**Prediction of rising stars in sports of cricket**

Classification models are divided into two classes.

* Generative (that randomly generate the observable data values while given some latent variables)
* Discriminative (that model the dependence of unknown variable over known variable).

**DATASET ACQUISITION:** The data is taken from espncricinfo that is a reliable web forum containing data of all cricket matches ever played since 1779. We make RSP both for batsmen as well as for bowlers based on real world ODIs dataset ranging from 2006 to 2013. Moreover, the predictions are made for the players who started their ODI international career during the span 2006-2013, were having maximum age of 30 years until 2013, and are still playing in their respective international ODI teams. More precisely, the data for the span 2006 to 2013 is used for RSP of 2013 and onwards. But, the first four ODI years data of each RS candidate is incorporated for RSP. The reason for taking of such a long span is to incorporate the data of Co-players (specially to check the effects of domain stars on RSs performance). During underlying era, a total of 645 batsmen 560 bowlers have performed in 1138 ODI matches. However, a pre-processing step was performed to huge amount of data for extracting more relevant information for RSP. In details, the players who played less than 20 innings were eliminated from the dataset because they did not play matches with all top ranked teams. After pre-processing, the batsmen were ranked in descending order w.r.t aggregate runs of individuals and bowlers are graded w.r.t. total wickets taken by individuals. Subsequently, top 200 instances for each domain (i.e., batsmen and bowlers) are picked, and their corresponding feature scores are extracted. Finally, two datasets for each domain are generated based on the following two metrics for RSP.

1. **WEIGHTED AVERAGE OF BATSMAN (WA(B)):** Runs (R), average (Avg) and strike rate (SR) are the three salient features that are considered to define the WA(B). Since all features are positively correlated with the batsman performance, hence, all of them are added. Besides, an equal threshold weightage of 33.33 is given to each factor as follows:
2. **WEIGHTED AVERAGE OF BOWLER (WA(BOW):** Wickets (W), average (Avg), economy (Eco) and strike rate (SR) are the four salient features that are considered to define the WA (Bow). Since, number of wickets are positively correlated with the performance of bowlers, hence, this feature is added, while the remaining three features are subtracted due to the negative correlation with the bowler’s performance. Besides, an equal threshold weight of 25 is given to each factor as follows:



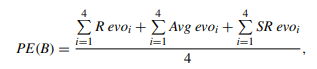
The second dataset is extracted for year wise performance check based on different characteristic nominated as performance evolution of a batsman PE(B), which incorporates runs, average and strike rate of a batsman. In similar manner, performance evolution of a bowler PE (Bow) considers wickets, economy, average and strike rate of a bowler.

First, we define the metric evolution that measures the ratio of change in the evolution indices (runs, average and strike rate of a batsman, while wickets, economy, average and strike rate of a bowler). Evolution of runs for a batsman i is defined as follows:



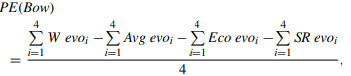
where, Rj are the runs made by batsman i during year j and Aj is denoting the average runs made by all the comparative batsmen during the same year j. Similarly, Avg evoi and SR evoi capture the evolution of a batsman i while considering its average and strike rate, respectively. Likewise, W evoi , Eco evoi , Avg evoi and SR evoi are computed to capture the evolution of a bowler i.

**PERFORMANCE EVOLUTION OF A BATSMAN (PE (B)):** Now, we formally present the mathematical notion for PE (B) that computes the year-based evolution score of a batsman i as follows:



where, 4 in denominator denotes first four years performance of an emerging batsman. The evolution indices appeared in numerator are added because all of them are positively correlated with the performance of a batsman.

**PERFORMANCE EVOLUTION OF A BOWLER (PE (BOW))**: Mathematical formulation of PE (Bow) is presented as follows:



where, 4 in denominator denotes first four years performance of an emerging bowler. Moreover, wickets taken by the bowler has positive correlation with its performance, therefore, it is added, while the remaining three features above the fraction have negative correlation with bowling performance, thus, that features values are subtracted.

Among 200 records belonging to each domain, 100 instances are representing players with the highest weighted average (a.k.a. RSs or positive samples), while other 100 instances are referring to players with the lowest weighted average (a.k.a. Not RSs or negative samples). Besides, 50 instances of positive samples and equally number of negative samples are randomly chosen for training and testing of datasets. Same kind of measures are considered for extracting the second dataset. Hence, both of the datasets fulfill the requirement of balanced dataset because of comprising equal number of negative and positive labeled records.