

2 Related Works

There have been some similar works done in the field of soccer match data analysis. Most of the works have used a single machine learning model for the evaluation of their model. Whereas, we have compared the accuracy of the model using different machine learning algorithms and also, we have tried to boost the accuracy of the model combining different set of features using feature engineering.

3 Data-set Description

The data is sourced from kaggle.com- https://www.kaggle.com/hugomathien/soccer. The dataset contains details of 25000+ matches in various european leagues and there are 10000+players with dierent attributes, in seasons ranging from 2008 to 2016, which amounts to a total of 184K instances of player attributes, and 1458 team attributes. The player and $team's$ data has been sourced from EA Sports FIFA games series. The scores, lineup, team formation and events have been sourced from-http://football-data.mx-api.enetscores.com/

There are 7 tables in data-set namely :

- Country - contains only 2 columns - id and name
- League - contains country id and name of all the 11 leagues
- Match - contains a comprehensive data of all the 25k+ matches played in the given 11 leagues form season 2008-2009. This data include number of goals, match id,match date, player coordinates and other such match related features
- Player - contains the biographical details, of all the 11k+ players in the European leagues, like date of birth,height,weight,etc.
- Player Attributes - contains the player attributes for all 11k+ players added every season from 2008-2015. These attributes include stamina, dribbling, passing, shot accuracy, etc.
- Team - contains the details of all 299 teams in all the European leagues combined. This data includes columns like team id, team name, etc.
- Team attributes - contains the team characteristics of the 299 teams, added every season from 2008-2015. The features in this table includes team features related to buildup play, chance creation and defense.

Exploratory Data Analysis(EDA):

The entire European Soccer Dataset is available in a SQLite format. So we had to perform few SQL queries in order to get the tables from the database and then these tables were stored in dataframe variables using the Python's pandas library. Initially, there were many null values in the player coordinates columns in the match table. So all of these rows were dropped from the match table. Also, there was bookie data in the match table, which turned out to be irrelevant for our prediction. So all these columns were dropped from the match table.

We have identified many duplicate rows with same team api id and player api id in the team attributes and player attributes table respectively. To tackle this issue, we sorted both the tables and extracted the latest data that has been added by FIFA recently and also removed the other rows with same api id.

We also checked the player attribute table for the kind of data it has and identified three columns had categorical values. The columns are preferred foot, attacking work rate, defensive work rate. We used Ordinal Encoder class from sklearn for encoding attacking work rate and defensive work rate with priorities equivalent to low, medium and high. And, we used label encoder for preferred foot.

Then we explored the relationship between individual skills and overall rating. We used scatter plots and correlation matrix for representing these relationships. We found two columns namely "potential" and "reaction" are very highly correlated. We believe it will be best to drop these features in order to avoid multi-collinearity issue.

We also joined few features from a different table "player" which had height, width and birthdate. We calculated the age basis on birthdate and tried to explore if we can get any interesting insight on the relation between these new features with over all rating.

4 Research Questions and their Results

There are three research questions that we have focused on, for our analysis of European Soccer Database

4.1 Research Question 1

This research question is focused on predicting the outcome of the match(Win, Lost, Draw) of the home team based on a given set of features.Initially we take all the features including the team attributes to train the model. The team attributes table is joined with the primary match table with a foreign key team api id, in order to combine all the features. We implemented three classification machine learning algorithms for this task - Logistic Regression, Naive Bayes Classifier and Random Forest Classifier. All the three models had the following accuracy :

Logistic Regression : 0.442

Random Forest Classifier : 0.399

Naive Bayes Classifier : 0.36

In order to boost the performance of the model and improve the accuracy, we implemented feature engineering and removed the team attributes table from the match features. After implementing the model with only match features such as player coordinates and overall ratings of each player, we were able to achieve an accuracy of 0.45. The accuracy using different classifier are :

Logistic Regression : 0.433

Random Forest Classifier : 0.45

Naive Bayes Classifier : 0.422

4.2 Research Question 2

Problem statement: To segregate players based on their playing attributes

Description: Based on the data of player attributes, we can segregate each player into clusters. This segregation would help the management of a team to find a player with more value for money during transfer window.

Model Training and practices:

We have used KMeans unsupervised clustering algorithm for segregation. Based on intuition, initially we trained the model with k=5, i.e., the number of clusters to be generated by the algorithm=5.

Observations: The algorithm, supporting our intuition, generated clusters of players based on their playing positions- Striker, Attacking Midfielder, Defensive Midfielder, Defender, Goal Keeper.

Evaluation- To evaluate our learning algorithm's performance we have used the elbow method. We trained the algorithm for different values of k(1,11), and observed that the cost function elbows at k=5.

To further evaluate our model, we have plotted the data points in all the clusters in the following two dimensional planes-

1) ('Reactions','GK Reflexes')

2) ('Interceptions','Standing Tackle')

3) ('Strength','Sliding Tackle')

4) ('Skill','Dribbling')

5) ('Heading Accuracy','Finishing')

There are no two dimensions where could perfectly illustrate the segregation between all the clusters. So, we have chosen the above dimensions to illustrate the segregation between at least two clusters.

Cluster Analysis:

We have analysed the valuations of top-twenty goal keepers against their ratings. We have illustrated this in the code using a scatter plot, where the x-dimension is the ratings, and the y-dimension is the valuations of each player.

Observation: We have observed that there is a huge gap in value between the highest rated player and the second highest rated player in class five. So, if we are looking for a bargain, the second highest rated player provides a better value for money, in comparison to the highest ranked player.

4.3 Research Question 3

Research Question : To predict the overall rating of the player.

Description: Based on the data-set we have we can make a model to predict the overall rating of a player based on various skill sets that the player possess. This model can be helpful to predict how a new player will perform in leagues and can help managers to make a decision about buying a player or not.

Model Training and Prediction:

Splitting the data into 70/30 ratio ,we trained three models. The accuracy and error on the Test set are as mentioned below.

Linear Regression :

Accuracy - 72.10

Error - 3.7

Decision Tree :

Accuracy - 96

Error - 1.35

Random Forest :

Accuracy - 98

Error - 1.35

Observing such high accuracy we realized that there is a high possibility that test data set has duplicate data from train data set. This is because the overall data set has multiple entries for the same player for different seasons. In order to avoid this possibility we dropped duplicates from our data-set and trained the model again. Below are the results

Linear Regression :

Accuracy - 68.10

Error - 3.53

Decision Tree :

Accuracy - 85

Error - 2.43

Random Forest :

Accuracy - 94

Error - 2.39

Since our accuracy on test data set is not very different from the accuracy on the training data-set itself we can say that our model is not over-fitting.

4.4 Technologies used

1) Python

2) SQL

4.5 References

1) https://towardsdatascience.com/

2) https://github.com/

3) https://www.kaggle.com/