**Off-Field Sports Analytics**

(Playing XI Selection for Cricket Team (ODI) based on Venue and Opponents)

Team Members:

1.Jaivignesh R

2.Sanman Malwankar

3.Sidesh Sawant

4.Udhav Patil

Abstract

Selecting a balanced playing XI in the game of cricket with the right mix of players of different specialization is a difficult decision making problem for the team management. To make the process more objective, optimization techniques can be applied to the process of selection of players from a given squad. Such an exercise has two dimensions. First, a suitable tool for performance measurement of cricketers needs to be defined (Predictive Tools). Secondly, for selecting a balanced team of XI players, an appropriate objective function and some constraints need to be formulated (using Past Data). Since the captain gets an obvious inclusion in the cricket team, the area specialization of the captain influences the selection of other ten positions in the playing XI. This study attempts to select the optimum balanced playing XI from a squad of players given the specialization of the captain using binary integer programming. To validate the exercise, data from the upcoming seasons for a Cricket.

Table of Contents

* Problem Statement
* Hypothesis Generation
* Data Exploration
* Data Cleaning
* Feature Engineering
* Model Building
* R Code
* Output Tables
* Conclusion
* References

Problem Statement:

The aim is to build a predictive model and find out the best suitable 11 players from Squad of 15 based on the Opponent, Pitch Condition and weather.

Using this model, We will try to understand the properties of Players Features like Batting Average, Bowling Economy, Runs, Matches played, Ground type and Opponent Strength which play a key role in increase the Winning Probabilities..

The Hypotheses

We came up with the following hypothesis while thinking about the problem. These are just our thoughts and come-up with many more of these. Since Our Game is Cricket(Batting and Bowling). lets make different sets for each.

**Batting Level Hypotheses:**

1. Run:<Defintion Required>
2. Mins:<Defintion Required>
3. Batting Average: <Defintion Required>
4. Number of Matches Played: <Defintion Required>
5. Position Played: <Defintion Required>
6. Highest Run Scored: <Defintion Required>
7. Venue Played:
8. Type of Pitch:

**Bowling Level Hypotheses:**

1. Wickets Taken:<Defintion Required>
2. Overs Bowled:<Defintion Required>
3. Bowling Average: <Defintion Required>
4. Number of Matches Played: <Defintion Required>
5. Type of Bowler: <Defintion Required>
6. Type of Pitch: <Defintion Required>
7. Venue Played:

These are just some basic 15 hypothesis we have made, but can think further and create some of our own. the data might not be sufficient to test all of these, but forming these gives us a better understanding of the problem and we can even look for open source information if available.

2. Data Exploration

The first step is to look at the data and try to identify the information which we hypothesized vs the available data.

**Sample Batting Data extracted from ESPN CRICINFO Website:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S.No | Runs | Mins | BF | 4s | 6s | SR | Pos | Dismissal | Inns | Opposition | Ground | Start Date |
| 1 | 26\* | 30 | 21 | 3 | 1 | 123.8 | 5 | not out | 2 | Zimbabwe | Harare | 12-Jun-10 |
| 2 | DNB | - | - | - | - | - | - | - | 2 | Zimbabwe | Harare | 13-Jun-10 |
| 3 | 28 | 28 | 19 | 5 | 0 | 147.36 | 1 | bowled | 1 | South Africa | Durban | 09-Jan-11 |
| 4 | 14 | 15 | 12 | 2 | 0 | 116.66 | 3 | caught | 1 | West Indies | Port of Spain | 04-Jun-11 |
| 5 | 4 | 5 | 5 | 0 | 0 | 80 | 4 | caught | 1 | England | Manchester | 31-Aug-11 |
| 6 | 15 | 21 | 16 | 3 | 0 | 93.75 | 3 | caught | 1 | England | Kolkata | 29-Oct-11 |
| 7 | 22 | 30 | 21 | 0 | 1 | 104.76 | 3 | caught | 2 | Australia | Sydney | 01-Feb-12 |

**Sample Bowling Data extracted from ESPN CRICINFO Website:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S.No | Overs | Mdns | Runs | Wkts | Econ | Pos | Inns | Opposition | Ground | Start  Date |
| 1 | 4 | 0 | 22 | 1 | 5.5 | 4 | 1 | v Zimbabwe | Harare | 12-Jun-10 |
| 2 | 4 | 0 | 48 | 1 | 12 | 5 | 1 | v Zimbabwe | Harare | 13-Jun-10 |
| 3 | 4 | 0 | 33 | 1 | 8.25 | 4 | 2 | v South Africa | Durban | 09-Jan-11 |
| 4 | 4 | 0 | 30 | 1 | 7.5 | 3 | 2 | v West Indies | Port of Spain | 04-Jun-11 |
| 5 | 4 | 0 | 37 | 0 | 9.25 | 4 | 2 | v England | Manchester | 31-Aug-11 |
| 6 | 4 | 0 | 20 | 0 | 5 | 1 | 2 | v England | Kolkata | 29-Oct-11 |
| 7 | 4 | 0 | 34 | 1 | 8.5 | 1 | 1 | v Australia | Sydney | 01-Feb-12 |

3. Data Cleaning

This step typically involves imputing missing values and treating outliers. Though outlier removal is very important in regression techniques, advanced tree based algorithms are impervious to outliers.

Imputing Missing Values

We found two variables with missing values –Runs and Postions. Lets impute the former by the average weight of the particular item. This can be done as: