Prediction Model for Sporting Betting Based on Cricket

A project by

Cesar Pena	Sneh Vartak
Jesi Joseph	Krushi Teja Reddy Padamati

for
Capstone - November 2023, CSUEB, California State University, Hayward, CA

INTRODUCTION

Cricket is a sport that was introduced to North America early in the 17th century, and in the 18th century, it spread to other parts of the world. Wikipedia describes Cricket as a bat and ball game with two teams of 11 players on a field at the center of which is a 22-yard (20-meter) pitch with wickets at each end, each having a pair of bails balanced on three stumps. In simple words, the team that scores the most runs wins the game. According to Fortuner Business insights, the global cricket equipment market size is projected to grow from $594.46 million in 2022 to $897.02 million by 2029, at a compound annual growth rate (CAGR) of 6.05%, and the cricket market is expected to reach $3.5 billion by 2029, at a CAGR of 28.6% during the forecast period of 2022–2029, according to market research provided in Meticulous Research.

This project addresses the gaps that exist between current sports betting business models of sports such as NFL, FIFA, etc., and cricket; and increases economic and knowledge opportunities among bettors and sports enthusiasts alike. In addition, based on our own inspiration, interests in sports, and the aspect of probabilities, the project attempts to predict and recommend the betting aspects with the best chances to individuals. As a result of this project, we intend to increase the winning probability of an individual betting on cricket teams and players, by applying different methods on the baseline and continually updated data to find the best winning probability.

RELATED WORK AND NOVELTY

Sports analytics grew with Cloud technologies across the globe and today, sports analytics are widely used in every facet of sports from fans to professionals. Economic activities outside of the sports business are emerging at this time through sports betting, though not widely found to be legal across all parts of the world. Cricket being a famous sport in many countries, is also experiencing sports betting; however, as the sports betting business has just begun to emerge, the research work, premade datasets, and models aren't abundant. Kaggle, Github, and Google searches showed more commercial information than research information and machine learning models [1], [2], & [3]. Descriptive and machine learning on bowler workload, score prediction, etc., but doesn't extend to the subject of sports betting [4]. We also found that Kapadia performed experimental work on winners based on winning the tosses and additional research work from Patil on forming a good team based on statistical data of players in order to improve the chance of winning [5][6]. Kampakis & Thomas attempted to predict the winner of a match back in 2015 with the PCA model, which we hope to improve the modern deep learning models [7].

Based on this research, we find that the existing work on cricket-based sports betting is limited, basic, and often lacks clarity for the normal sports betting community in relation to sports such as NFL, NBA, FIFA, etc., and hence it could be improved in all facets of sports betting using the current deep learning models [8]. Our project collects historical data of various teams and at the same time keeps the data afresh with continual updates, creates maps among competing teams, predicts sports bets based on cricket teams and players of these teams, and recommends bettors options with higher chances of winning.

BUSINESS PLAN SUMMARY

Our project would provide artificial intelligence-based suggestions for cricket betting and generate revenues for the bettors by providing the winning recommendations for money line, spread, over/under, and player metrics.

DATA

Data Collection: For the datasets for the cricket betting project, we web-scraped team and player data from ESPN, and ICC Cricket and the sports betting data from Oddsportal, Cricket Betting, Draft Kings, and FanDuel, using web-scraping techniques in Python/R. We chose to restrict the data collection only to the top eight teams (Australia, Bangladesh, England, New Zealand, Pakistan, South Africa, Sri Lanka, and India) and last few years. Over 5 weeks of data gathering through web scraping and manual collection, we amassed over 200 record sets for ICC ODI games for the teams we chose for our predictions. This was to ensure that the player and team data of the distant past didn't skew our predictions. The datasheet for the moneyline and spread predictions.

