Model Card for Predicting Pakistan Cricket Match Outcomes*

Muhammad Abdullah Motasim

Invalid Date

1 Model Card for Predicting Pakistan Cricket Match Outcomes

1.1 Model Details

- **Model Name**: Pakistan Cricket Match Outcome Predictor
- **Algorithm**: Logistic Regression using Bayesian Generalized Linear Model (via `stan_glm`)
- **Purpose**: To predict the likelihood of Pakistan winning a cricket match based on historical match data.
- **Version**: 1.0
- **Developed By**: Muhammad Abdullah Motasim
- **Frameworks Used**: R (`rstanarm`, `dplyr`)

1.2 Intended Use

- **Primary Application**:
  - Predict match outcomes for Pakistan cricket matches.

*Model can me found at model/win_chance.rds and scripts/05-model_data.R

  - Analyze factors influencing match results, such as opponent team, toss outcomes, and time trends.

- **Intended Users**: Cricket analysts, statisticians, fans, and sports strategists.
- **Use Cases**:
  - Enhancing cricket strategies by identifying key predictors of success.
  - Historical performance analysis of Pakistan's cricket team.

1.3 Model Metrics

- **Training Data Accuracy**: Not explicitly calculated; the model was fit on 80% of the data split.
- **Testing Data Accuracy**: 82% (rounded from 0.82).
- **Confusion Matrix**:
  - **True Positives**: Correctly predicted Pakistan wins.
  - **True Negatives**: Correctly predicted losses.
  - **Errors**: Misclassifications in either class.
- **Model Coefficients**:
  - **Year**: `0.03 ± 0.04` (indicates a minor increasing trend in win probability over time).
  - **Opponent Teams**:
    * Strongest predictors:
      · **Bangladesh** (`3.58 ± 1.35`) indicates a higher probability of wins.
      · **West Indies** (`2.20 ± 0.86`) also shows a strong win probability.
      · **Zimbabwe** (`2.47 ± 1.51`).
    * Moderate predictors: **England**, **Sri Lanka**.

    * Weak/uncertain predictors: **India**, **South Africa**, **New Zealand**.

  - **Toss Win**: `0.59 ± 0.43` (winning the toss slightly increases the probability of winning the match).

1.4 Training Data

- **Source**:
  - Player and match data files: `cleaned_player_data.parquet`, `cleaned_match_data.parquet`, and `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**:
  - Filtering for Pakistan matches.
  - Encoding binary variables (e.g., match win/loss, toss win/loss).

1.5 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.

- **Transparency**: Coefficients and assumptions are clearly documented to ensure interpretability.

1.6 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.
- **Recommendations**:
  - Include additional contextual features (e.g., match location, player statistics) for better predictions.
  - Validate the model on more recent datasets to ensure reliability over time.
  - Be cautious when using the model for critical decision-making, as it simplifies cricket's multifaceted nature.

1.7 Acknowledgments

This model was inspired by Pakistan's rich cricketing history and aims to provide insights into factors affecting their performance. The Bayesian approach ensures probabilistic interpretability and robustness in predictions.

1.8 Contact Information

- **Developer**: Muhammad Abdullah Motasim
- **Email**: abdullah.motasim@mail.utoronto.ca
- **GitHub Repository**: https://github.com/abdullah-motasim/Analyzing-Pakistans-Cricket-Data/tree/main