Technical Project Proposal

Title:	Sports Data Analytics on International Football Results
Author:	Sweta Sindagi
SUID:	L00149560
Supervisor:	Dr Shagufta Henna
Degree:	MSc in Big Data Analytics and Artificial Intelligence

1 Title

Sports Data Analytics on International Football Results

An insightful journey through the simulations to discover the findings for International Football Results using Sports Data Analytics.

2 Problem Description

The International football matches from 1972 to 2019 results are incorporated in this dataset. Discovering chronological sports statistics applied to achieve competitive gain for a team or player which uses player performance data and it is to develop innovative metrics to measure key skills required by a player to succeed in the matches. This helps to identify good players amongst those who are relatively undervalued in the market. The sports analytics is about using data to derive player and team statistics to achieve a competitive outcome which focus on On-field analytics. It uses the data in explanatory predictive manner to improve the performance of the team or an individual player. Historical data from matches are analyzed to quantify skills and weaknesses in terms key metrics. Team players selection and Scouting are performed using performance matrix. And, Off-field analytics can be used for increasing ticket and merchandise sales by analyzing sales to gain customer groups preferences to send promotional offers.

3 Description of Data

The dataset is extracted from the [1] which includes International football matches results starting from 1972 to 2019.

Why is it Challenging

results.csv contains the columns for Matches outcomes.

1. Current name of the team is used for home and away teams columns which is to trace team history statistics.

2. The Country names used at the time of the match which is indicated by the Neutral column.

4 Methodology

Approach:

The preliminary phase to make a pathway using dataset is to clean before process and analyze the data which is performed by Open Refine in the Spark pipeline which is perfomed using OpenRefine Distribution [2]. This comprises of Faceting and Atomization. Clustering technique to group similar data is applied to handle data inconsistencies in the categories in the dataset.

Machine Learning Algorithm:

1. Logistic Regression Classifier performed to predict a dependent variable in a given historical football matches data such that the dependent variable is categorical [3].

2. It is possible to use Support Vector Machine which is a supervised classification method that separates the data using hyperplanes [4].

Programming Paradigm:

Apache Spark SQL data analysis using data from International Football matches results. It is to propose Spark SQL used to visualize the historical Sports data by building the model using the Spark's machine learning API [5]. In addition to that, MLib Classifier methods are used to produce knowledge base about the past trends.

Platform details:

- Standalone Cluster in Ubuntu Machines
- Hadoop 3.2.0
- Tableau
- PySpark SQL 2.4.0

5 Goals

The major aim of this project proposal is to obtain solutions applying Data Analytics to the following queries :

1. Historical statistics applied to gain competitive advantage for a team or player.
2. The outcome used for player selection for a game against a particular opponent.

3. To uncover the best team of all time.

6 Link to the GitHub

The link given below is used for controlling the source code for this project: https://github.com/sindagisweta/SportsAnalytics

References

[1] International football results from 1872 to 2019. [Online]. Available: https://kaggle.com/martj42/international-football-results-from-1872-to-2017

[2] T. F. Kusumasari and Fitria, "Data profiling for data quality improvement with OpenRefine," in *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, pp. 1–6, ISSN: null.

[3] K. Hallmann and C. Breuer, "The impact of image congruence between sport event and destination on behavioural intentions," vol. 65, no. 1, pp. 66–74. [Online]. Available: https://doi.org/10.1108/16605371011040915

[4] K. Zhang, H. Xu, J. Tang, and J. Li, "Keyword extraction using support vector machine," in *Advances in Web-Age Information Management*, ser. Lecture Notes in Computer Science, J. X. Yu, M. Kitsuregawa, and H. V. Leong, Eds. Springer, pp. 85–96.

[5] Data science how-to: Using apache spark for sports analytics. [Online]. Available: https://content.pivotal.io/blog/how-data-science-assists-sports