Feature	Type	Description
Team1_Win_Loss_Ratio	Float 64 bit	Team 1's win/loss ratio against Team 2
Team2_Win_Loss_Ratio	Float 64 bit	Team 2's win/loss ratio against Team 1
Avg_Max_Temp	Float 64 bit	Average max temperature of city where game is played
Avg_Min_Temp	Float 64 bit	Average minimum temperature of city where game is played
Avg_Rain	Float 64 bit	Average rain in city where game is played
Team1_Bat_Avg	Float 64 bit	Team 1's batting average
Team1_Bat_Avg_SR	Float 64 bit	Team 1's average batting strike rate
Team1_Bowl_Avg_Econ	Float 64 bit	Team 1's average economy of bowling
Team1_Bowl_Avg_Wkt	Float 64 bit	Team 1's average wickets per bowler
Team1_Bowl_Avg_SR	Float 64 bit	Team 1's average strike rate per bowler
Team2_Bat_Avg	Float 64 bit	Team 1's batting average
Team2_Bat_Avg_SR	Float 64 bit	Team 1's average batting strike rate
Team2_Bowl_Avg_Econ	Float 64 bit	Team 1's average economy of bowling
Team2_Bowl_Avg_Wkt	Float 64 bit	Team 1's average wickets per bowler
Team2_Bowl_Avg_SR	Float 64 bit	Team 1's average strike rate per bowler
Team1	Integer 64 bit	Team 1
Team2	Integer 64 bit	Team 2

The datasheet for over/under prediction is given below and it is a different dataset compared to the datasheet shown above. The dataset focuses on players and teams, to predict the runs to be scored.

Feature	Type	Description
Name_Idx	Integer 64 bit	Players name of a Team, whose Over/Under is predicted
M	Integer 64 bit	Number of matches played by the batsman
Inn	Integer 64 bit	Number of innings played by the batsman
NO	Integer 64 bit	Number of "not-outs"
Runs	Integer 64 bit	Number of runs played by the batsman
HS	Integer 64 bit	Highest score of the batsman
Avg	Float 64 bit	Total average runs of the player
BF	Integer 64 bit	Total number of balls faced by a player to score the runs
SR	Float 64 bit	The strike rate of the player

Datasets: We classified the scraped dataset into team, player, and weather datasets for the prediction of money line, spread, over/under, and player statistics. Team and weather datasets were to be used for moneyline and spread, whereas player dataset was to be used for over/under and player predictions.
- Team dataset: Data features of teams consisted of country, statistics on batting and bowling, date of the game, location, final scores, winners, and type of winning based on who elected to bat or bowl first (won by runs or wickets). Betting data for each match is also included, which shows the betting odds for each team winning. We added some calculated fields for enhancing the data set; we added a Win/Loss ratio for each team throughout these 200 matches.
- Weather dataset: For cities that host cricket stadiums around the world, we gathered monthly rainfall and temperature averages per month in the weather dataset.
- Player dataset: Player data features are runs made, balls faced, number of 50s, etc by a batsman and the number of overs bowled, number of dot balls, etc by a bowler. The batting data was gathered for the top 5 batsmen from these teams. The top 5 batsmen were chosen by selecting the players with the most runs in the World Cup. Each player's performance versus the other teams was collected. Batting metrics include runs, batting average, and strike rate. 4s and 6s. The weather data includes the average maximum and minimum temperature each month for cities where cricket stadiums are located.

Dataset Features: By applying descriptive analysis on the dataset, we were able to identify the related datasets that could explain our regression (dependent) variable from which we derive moneyline, spread, over/under, and player statistics. The coefficient of determination (r^2) and clusters leads us to the following dataset features (independent variables), which are to be used as the inputs to the machine learning models.

Moneyline and Spread
- Dependent feature: We choose to predict the winner as the dependent variable.
- Independent features: The primary features are both team's win-loss ratio against each other, batting average, batting average strike rate, bowler's average, average wickets taken by bowlers, and bowler's average strike rate. We

combined the weather data features, average maximum and minimum temperature, and average rain for the final prediction.

<u>Over/Under</u>
- ● Dependent feature: We selected to predict the average number of runs per player of a team and then combined all predictions to estimate the team score.
- ● Independent features: The features to predict the team score for all players of a team are 'Format of the matches played, number of matches played, not-outs, runs scored, average run based on the total number of games played, balls faced to score the run as an average per match and the number 100s, 200s, 50s, 4s, and 6s scored.

