Model Card for Predicting Pakistan Cricket Match Outcomes*

Muhammad Abdullah Motasim

December 10, 2024

1 Model Details

- **Developed By**: Muhammad Abdullah Motasim

- **Contact**: abdullah.motasim@mail.utoronto.ca
- **License**: MIT
- **Model Name**: Pakistan Cricket Match Outcome Predictor

- **Version**: 1.0

- **Algorithm**: Logistic regression using a Bayesian generalized linear model, trained on an 80-20 train-test split of data sourced from `cricketdata` (Hyndman et al. 2024).
- **Features**:
  - `Year` (numerical)
  - `Opponent Team` (categorical: `team2`)
  - `Toss Win` (binary: 1 for won, 0 for lost)
- **Purpose**: To predict the likelihood of Pakistan winning a cricket match based on historical match data.

- **Frameworks Used**: R (R Core Team 2023) with packages: `tidyverse` (Wickham et al. 2019), `rstanarm` (Vehtari et al. 2022), `arrow` (Richardson et al. 2024), `caret` (Kuhn

*Model can me found at https://github.com/abdullah-motasim/Analyzing-Pakistans-Cricket-Data/tree/main/models

and Max 2008))

- **Questions or Comments**: abdullah.motasim@mail.utoronto.ca

2 Intended Use

- **Primary Intended Uses**:
  - Predict match outcomes for Pakistan cricket matches.
  - Analyze factors influencing match results, such as opponent team, toss outcomes, and time trends.
- **Primary Intended Users**: Cricket analysts, statisticians, fans, and sports strategists.
- **Out-of-scope uses**:
  - Real-time decision-making or betting applications.
  - Inferences about non-Pakistan matches or players.

3 Factors

3.1 Groups

- **Definition**: Groups refer to distinct categories represented in the evaluation data. For this model, the relevant groups are defined based on:
  - Opponent teams (`team2`).
  - Match year (`year`).
  - Toss win outcomes (`toss_win`).
- **Intersectional Analysis**: The model evaluates performance variations across combinations of year, opponent, and toss outcomes to assess fairness and representativeness.

3.2 Instrumentation

- **Input Data Instruments**:

  - Match data sources include cricket archives and datasets standardized for analysis.
  - Input data preprocessing ensures consistent formats, removing noise in features.
- **Potential Variability**: Accuracy might vary depending on the completeness and standardization of match records.

3.3 Environment

- **Context of Deployment**: The model is intended for offline analysis and reporting.
- **Environmental Factors**: External influences (e.g., match conditions, venue) are not explicitly modeled but may implicitly affect historical data trends.

3.4 Relevant Factors

- **Definition**: Foreseeable salient factors for which model performance may vary include:
  - Opponent teams: Performance trends differ significantly based on the opposing team's skill level and strategies.
  - Match years: Historical shifts in team dynamics and cricket rules may influence predictions.
  - Toss outcomes: Toss-winning trends are crucial in determining match results.
- **Determination**: These factors were identified based on cricket domain knowledge and statistical analysis of historical data trends.

3.5 Evaluation Factors

- **Definition**:
  - The evaluation reports include model performance across:
    * Opponent teams.
    * Toss win outcomes.
    * Yearly trends.
- **Reasoning for Selection**:

  - These factors align with the model's intended use case: analyzing historical patterns and trends in cricket.

  - Relevant factors like player-specific attributes were excluded due to limited data availability.

- **Relevance vs. Evaluation Factors**: While player-specific attributes or match venue conditions might be relevant for understanding outcomes, they were not included due to the lack of annotated datasets. Incorporating these factors would improve granularity but requires further data collection and standardization.

4 Model Metrics

- **Model Performance Measures**:
  - Accuracy: 0.70.

  - Confusion matrix: Stored in `models/confusion_matrix_win_chance.rds`.

- **Decision Thresholds**:
  - Binary classification threshold of 0.5 for predicting wins.

- **Evaluation Variations**:
  - Cross-validation to analyze robustness.

  - Subset evaluation by year and opponent.

5 Evaluation Data

- **Datasets**:
  - Cricket match data sourced from cricketdata R package, cleaned and aggregated into:
    * Player data (`data/02-analysis_data/cleaned_player_data.parquet`)

    * Match data (`data/02-analysis_data/cleaned_match_data.parquet`)

    * Pakistan-specific matches (`data/02-analysis_data/cleaned_pakistan_match_data.parquet`)

- **Motivation**: To understand performance patterns of Pakistan's cricket team over time and against different opponents.

