Cricket Winner Prediction - Domain Based Feature Engineering and Analysis

Nazmus Sakib Ahmed
Inst. of Information Technology
University of Dhaka
Dhaka, Bangladesh
bsse1108@iit.du.ac.bd

Samiha Tahsin Noshin
Inst. of Information Technology
University of Dhaka
Dhaka, Bangladesh
bsse1107@iit.du.ac.bd

Shafiq Us Saleheen
Inst. of Information Technology
University of Dhaka
Dhaka, Bangladesh
bsse1125@iit.du.ac.bd

BM Mainul Hossain
Inst. of Information Technology
University of Dhaka
Dhaka, Bangladesh
mainul@iit.du.ac.bd

***Abstract*—Cricket, as one of the world's most captivating sports, presents a challenging task: predicting match winners. Our research delves into this intriguing challenge, aiming to forecast cricket match outcomes by analyzing a blend of external and internal factors related to the teams. We conducted experiments in feature engineering, seeking to enhance prediction accuracy. Despite modest a modest score, the overall accuracy remains a focal point for improvement. Our correlation analysis revealed limited association between the feature vector and match winners. However, it shows the way to the next steps in order to improve the prediction. This research not only contributes to the challenging task of predicting cricket match winners as well as serves as a testament to the evolving synergy between sports and data analytics. As the boundaries of accuracy continue to be pushed, our exploration paves the way for a deeper understanding of the intricate dynamics within the realm of sports analytics.**

***Index Terms*—cricket, machine learning, analytics**

I. Introduction

Cricket, as one of the world's most popular sports, stands out with its irresistible blend of skill, strategy, and suspense. It has become a captivating symbol of competition and endurance, featuring various formats, fiercely contested matches, and a global following. With its rich history dating back to the 16th century, cricket has continued to evolve over the centuries.

In today's sports landscape, where success depends on every goal, point, and second, data analytics is reshaping the game [1] [2] [3]. Athletes, coaches, and sports organizations are increasingly leveraging data-driven insights to gain a competitive edge. This research explores the profound impact of data analytics in sports, covering performance analysis and management. Athletes no longer rely solely on raw athleticism; data-driven insights inform tailored training, real-time tracking, and injury prevention, pushing the boundaries of human performance.

In recent years, cricket has also witnessed a data revolution. [4] With an ever-increasing accumulation of information during matches, this data provides researchers and sports enthusiasts with a compelling opportunity to learn more about the game. This includes ball-by-ball statistics, players' profiles, environmental factors, and more, presenting an exciting opportunity for researchers and enthusiasts. Several data science researchers have worked with cricket data, unleashing new dimensions in the sport.

The fusion of data analysis and cricket has led to groundbreaking discoveries in various academic investigations, expanding the sport beyond its traditional boundaries. The synergy between these academic activities and artificial intelligence techniques has unveiled previously unknown aspects of player performance, player management strategies, and even match result predictions. These discoveries could have a profound impact not only on players and teams but also on the enthusiasm and engagement of passionate cricket fans.

The ever-evolving landscape of AI research has transformed the sport of cricket over time. In-depth studies by researchers have revealed subtle inconsistencies that were hidden in plain sight, skillfully leveraging the vast amount of data accumulated during cricket matches and the power of machine learning algorithms. These efforts are leading to the development of a more intelligent, strategic, and dynamic approach to cricket.

In this paper, we attempt to predict match winners based on team, venue, and date-related data. Furthermore, we apply domain knowledge to generate features that enhance accuracy. With ample feature engineering we get a modest score in terms of accuracy. We collected the data from various sources, including match information from one-day internationals (ODI) over many years and their ball-by-ball data. In addition to the features we have used, this dataset holds further potential to achieve even better outcomes.

II. Related Works

Outcomes of matches in various sports have been the focus of several research endeavors. For instance, Kumash

Kapadia et al. [5] explored the prediction of winners in IPL T20 matches using a variety of algorithms, including Naïve Bayes, Random Forest, K-Nearest Neighbors, and Model Tree. Feature selection, a critical step in their research, was achieved through methods such as Correlation, Information Gain, Relief, and the Wrapper technique. In their study, the home-based model achieved a maximum accuracy of 57% with Naïve Bayes, while the toss-based model reached its highest accuracy of 62% through the K-Nearest Neighbors (KNN) algorithm.

Jhanwar [6] achieved a 71% prediction accuracy in determining the winner of One Day International (ODI) cricket matches. His approach involved employing binary classification models, including Logistic Regression, KNN, Random Forest, and Decision Trees. However, it's worth noting that a cross-validation procedure was notably omitted in his work. Additionally, Jhanwar's research focused on predicting match winners based on various factors, such as the end-of-over situation, recent and historical player performances, and other essential statistics crucial for match outcome prediction.