<u>Player Statistics</u>
- ● Dependent variable: This is the player statistic to be predicted and we chose it to be the expected runs to be scored for batsmen and the wickets taken for bowlers.
- ● Independent variables: To predict the possible runs from a player, we used runs scored against a specific opponent, the minutes it took to score, balls faced, number of 4s & 6s scored, strike rate, position played, dismissals, innings played, and the venue as the independent variables.

Data Preparation: We processed the input features out of our "Team Datasource", "Team-Player Datasource for Over/Under", and "Player Datasource" Google Sheets, published as CSV web resources. As part of input processing, we encoded the team names into numbers, and the rest of the features were transformed using Scikit's encoder and standard scaler encoders. We also dropped features that were not correlating based on descriptive analysis results and thus selected 16 features as a result for input processing.

APPROACH AND MODELS

Approach: Based on the metrics of interest (regression variables):
- ● Spread: Define a favorite and an underdog (aka Handicap):
- ● Moneyline: Bet on the winner based on Spread (aka Winner, Outright, winning odds)
- ● Totals or Over/Unders (aka Expected Points Added or Scored)
- ● Player Bets

our approach for creating models that predict the sports metrics is as follows:
1. Data Collection using Web Scraping: Our data was web-scraped on websites such as ESPN, ICC Cricket, etc. This data consists of information about teams and players in those teams.
2. Data Cleaning and Transformation in Python Pandas: We combined the data to ensure that the data size was adequate for our deep learning exercise; we applied data cleaning methods to ensure that the data quality was good.
3. Algorithms used: We used different algorithms for regression models but finalized them with Random Forest, Gradient Boost, and XGBoost algorithms. We dropped Logistic Linear Regression, LSTM, and Auto Encoder models after testing as they produced inconsistent results. Our goal is to apply multiple models to ensure that we apply the optimal model and select a model that has superior performance and accuracy to find the best results.
4. Build a report using Google Sheets: The Google Sheets dashboard provides recommendations to the bettors and performance data about the chosen bets.

We used the model pipeline containing training, testing, and prediction model phases to complete the sports betting predictions. This is shown below:
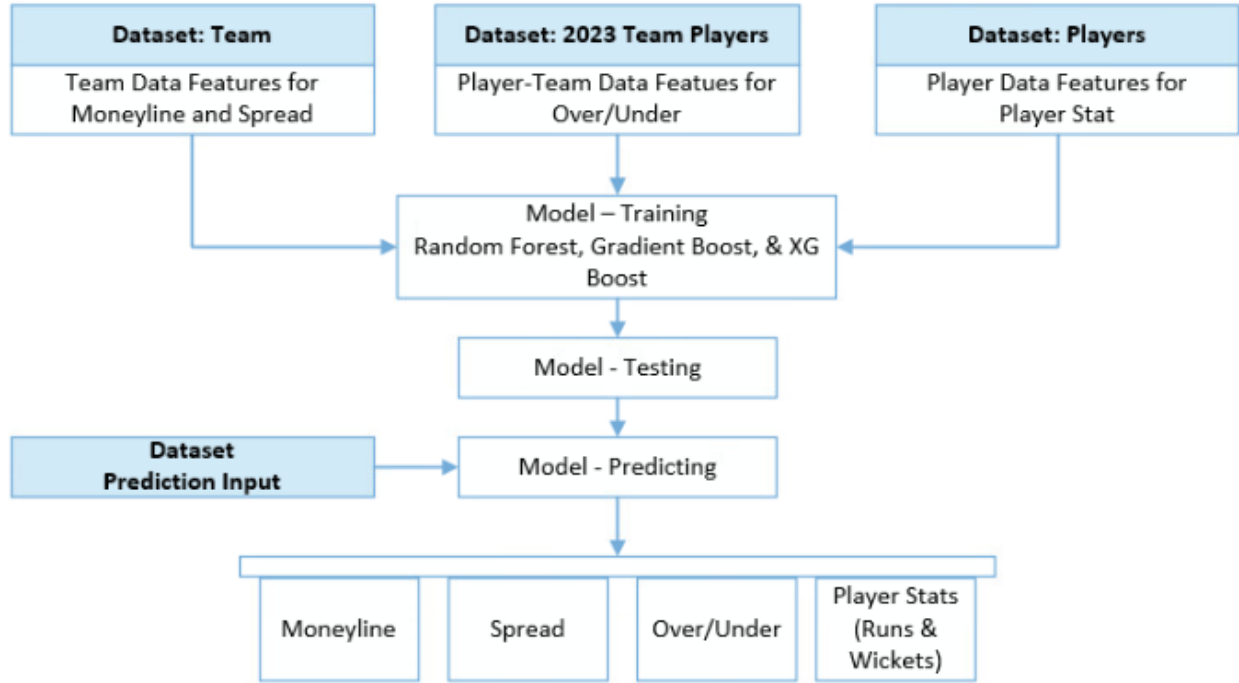
(Figure: Cricket betting project model)

