Intermediate report: Predicting the outcome of ODI Cricket matches

Samiran Roy 153050091
Sushant Shambharkar 153050081
Ganesh Bhambarkar 153050072
Surendra Singh Rao 153050069

February 2016

1 Description

Our task is to investigate to what degree we can predict the outcome of cricket matches, specifically ODI Matches. Given the popularity of this multi-million dollar industry, there is a strong incentive for match fixing and underground betting. The complex rules surrounding the game, along with the numerous parameters affecting the game, including but not limited to cricketing skills and performances, match venues, toss, weather conditions present significant challenges for accurate prediction. This problem has been well investigated for games like basketball and soccer, but yet to be researched for cricket.

2 Dataset

2.1 Match data

We will be using the dataset from `cricsheet.org`

The data-set provides ball by ball data for matches. The results of the games can be a *win*, *tie* or *no result.* Each YAML file contains data about one game. In case of *no result*, the reason is not provided. Although the *city* where the games took place is present in most files, about 10% of files do not specify the city. We'll need to manually fill the city by looking at the venue in that case. Some matches span for 2 days. We'll consider only the starting date as our feature for prediction. The data of total 1,164 ODI matches and total 500 T20 matches are present.

2.2 Weather data

`https://www.wunderground.com/history/index.html` provides daily weather data. We can query the data for any city on a particular day using a simple

URL change. Then we can parse the information from the website. This allows us to automate the gathering of the data per city, per day and put it as features in our data.

For example, this link gives the weather data for Mumbai on 21 March 2006: `https://www.wunderground.com/history/airport/VABB/2006/3/21/DailyHistory.html?req_city=Mumbai&req_state=&req_statename=India&reqdb.zip=00000&reqdb.magic=1&reqdb.wmo=43003`

The website provides following information which is relevant to our project:

- Temperature
- Moisture
- Precipitation
- Sea Level Pressure
- Wind

3 Current progress

We have written code to parse YAML files and produce one CSV file which contains following columns:

- Year : The year of the game
- Month : The month of the game
- Day : The day of the game
- City : The city where the game took place
- Venue : The venue where the game took place
- FirstTeam : The name of the first team.
- SecondTeam : The name of the second team.
- FirstToBat : The team which bats first. 0 : first team, 1: second team
- Result : The result of the game. 0: win, 1: tie, 2: no result
- Winner : The winner team if any.

Other than this, we have designed the flow of the code and main functions for which code needs to be written.

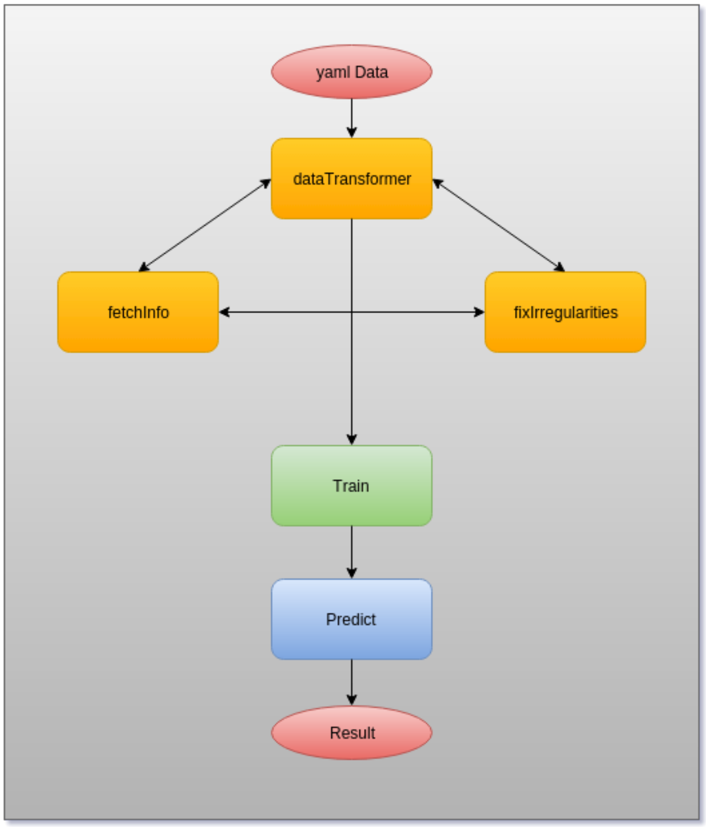

Figure 1: Flow diagram for the project

Plan for the rest of the semester

Feature selection using Pearson correlation, recursive feature elimination and random forests.
We need to figure out a way to model team/player weaknesses from this data
Building a machine learning model prediction. The algorithms we will try are:

- Random Forests
- SVM
- Neural Networks
- Bayesian Methods

We shall derive inspiration from the following papers[?][?][?][?]: