Solution Framework

Private LB - Rank 3

Nikhil Kotra
AV-Username - nikhil1e9
Email - nikhil@ph.iitr.ac.in

Find the full code implementation here - AV_match_prediction.ipynb
Detailed inference and walkthrough - Inference_notebook.ipynb
Github repo - https://github.com/nikhil-1e9/AV-Hackathon

Problem Statement

The task is to make precise predictions regarding the runs scored and wickets taken by each player who has been carefully selected to represent their respective teams, India and Australia 15 members squad in the highly anticipated ICC World Cup 2023 clash on Oct 8, 2023. Make use of data science models and techniques based upon extensive historical data encompassing both player and team performance, to make well-informed predictions.

About Data

The dataset contains the batting and bowling statistics of the 30 players selected for the ICC World Cup 2023 of both teams India and Australia. The dataset contains the batting and bowling stats of each ODI played by the cricketer throughout his career.

Approach

Preprocessing the data

The data contains 2575 non-null values and no null values for each column. Preprocessing steps include:

- Cleaning and converting runs_scored and wickets columns to int datatype.
- Replacing DNB, TDNB runs values to -1 to indicate the player didn't bat.
- Replacing missing wicket values to 0.
- Converting match_date column to datetime format.
- Extracting ground from the opposition columns.
- Cleaning opposition column to only include the opponent team.

Exploratory Data Analysis

In this step I analyzed and visualized the data by making some graphs and plots. This helped to know the distribution of the features and what to expect from our predictions. This also helped in choosing potential players for the playing XI of both teams. The steps include:

- Extracting count of matches played by each player in his career to identify experienced players.
- Visualizing distribution of runs scored and wickets taken by each player in his career to see how these varied with time.
- Extracting average runs and average wickets counts of all players from data to identify good and bad players.
- Extracting potential wicketkeeper options for both teams based on total stumpings made.
- Analyzing and comparing the spin bowlers to fast bowlers based on the past records on the Chennai pitch.

Ground Analysis

Extracted historical ground data for the MA Chidambaram Stadium in Chennai from https://stats.espncricinfo.com/ and stored the table into excel file from the website. The dataset can be found here - https://github.com/nikhil-1e9/AV-Hackathon/blob/main/Ground_Stats.csv. This dataset helped me to get a good idea of the total scores that can be made on the pitch by analyzing the past performances. I extracted the average total scores made by both India and Australia historically on this ground. I took only the first innings scores when calculating the average as these are more important because the team can score as

many runs playing first but in 2nd innings the runs are constrained to only how much runs the opponent team made so will not give an accurate estimate. Also matches with no result were excluded from this analysis.

Modeling

I experimented with different techniques. Some of these seemed to work really well and some gave very bad results.

Model 1: Average or Median of last N performances

In this model I just extracted the average or median runs scored and wickets taken by players in the last n number of matches they have played. This involved sorting the dates in descending order and selecting the mean or median of first n columns.

Model 2: Average of last N performances in IND vs AUS matches only

This is the extension of the first model where I selected the rows which contain only India vs Australia matches. This didn't worked very well because of very less amount of datapoints.

Model 3: Weighted average by date

This model assigns a weight to every individual match date of each player, with higher weight given to recent matches and lower weight given to older matches. This model tries to capture the form of the player as the runs he scored 5 or 10 years ago don't matter much and his recent performances depict his form and he is more likely to perform likewise. The weights are calculated as follows:

First, the total matches played (n) for the player is calculated and a separate data frame is created. Then to each entry or row for an individual player's dataframe the row number is multiplied to the inverse of n i.e. the weight for the first entry is 1/n, for the second is 2/n, and so on upto n/n=1 which is the weight given to the most recent match played. This way the older matches get the least weightage and the more recent matches get higher weightage.

After the weights are assigned, the weighted average runs and wickets are calculated for each player based on the following formulae:

Wtd avg runs = $\frac{\sum_{i=1}^{n} runs_i * wts_i}{\sum_{i=1}^{n} wts_i}$, **Wtd avg wickets** = $\frac{\sum_{i=1}^{n} wickets_i * wts_i}{\sum_{i=1}^{n} wts_i}$

This approach worked really well and is likely the best among all the three models.

Model 4: Prophet

In this approach I used Facebook's Prophet library which deals with time-series data quite effectively by automatically predicting patterns, seasonality etc. in data. This approach is very robust for time-series applications but this didn't work very well with such small amount of data. Also the dates are not consecutive and have huge number of gaps in between. The generated predictions of some players looked accurate and reliable but others have very unrealistic predictions. For example, the model predicted that Marcus Labuschagne will score over 300+ runs and Cameron Green will take 8 wickets which is quite impossible.

Predicting playing XI

The playing XI of both teams was predicted based on the historical performances of the players such as average runs scored, average wickets taken, stumpings done etc. I also analyzed the previous matches played on the Chennai pitch and came to the conclusion that spinners will perform better at the pitch compared to fast bowlers. This led to giving higher chance to spinners to be included in the team instead of pacers. The past performance analysis for the players led me to shortlist 13/15 players from each side. Shortlisting the last 2 players was a challenge and the unprocessed predictions from the weighted average model, historical data analysis, as well as my own cricket knowledge helped me in shortlisting the final playing XI for both teams.

Final predictions and Post Processing

Weighted average model was selected for the final predicitons with a bit of post processing. After making predictions from the weighted average model, these predictions were processed one final time depending on the potential playing 11 of both

teams calculated in the previous step. Therefore only 22 rows contain non-zero values and the 8 rows for runs and wickets columns for players not playing were set to zero.

Also Indian player runs scored have been multiplied by a weight factor based on the total predicted scores and ground data analysis to balance the predictions. The weight factor was calculated as **wt_factor = (India's avg score in Chennai/Total predicted score from model).** This was then multiplied to the runs predictions for India to obtain the final predictions for submission.

Conclusion and Summary

Some of the players have very less number of matches so it was really difficult to make accurate predictions with only this much amount of data. Weighted average model worked nicely but no method can predict the scores and wickets perfectly. Only after the match we will know what will happen.

As a final note, I would like to thank the Analytics Vidhya team for organizing this amazing hackathon. I had a lot of fun and learnings from this competition. Looking forward to see more such hackathons soon :)

For further queries or clarifications you can contact me at my mail- nikhil@ph.iitr.ac.in or at LinkedIn - https://www.linkedin.com/in/nikhil-kotra/