Models: In regression, for the squared error loss, it is common to combine the random outcome of models into an ensemble and to average their outcomes:

$$\psi L, \theta1,..., \theta M(x) \;=\; \frac{1}{M} \sum_{m=1}^{M} \phi L, \theta m(x), \; where\; \phi L, \theta m\; is\; a\; predicted\; outcome\; of\; M\; samples$$

[13]

We chose various machine learning models to establish a linear relationship with our input datasets and the outcome variables. The predictors (Yi) are the player and team metrics, and the general form of a linear relationship between the regressor and predictor variables for player and team betting parameters is

$$Yi \;=\; \beta o \;+\; \beta iXi \;+\; \varepsilon i$$

where:

- *Yi is the population-dependent variable,*
- *βo is the population Y intercept,*
- *βi is the coefficient of population slope,*
- *Xi is the independent variable, and the ε is the random error.*

Moneyline: The moneyline-specific regression model for our sports betting is based on the odds from which we derive the winning probability.

$$Yodds(team\; 1,\; 2) \;=\; \left\| \left(\frac{100}{\left(\beta o + \beta i \sum_{i=1}^{i=5} (X) + \varepsilon i \right) - 1} \right) \right\|,\; if\; P(Y) > 0.5$$

$$Yodds(team\ 1,\ 2)\ =\ \left|\left(\frac{100}{1-\left(\beta o+\beta i\sum(X)+\varepsilon i\right)}\right)\right|,\ \text{if P(Y)} < 0.5$$

$$M\ =\ \frac{Yodds(team\ 1,2)}{100}\times\ B,\ \text{if}\ Yodds(team\ 1,\ 2) > 1$$

$$M\ =\ \left(B\ \div\left(\frac{Yodds(team\ 1,2)}{100}\right)\right)\ +\ B,\ \text{if}\ Yodds(team\ 1,\ 2) < 1$$

Where:
- *X is the dataset containing win-loss ratio, batting average, batting average strike rate, bowler's average, average wickets taken by bowlers, bowler's average strike rate, average maximum temperature, average minimum temperature, and average rain,*
- *M is the money line, and*
- *B is the betting amount in dollars*

Spread: The spread model for our sports betting is also based on odds and winning probability with the prediction of outcome from one team.

$$S\ =\ \left|\frac{Y}{(Y+100)}\right|,\ \text{if}\ Y < 1$$

$$S\ =\ \left|\frac{100}{(Y+100)}\right|,\ \text{if}\ Y > 1$$

Where
- *S = Spread for the chosen team*

Over/Under: The model for the expected average of runs of a team, which is set as a reference for over/under, is calculated by summing up the individual player's expected average score.

$$OU\ =\ \sum_{i=0}^{i=n}\left(\beta o\ +\ \beta i\sum(X)\ +\ \varepsilon i\right)$$

