Solution Framework

Private LB - Rank 3

Nikhil Kotra

AV-Username - nikhil1e9

Email - [nikhil@ph.iitr.ac.in](mailto:nikhil@ph.iitr.ac.in)

Find the full code implementation here - [AV\_match\_prediction.ipynb](https://colab.research.google.com/drive/1D_ep1hxgvp__350aOH-uroKFkblB-MOD?usp=sharing)

**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.
* 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. These include:

* Extracting count of matches played by each player in his career.
* Visualizing distribution of runs scored and wickets taken by each player in his career.

**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. 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.

**Final model**

Weighted average model was selected for the final submission with a bit of post processing. The predictions were processed one final time depending on the playing 11 of both teams. Therefore only 22 columns will contain non-zero values and the 8 columns for players not playing will be zero. Also Indian player runs scored have been scaled by a factor based on the total predicted scores to balance the predictions.

**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.

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