CRICTO-PAUL

Project Proposal

Based on Project 7

Group Members:

1) Soham Mehta	111496015
2) Sai Prasanth Gumpalli	111480980
3) Venkata Kedarnath Pakala	111014006

1. ABSTRACT

This project aims to whether it is possible to find the outcome of a cricket match and predict a winner based on past data and given features. It aims to predict the result of a future match similar to Octopus Paul but in a much informed and sophisticated manner. A vast amount of ball-by-ball data of every match played since 1800's is available online for analysis. This data can be combined to find the batting, balling, fielding averages of each team in a given year. Also, the ranking of players in each team can help in identifying the team's performance during that match. Other features include knowing whether match is being played at home ground or away, the past performance of a team in that particular ground and the net run rate of a team against the competitor. These all parameters can be combined to identify the current strength of a given team and its odds to win in a given scenario.

2. DATA

2.1 Data Identification

Being one of the most followed sport in world, there is a plethora of data available online on sports website providing the scorecard of each match ever played. However, not all this data can be made to use. A careful selection of features have to be done to make an input vector having relevant information suitable for prediction. The most widely available sources for cricket information are:

1) https://www.icc-cricket.com : Official website of the International Cricket Council where all the records, rules and rankings are available
2) http://www.espncricinfo.com : Each and every cricket match data is available in the form of a scorecard.
3) https://cricsheet.org/downloads : The ball-by-ball data of all the matches in every format (One day, T20, Test) is available to download in YAML format.
4) http://www.howstat.com/cricket/home.asp : Aggregated data of teams and their performances year wise and series wise is available for viewing.
5) http://www.cricbuzz.com/ : This is the most used website in India to watch live updates in cricket match. Due to such a large audience, they have detailed analysis of each ball along with written commentary.

2.2 Data Scraping

A lot of data is available online however there is no single dataset which can be downloaded from a given website as a csv which can be worked upon. Gathering the data is the toughest challenge. The data mentioned in the above sources can be obtained by scrapping them. This might result in consistent and redundant data which will require extensive cleaning. There is also an API available for downloading cricket statistics from ESPN. It can be accessed from the URL in the form:
http://stats.espncricinfo.com/ci/engine/stats/index.html?class=1;team=6;template=results;type=batting

class=1; **team**=6; **template**=results; **type**=batting

If we break down this URL to get more statistics on cricket, we can choose the following parameters.

class
1=Test
2=ODI
3=T20I
11=Test+ODI+T20I

team
1=England
2=Australia
3=South America
4-West Indies
5=New Zealand
6=India
7=Pakistan and
8=Sri Lanka

type
batting
bowling
fielding
allround
fow
official
team
aggregate

2.3 Data Cleaning

The main issue here is not the availability of data but it is about the availability of data in the right format. Each website has a different way of representing the match data. Also, not all data is required by us for example ball-by-ball analysis is not that important as match net run rate and individual team and player statistics. So the data obtained from the scrapped websites has to be cleaned to obtain the most relevant features. This relevant features can then be trained to understand the relationship between the team and opponent, players in each team, team averages, player averages, player rankings, ground conditions and odds of each match.

3. FEATURES

Given 2 teams , we are trying to predict the win percentage of the first team. For this we need to determine various factors that directly impact team's victory. The victory of team depends on the following statistics:

Team Features

1.	Win Percentages	(Number of matches won out of 100 matches)
2.	Batting average	(Runs scored in 100 matches / Wickets lost)
3.	Bowling Average	(Runs scored in 100 matches/ Wickets taken)
4.	Batting Run Rate	(Runs scored in previous 100 matches/ Overs batted)
5.	Bowling Economy Rate	(Runs conceded in previous 100 matches / Overs bowled)
6.	Batting Wicket rate	(Wickets lost in previous 100 matches/ Balls batted)
7.	Bowling Strike rate	(Wickets taken in previous 100games ./ Balls bowled)
8.	Ground type	(home ground / away)
9.	Win percentage at ground	(The times team won at that ground)
10.	Win vs Opponent	(Total wins/loss against the competitor)

Player Features

1. ICC Ranking	The rank of the player given as per ICC
2. Batting Average :	Runs scored /number of times out
3. Batting strike rate :	Runs scored per balls faced
4. Bowling average :	Runs conceded / wicket taken
5. Bowling economy rate	Runs conceded / overs bowled
6. Bowling strike rate	balls bowled / wickets taken
7. Number of records	Total records held by player

4. MODELS

This is a classification problem predicting the outcome of a match. We will try to use the following models to predict the outcome and decide which one works as the best:

Naive Bayes
Logistic Regression
Support Vector Machine
Neural Networks

4.1 Modelling the data

Individual team's data will be given as input to all the classifiers for training purpose like the number of matches won, number of matches lost, ranking of the team over the years, batting averages, bowling averages of the team, net run rate of the team. Classifier will train the model over the win percentage of that team in that year with respect to previous years.

So when the match actually is to happen we will simply provide the individual team's data as features to the classifier to get the win percentage of each team and from that we will predict the outcome of the match.