Where:
- *X is the dataset containing* runs scored against a specific opponent, the minutes it took to score, balls faced, number of 4s & 6s scored, strike rate, position played, dismissals, innings played, and the venue*,*
- *OU is over/under, and*
- *n is the number of players in the chosen team*

Player Statistics: This model predicts the runs a batsman will score and the number of wickets a bowler will take in a given match against a particular team.

$$X = \beta_0 + \beta_1 Opposition + \beta_2 Ground + \beta_3 Innings\ +\ \varepsilon$$

Where:

- X is either runs or Wickets
- β_0 is the intercept term.
- B_1 is the coefficient for the opposition team.
- B_2 is the coefficient for the ground where the match will be played.
- B_3 is the coefficient for the innings.
- ε is the error term.

The coefficients β_1, β_2, β_3, and β_4 are determined by the machine learning model and represent the expected change in the predicted number of wickets for a one-unit change in the corresponding feature. For example, if β_1 is positive, then we expect the predicted number of wickets to increase as the opposition team's strength increases. The error term ε represents the uncertainty in the prediction and accounts for factors that are not included in the model.

ANALYSIS

After creating the models, we were able to test them using the ICC World Cup series that ended on 11/19. We were able to use 6 games as the test cases and gathered prediction metrics on the money line, spread, over/under, and player metrics. Interestingly the accuracy of the real results closely matched with the model's prediction accuracy. Our models generally produced 83% accuracy, with data and training we got the MSE (mean squared error) as low as 24%, MAE (mean absolute error) 45%. We believe if the dataset is updated continually with fresh data, both MAE and MSE could tend toward zero. Using our project model (Figure: Cricket betting project model), We were able to predict the team betting moneyline, spread, and over/under for the chosen eight teams India, England, Australia, South Africa, New Zealand, Pakistan, Bangladesh, and Sri Lanka.

Moneyline: The team and weather input features were applied to the three regression models for the prediction of a binary outcome of win or loss and obtaining the regression. A hypothetical $100 bet on the money line based on our model prediction is shown here:

Date	Team	Opponent	Predicted Winner	Real Winner	Moneyline Bet on Winner	Payout on Moneyline Bet
11/19/2023	India	Australia	Australia	Australia	$100	$104.33
11/16/2023	Australia	South Africa	Australia	Australia	$100	$112
11/15/20023	India	New Zealand	New Zealand	India	$100	$203.33
11/11/2023	England	Pakistan	England	England	$100	$116.33
11/10/2023	Australia	Bangladesh	Australia	Australia	$100	$111.33
11/9/2023	New Zealand	Sri Lanka	New Zealand	New Zealand	$100	$116.67
				Total	$600	$763.99

We chose the bets based on winning possibilities and hence could be safe bets. However, if the bettor has a high-risk appetite, she/he could choose to bet on the losing team or a combination of both. 27.33% payout is expected for conservative betting, based on our model predictions.

Spread: Our spread prediction model is very similar to moneyline, in that we find out the probabilities of winning and losing and then we derive the spread. A hypothetical betting based on spread predictions of our models is given below. Incidentally, the prediction spreads of our models ended up being correct in cases where the winners were by runs. This is a low-risk bet that earned $10 if the bettor followed our predictions, which is that the team run spread would be within the given value in the column "Prediction Spread". Bettors have the freedom to choose relatively risky bets based on their risk preferences.

Date	Team	Opponent	Team Runs	Opponent Runs	Real Spread	Predicted Spread	Spread Bet	Payout on Spread Bet
11/19/2023	India	Australia	240	241	1	68	$100	$110
11/16/2023	Australia	South Africa	215	212	3	69	$100	$110
11/15/20023	India	New Zealand	397	327	70	68	$100	$110
11/11/2023	England	Pakistan	337	244	93	69	$100	$110
11/10/2023	Australia	Bangladesh	307	306	1	24	$100	$110
						Total	$500	$550

Over/Under: For Over/Unders which is defined as the expected value of bet (.50 x Under + .5 x Over), we applied the different regression models through the team and player metrics, including spread. A hypothetical betting based on our model's over/under predictions is shown below:

Date	Team	Opponent	Team Runs	Opponent Runs	Predicted Over/Under	Over/Under Bet on Under	Payout on Over/Under Bet
11/19/2023	India	Australia	240	241	318	$100	$200
11/16/2023	Australia	South Africa	215	212	237	$100	$200
11/15/20023	India	New Zealand	397	327	260	$100	$110
					Total	$300	$510

With the recommendation of "under" and the payout of $100 on under, and $10 on over, this payout's variance is high.

