

CS322:Big Data

Final Class Project Report

**Project (FPL Analytics / YACS coding): FPL Analytics**  **Date: 01-12-2020**

|  |  |  |  |
| --- | --- | --- | --- |
| SNo | Name | SRN | Class/Section |
| 1 | Archana J | PES1201801306 | I |
| 2 | Swati Naik | PES1201801489 | I |
| 3 | Srujan KN | PES1201801829 | I |
| 4 | Sumukha MK | PES1201801995 | I |

## Introduction

The aim of our project is to analyze various real time events that happen in a football league (English Premier League) and come up with necessary results .These events are associated with many metrics such as players’ free kick effectiveness, pass accuracy , duel effectiveness , etc. We use streaming data for the analysis.

## Related work

Any background study material that you may have read and referenced

We referred to PESU Academy videos and slides to learn about Streaming Spark and MLLib. We also referred to [www.apache.org](about:blank) to understand how to use various modules in Spark, Streaming Spark and MLLib in Spark.

## Design

Talk about the design of the system, algorithms used, and models implemented. Block diagrams are preferred wherever applicable.

Data was streamed through port 6100 and we used Streaming Spark to read the streamed data. We performed various actions and transformations on the DStreams created by Streaming Spark. We calculated all the metrics such as free kick effectiveness, pass accuracy, duel effectiveness, own goal, shots on target , etc for each player from all the matches played by him.

The user can request for 3 queries in the form of json files. The first query is to predict the winning chance of a team given player names of two teams and the date of the match. For predicting the same, we have calculated the average chemistry coefficients of each team and then multiplied it with the ratings of the respective team players. But for this method, there are two issues. One is that a player might not have played enough matches in the past. To solve this issue, we performed KMeans clustering on all players who have played less than 5 matches. For this we converted the input data (playerId, number of fouls, number of goals, number of own goals, pass accuracy, shots on target and the player rating) to Vectors using VectorAssembler module in Spark. Then we used KMeans function to approximate the chemistry coefficient for players who have played less than 5 matches.

The second issue with this approach is that a player’s rating depends on the player’s age. So to solve this problem, we used Linear Regression to find the rating of the player with respect to his present age. Here the input data (PlayerId, Age and Rating) were given as the input to the Regression model to train the model. So while processing the query, we find the present age of each player and find the rating from the model.

The second request to provide the Player’s Profile given the Player’s name. We used the data output (stored in PlayerProfile files) from the model which we trained from the Streamed data.

The third query is to find the Match data given the match’s date and label. This is also answered from the data we processed from the Streamed data which is stored in MatchData.json file.

## Results

Discuss the results you got. What inferences could you draw from the results? Was any result unexpected? Any fine-tuning done to parameters so that the results changed?

By extracting the data from the streamed data, we calculated results for the below metrics:

1. Pass Accuracy
2. Duel Effectiveness
3. Free Kick Effectiveness
4. Shots on Target
5. Foul Loss
6. Own Goal

After computing the results for pass accuracy, duel effectiveness , free kick effectiveness , shots on target, we observed the output values were bounded between 0 and 1 for each of the specified metrics. No results were unexpected, and hence fine tuning as such was not required. This output data was stored in PlayerProfile files using saveAsTextFiles(). And then we computed the chemistry between each player and stored in Chemistry.txt files using the function saveAsTextFiles().

The outputs for each query are stored into json files if the query is successful, else it prints the output in the terminal.

## Problems

Mention problems faced and how were they solved

One of the problems we faced initially were to extract data from the streamed data for further analysis. Since we are not reading from a stored dataset which is simpler, Extracting Data values for metric calculation was a bit of a challenge. Proper usage of mllib for clustering and prediction also involved decent effort.

## Conclusion

What was your main learning from this project?

We got to learn a lot of things from this assignment. Analysis of streaming data is one of the main learning aspects in this project , Practical use of Dstreams and RDDs and application of Mllib in Spark are also few key take-aways.

## EVALUATIONS:

|  |  |  |  |
| --- | --- | --- | --- |
| SNo | Name | SRN | Contribution (Individual) |
| 1 | Archana J | PES1201801306 | 100% |
| 2 | Swati Naik | PES1201801489 | 60% |
| 3 | Srujan KN | PES1201801829 | 30% |
| 4 | Sumukha MK | PES1201801995 | 30% |

## (Leave this for the faculty)

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Evaluator | Comments | Score |
|  |  |  |  |

## CHECKLIST:

|  |  |  |
| --- | --- | --- |
| SNo | Item | Status |
| 1. | Source code documented |  |
| 2. | Source code uploaded to GitHub – (access link for the same, to be added in status 🡪) |  |
| 3. | Instructions for building and running the code. Your code must be usable out of the box. |  |