- **Preprocessing**:
  - `Data Reshaping`: Renamed and reorganized columns in the batting, bowling, and fielding datasets to standardize the data (e.g., bat_innings, bat_runs) and removed unnecessary columns.
  - `Match Data Cleaning`: Converted column names to lowercase, selected relevant columns, and cleaned date formats; replaced missing "winner" values with "Draw".
  - `Team Reordering`: Ensured Pakistan was always listed as team1 in matches involving Pakistan, swapping team order when necessary.
  - `Filtering`: Subsetted match data to include only matches where Pakistan was team1 and filtered player data to include only Pakistani players.
  - `Data Aggregation`: Combined batting, bowling, and fielding data, then summarized player statistics (e.g., total innings, runs, wickets) while calculating batting averages, handling edge cases.

6 Training Data

- **Source**:
  - Player and match data files:
    * `cleaned_player_data.parquet`
    * `cleaned_match_data.parquet`
    * `cleaned_pakistan_match_data.parquet`.

- **Structure**: Includes match-level variables such as winner, toss winner, opponent team, and match year.

- **Size**: Split into training (80%) and testing (20%) datasets.

- **Preprocessing**:
  - `Added Binary Columns`: Created two new binary columns (pakistan_win and toss_win), encoding whether Pakistan won the match and toss, respectively (1 for win, 0 for loss).
  - `Calculated Aggregates`: Summarized match-level performance by calculating total matches, total wins, win rate, and tosses won by Pakistan for each year.

7 Quantitative Analyses

- **Unitary Results**:
  - Overall accuracy and metrics calculated for each year and opponent group.

- **Intersectional Results**:
  - Performance evaluation across intersections of toss win outcomes and opponents.

8 Ethical Considerations

- **Bias**: The model's coefficients reveal some bias towards opponent teams (e.g., higher probabilities against weaker teams like Bangladesh and Zimbabwe). This reflects historical trends but might perpetuate biases in cricket analysis.

- **Data Limitations**:
  - The dataset does not include contextual factors like weather, player fitness, or ground conditions, which could influence outcomes.
  - Matches against less frequent opponents may lead to inflated coefficients (e.g., Ireland).

- **Fairness**: The model does not account for changes in cricket rules or playing conditions over time.
- **Use Case and Impact**: While the model is primarily designed for analyzing Pakistan's historical match data, it could also be used to predict outcomes in future matches, influencing decisions in strategic planning, media analysis, or fan engagement. However, it should not be used for critical decision-making, such as betting, as its predictions are based on historical patterns and do not account for all real-time variables.
- **Mitigation**: To mitigate the potential risks of bias and inaccuracies, it is recommended to:
  - Use the model as a complementary tool rather than the sole decision-making basis.
  - Incorporate additional contextual features (e.g., player statistics, match conditions) in future model iterations to improve accuracy and fairness.
- **Potential Risks**:

- Misuse: There is a risk that the model could be used in ways not intended, such as in gambling or betting, where the consequences of incorrect predictions could have financial or reputational implications.
- Over-reliance: Relying too heavily on the model's predictions without considering external factors (e.g., player injuries or weather) could lead to misinformed decisions.

9 Caveats and Recommendations

- **Caveats**:
  - The model assumes linear relationships between predictors and outcomes, which may oversimplify complex dynamics in cricket matches.
  - The data includes only matches where Pakistan participated, limiting generalizability to other teams.
  - The "toss win" variable shows moderate predictive power, but its causal relationship with match outcomes is not guaranteed.
  - Excludes critical factors like player performance and weather conditions.
- **Recommendations**:
  - Include additional contextual features (e.g., match location, player statistics) for better predictions.
  - Use the model as a supplementary tool for research and analysis.
  - Regularly update the model with new data for better performance and relevance.

References

Hyndman, Rob, Charles Gray, Sayani Gupta, Timothy Hyndman, Hassan Rafique, and Jacquie Tran. 2024. *Cricketdata: International Cricket Data*. https://github.com/robjhyndman/cricketdata.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Vehtari, Aki, Daniel Simpson, Michael Betancourt, Paul B. Gelman, Jonathan Goodrich, Ben McElreath, Jonah Gabry, and Peter J. L. Chodera. 2022. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.