Player Metrics: We predicted the individual player statistics (batting average, number of wickets taken, number of catches, etc.) based on the individual's statistics data and applied regression models against them. Based on the regression model that yields the best prediction results, we calculated the highest possibility for scoring runs or taking wickets from that player [12]. To calculate player performance there are many unknown factors that one has to consider and consistently add or remove features like Player form, Player fitness level, Player Injuries, and Match conditions. The Player dataset consists of 2 sub datasets; the batsmen data set consists of Runs, Minutes, Balls_faced, 4's, 6's, Strike_Rate, Position, Dismissal, Innings, Opposition, and Ground data set; but, the most important features are Opposition, Position, Innings, Ground which are used to predict Runs for a future game.

Player's Team	Payer	Opponent	Innings	Prediction	Actual
India	Rohit Sharma	New Zealand	1	51 Runs	47 Runs
India	Jasprit Bumrah	New Zealand	1	1 Wicket	1 Wicket
India	Virat Kohli	Australia	2	60 Runs	54 Runs
India	Mohammed Siraj	Australia	1	1 Wicket	1 Wicket

CONCLUSION AND FUTURE WORK

We predicted the money line, spread, over/under, and player metrics based on 3 years of data we collected on the top eight teams, using random forest, gradient boost, and XGBoost algorithms. In general, we found consistency and agreement in predicting among these chosen models. Throughout the course of model creation, testing, and implementation, we continually learned that relevant and fresh data is the key part of the success of the models and accuracy in predicting betting outcomes. We noticed that there was a very high correlation between prediction and reality, indicating the model performance was adequate. We recommend this model for bettors to use as a strong guide before betting, with proper testing, improved and continually updated input features, controlled adoption of this model for practical use, and adapting the model based on the learning by experience.

We observed that the following could be considered in the future for expansion and enhancement:

- Expand the model to include large betting options such as props (various bets based on cricket-related metrics), and parlays (bets based on multiple games)
- Extend this model for the betting opportunities with women's cricket teams, and players
- Enhance the model accuracy by including advanced and timely input features on teams and players (features like team and player's form, fitness level, injuries, and so on).

The timing of the ICC World Cup greatly helped us to practically put the model to the test.

CONTRIBUTIONS

Cesar: Gathered batting data for player performance against each of the other teams, formatted to only include metrics relevant to the model. Calculated W/L ratio for team data. Collected team performance data for 2017-2020. Attended team meetings and contributed to the report.

Sneh: Spearheaded data gathering and preprocessing efforts for our project. Using Python, scraped player data for various teams and cleaned bowling data, selecting model-relevant metrics; also collected and transformed stadium, weather, and bowling data to seamlessly integrate into the team dataset. Additionally, compiled betting odds for teams. Extended support to assisting team members, and actively contributed to meetings and the final report.

Krushi: Web-scraped data for teams, batsmen, and bowler's international performance data to build and train the model to predict the winner between 2 teams, the number of runs a batsman will score in his next match against a particular team concerning the ground, the number of wickets a bowler is going to pick against a particular team concerning the ground. Participated in team meetings, planning sessions, and gathering requirements for the project's success. Contributed to writing the final report.

Jesi: Developed web-scraping code for the team and over/under models, processed the input datasets, tweaked the model code, and generated money line, spread, and over/under predictions; participated in team meetings that occurred at least thrice a week, planning sessions, and agile software development process with the team; contributed to writing the final report.