Neeraj Pathak and Hardik Wadhwa [7] employed Naïve Bayesian, Support Vector Machine, and Random Forest algorithms to forecast the outcomes of One Day International (ODI) matches. Their findings revealed that Support Vector Machine (SVM) outperformed the other methods with an accuracy rate of 61.67%. Notably, when dealing with imbalanced data, Naïve Bayesian algorithms demonstrated promising results. This research sheds light on the effectiveness of these predictive techniques in the context of ODI match predictions.

III. Data Collection

The data utilized in this study was acquired from two distinct sources: Kaggle and Cricsheet. [9] Kaggle provided two primary datasets, named "matchdata" and "matchinfo," containing tabulated information related to One Day International (ODI) cricket matches spanning the years 2002 to 2023. The column descriptions are given in table I and table II. Additionally, we leveraged JSON files sourced from Cricsheet to obtain the roster list.

Column Name	Description
match_id	Unique identifier for each ODI match (foreign key for id column of matchinfo).
season	The cricket season in which the match took place.
start_date	The date on which the match commenced.
venue	The stadium or venue where the match was played.
innings	Indicates the innings number (e.g., 1st innings, 2nd innings).
ball	The specific ball number in the over.
batting_team	The team at the batting crease during the innings.
bowling_team	The opposing team responsible for bowling.
striker	The batsman currently facing the ball.
non_striker	The batsman at the non-striker's end.
bowler	The bowler delivering the ball.
runs_off_bat	Runs scored off the bat, excluding extras.
extras	Additional runs attributed to extras (e.g., wides, no-balls).
wides	The number of wides bowled.
noballs	The number of no-balls bowled.
byes	Runs scored as byes.
legbyes	Runs scored as leg byes.
penalty	Penalty runs, if any.
wicket_type	The type of wicket taken (e.g., caught, bowled).
player_dismissed	Name of the dismissed player.
other_wicket_type	Additional wicket type information.
other_player_dismissed	Name of the dismissed player (in case of additional wickets).
cricsheet_id	Unique identifier from the Cricsheet database.

TABLE I
COLUMNS IN THE "MATCHDATA" DATASET

Column Name	Description
id	Unique identifier for each match.
season	The cricket season in which the match took place.
city	The city where the match was held.
date	The date of the match.
team1	The first team participating in the match.
team2	The second team participating in the match.
toss_winner	The team winning the toss.
toss_decision	The decision made by the toss-winning team (e.g., batting, bowling).
result	The result of the match (e.g., 'normal,' 'tie,' 'no result').
dl_applied	Whether the Duckworth-Lewis method was applied (1 if applied, 0 if not).
winner	The winning team of the match.
win_by_runs	The margin of victory in terms of runs.
win_by_wickets	The margin of victory in terms of wickets.
player_of_match	The player awarded 'Man of the Match.'
venue	The venue where the match took place.
umpire1	The name of the first umpire.
umpire2	The name of the second umpire.
umpire3	The name of the third umpire (if applicable).

TABLE II
COLUMNS IN THE "MATCHINFO" DATASET

IV. Data Preprocessing

A. Data Cleaning

To begin with, we established our initial feature vector using the matchinfo dataset. This dataset exhibited a considerable number of missing values, primarily stemming from situations where a match did not have a clear winner due to various external factors. To ensure the data's integrity, we embarked on a data cleaning process. One of our initial steps was to identify and handle teams that had participated in 30 or fewer matches, classifying them as outliers due to insufficient data for meaningful analysis. Subsequently, we removed columns that did not significantly influence our target variable, 'umpire1,' 'umpire2,' and 'umpire3.'

Moreover, we recognized that certain columns, such as 'player_of_the_match,' 'win_by_wickets,' and 'win_by_runs,' only had meaningful values in matches with a clear winner. Consequently, these columns were excluded from our input features. We also discarded the 'result' column, as it became irrelevant after excluding matches with inconclusive outcomes or ties. Additionally, the 'dl_applied' column was removed from consideration, as we exclusively focused on normally concluded matches.

