

Mini Project - Data Analysis and Prediction

Task: Create a Score predictor Model.

The dataset is comprised of First Innings statistics of T20 international Cricket Matches. There are 15 Columns in the dataset:

Batting Team, Bowling Team, Venue, Current score, Is Powerplay, Balls_left to be bowled, Wickets_left, Current Run Rate, Top_Order, Middle Order, Lower Order, Tail, Last_five (Score in last five overs), Death_Overs (Last 5 Overs), Final Score (i.e. Target set for second Innings)

Add more following features based on given criteria:

- Pressure: If the batting team is under pressure by scoring at a run rate of 7 or below or not
- Aggression: Whether the batting team is batting aggressively by scoring 45 or more runs in the last 5 overs

Be creative and come up with more features. (bonus)

Steps to follow (after adding above features):

- 1) Data cleaning
- 2) Data Analysis (More Question in the later part)
- 3) Split between train and test
- 4) Create a linear regression model
- 5) Evaluate performance using MSE, MAE and R2 Score

Provide Insights from data based on following question:

- 1) Other than home ground, where did Sri Lanka play the most number of matches?
- 2) Which Team has the highest score and opposition against which team did it occur?
- 3) Which team has the highest run rate?
- 4) Which batting order (Top Order, Middle Order, Lower Order, or Tail) performed the best in terms of average runs scored?
- 5) What is the average pressure score in matches where the Tail was batting?
- 6) In how many matches was the Aggression Mode set to 1? describe the teams with most aggressive Matches
- 7) What is the average runs scored by India in the last five overs (Death Overs) across all matches?
- 8) Highest and Average score in Johannesburg?
- 9) Highest and Average target set against Pakistan?
- 10) Visualize the highest vs lowest score of a team.
- 11) Average Number of wickets fallen in death Overs?
- 12) Average target set by Pakistan in their homeland? Note: Homeland may include cities of Pakistan and UAE.
- 13) describe the average runs scored by each team under pressure

- 14) describe the average runs scored by each team in death overs
- 15) describe the average target by each Team.

You might want to describe them visually, for this you can use matplotlib or plotly libraries.

Deliverables:

- Jupyter notebook
- Document each question inside the Jupyter notebook