USE OF LLM

We used Google Bard and Chat GPT 4 throughout projects and found them to be very useful in "doing" the job. Our main use case was to get skeletal Python code for our model, which both LLMs handled extremely well. However, the skeletal code had to be fitted to our purpose, that is in every phase of modeling the skeletal code had to be "inserted" and modified to create outputs. The following was our approach to using LLMs:

- We have used both Chatgpt and Bard models for any help to make sure they are accurate and they are 95% accurate but their approach to the solution differs in some situations.
- We took help from Chatgpt and Bard for data scraping, Chatgpt helped us with the code needed for data scraping, and Bard helped scrap the website itself.
- Bard and Chatgpt were also helpful in combining multiple spreadsheets.
- The only concern we had while using LLMs was our prompt because that changed the output and the more clearly organized the prompt was the better and clearer the final result was.

Our view on the use of LLMs is that LLMs are very useful to quickly create building blocks, but not the solution in themselves. They are enablers and provide huge help in reducing the complexity. In our case, we were able to code the solution easily once our design was complete, data acquired, and outcomes were defined.

IMPLEMENTATION

Moneyline and Spread:
- https://colab.research.google.com/drive/1oLjR8jHT1ZKFfPbRJartrgb8M9VhoIBC?authuser=2#scrollTo=Mt9owiibUg21
- https://gist.github.com/kjfam/2af91ba2864e351dca7688eec5502fb7

Over/Under:
- https://colab.research.google.com/drive/1y-bv5-icbE7SJANHibNj2_kGEat8HpuZ?authuser=2#scrollTo=0CEZfO_nwNN0
- https://gist.github.com/kjfam/25b767ff8beec9200117645a0f8f5e70

Player Stat
- https://colab.research.google.com/drive/1w1Cp05qoLin4y522-Yt-nCLz7d_jkk92?authuser=2#scrollTo=4eb7aeb6

REFERENCES

[1] Kaggle: https://www.kaggle.com/search?q=cricket, 2023
[2] Google:
https://www.google.com/search?q=sports+betting+algorithm+for+cricket&rlz=1CAPOUW_enUS

897&oq=sports+betting+algorithm+for+cricket&aqs=chrome..69i57j33i160l4j33i22i29i30j33i15i22i29i30.16542j1j7&sourceid=chrome&ie=UTF-8, 2023

[3] Hydnman, Rob J: https://github.com/robjhyndman/cricketdata, 2023

[4] Wickramasinghe: Applications of machine learning in cricket: A systematic review, https://www.sciencedirect.com/science/article/pii/S2666827022001104, 2022

[5] Kapadia, Abdel-Jaber, Thabtah, & Hadi: Sport analytics for cricket game results using machine learning: An experimental study, https://www.emerald.com/insight/content/doi/10.1016/j.aci.2019.11.006/full/html, 2020

[6] Patil, Sequeira, Gonsalves, Singh, and Fr. Rodrigues: Cricket team prediction using machine learning techniques, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3572740, 2020

[7] Kampakis & Thomas: Using Machine Learning to Predict the Outcome of English County Twenty over Cricket Matches, https://arxiv.org/abs/1511.05837, 2015

[8] SR, Mahashree: IPL 2023: A closer look through the lens of data analytics, 2023

[9] ESPN: Records for India, https://www.espncricinfo.com/records/team/india-6, 2023

[10] ICC Cricket: Men's Cricket Rankings, https://www.icc-cricket.com/rankings/mens/overview, 2023

[11] Levine: Beating Vegas - Creating a dynamic sports betting model, 2019

[12] Gifford and Bayrak: A predictive analytics model for forecasting outcomes in NFL games using decision tree and logistic regression, Decision Analytics Journal, 2023

[13] Louppe, Gilles: Understanding Random Forests - From theory to practice, https://arxiv.org/pdf/1407.7502.pdf, Last Accessed: Nov 2023

[14] Cricket Equipment Market Size https://www.fortunebusinessinsights.com/cricket-equipment-market-104485

[15] Cricket Market by Product https://www.meticulousresearch.com/product/crickets-market-5247

[16] Cricket https://en.wikipedia.org/wiki/Cricket

ACKNOWLEDGEMENTS