B. Data Transformation

As we are predicting the match winner, we transformed the 'winner' column into a binary space where it equaled 1 when 'team1' emerged victorious and 0 when 'team2' claimed the win. Furthermore, we transformed the 'venue' and 'city' columns into new features, 'ishome1' and 'ishome2,' respectively. These binary variables indicated whether 'team_1' and 'team_2' were playing on their home turf or not. If a team was playing at their home ground, the corresponding 'ishome' value was set to 1; otherwise, it remained at 0. Because, according to studies, being at home effects the outcome of the result of a sporting event. [10]

To capture the significance of the toss outcome, we introduced a new feature called 'tossDecision,' which stored the toss result for 'team1.' This feature could assume one of four values: 'wonAndBat,' 'wonAndField,' 'lostAndBat,' or 'lostAndField.' We hypothesized that the decision made after winning the toss could be an informed choice influenced by field conditions, potentially holding valuable insights into the match outcome. Sood, G., & Willis, D. [11] argue that toss results has a significant impact on match results. We transform the 'date' values to month numbers ranging from 1 to 12. As, in a month, the weather may be similar all around and we were unable to obtain weather data, this may contain some information about the weather passively.

Lastly, we applied one-hot encoding [8] to categorical columns, namely 'team1,' 'team2,' and 'tossDecision,' to prepare the data for machine learning analysis.

C. Feature Creation

To enhance the accuracy of the models created features from the matchdata dataset as well as external data from cricksheet.

1) Economy: The term "economy" in cricket is often linked to the bowling component of the game. It alludes to a figure called the "economy rate," which expresses a bowler's effectiveness in terms of runs conceded. An important statistic for assessing a bowler's efficacy and performance is their economy rate. It is measured as the average runs a bowler gives up in an over, which is a set of six permitted deliveries. The following formula may be used to get the economic rate:

(Total Runs Conceded / Total Overs Bowled)

The phrase "economy rate" is typically used in relation to the bowler or bowling side and is one of several metrics used to determine a bowler's performance. An economy rate that is lower is better for the bowling side. The terms economy rate and run rate have the same meaning, while economy rate is related to a bowler's individual performance and run rate is typically used to describe a team's overall performance.

2) Expected Runs: We took the sum of the average runs by the players present in the squad in a match.

3) Expected Wickets: We took the sum of the average wickets taken by the players present in the squad in a match.

4) Expected Economy: The sum of the economy of 6 players with the least economy in the team.

5) Average Run: Runs scored on average in previous matches.

We introduce columns expRuns_team1, expWickets_team1, expEconomy_team1, expRuns_team2, expWickets_team2, expEconomy_team2 for expected runs, wickets, and economy (runs given) for team1 and team2, respectively. We take the difference between these columns to reduce the dimension of the features.

V. Models

A. Logistic Regression

Logistic Regression is a statistical method used for binary classification. It aims to find the relationship between dependent and independent variables and expresses this relationship as a probabilistic value within the (0, 1) range. The logistic regression model is typically expressed using the sigmoid function:

$$\text{Logit} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Where: - $\hat{y}$ is the predicted value. - $b_0, b_1, \ldots, b_n$ are the corresponding weights associated with the independent variables in the model. - $(y - \hat{y})$ represents the error or loss value of the model.

The loss function for logistic regression is the Log Loss, defined as:

$$\text{Log Loss} = -\frac{1}{D} \sum_{(x,y)} [y \log(y') + (1 - y) \log(1 - y')]$$

Where: - (x, y) represents the labeled pairs from the dataset. - y is the label in a labeled example, which is within the range (0, 1). - y' is the predicted value according to the given set of features in x.

In logistic regression models, regularization is of utmost importance. Without regularization, the loss can continue to be driven toward zero in large dimensions due to the asymptotic nature of logistic regression.

B. Decision Tree

A Decision Tree is one of the most powerful tools in machine learning for data classification and regression. It iteratively splits data into branches or nodes based on defined conditions until a stop threshold has been met, such as maximum tree height, minimum number of samples in one node, and the average number of leaf nodes. The prediction is represented by each leaf node of the tree.

Two common error calculation metrics used for decision trees, depending on the type of task (classification or regression), are Gini Impurity for classification and Mean Squared Error (MSE) for regression. An important measurement concept in decision trees is known as entropy, which measures impurities and uncertainties in an observation group and determines which data will be split by the Decision Tree.

	Metric	Logistic Regression	Decision Tree	Random Forest	Support Vector Machine
Without player and team centric features	Accuracy	0.71	0.62	0.65	0.70
	Precision	0.72	0.62	0.65	0.72
	Recall	0.71	0.61	0.65	0.70
	F1 Score	0.71	0.61	0.65	0.69
With player and team centric features	Accuracy	0.71	0.64	0.70	0.72
	Precision	0.72	0.64	0.70	0.73
	Recall	0.71	0.64	0.70	0.72
	F1 Score	0.71	0.64	0.70	0.71

TABLE III
RESULT TABLE

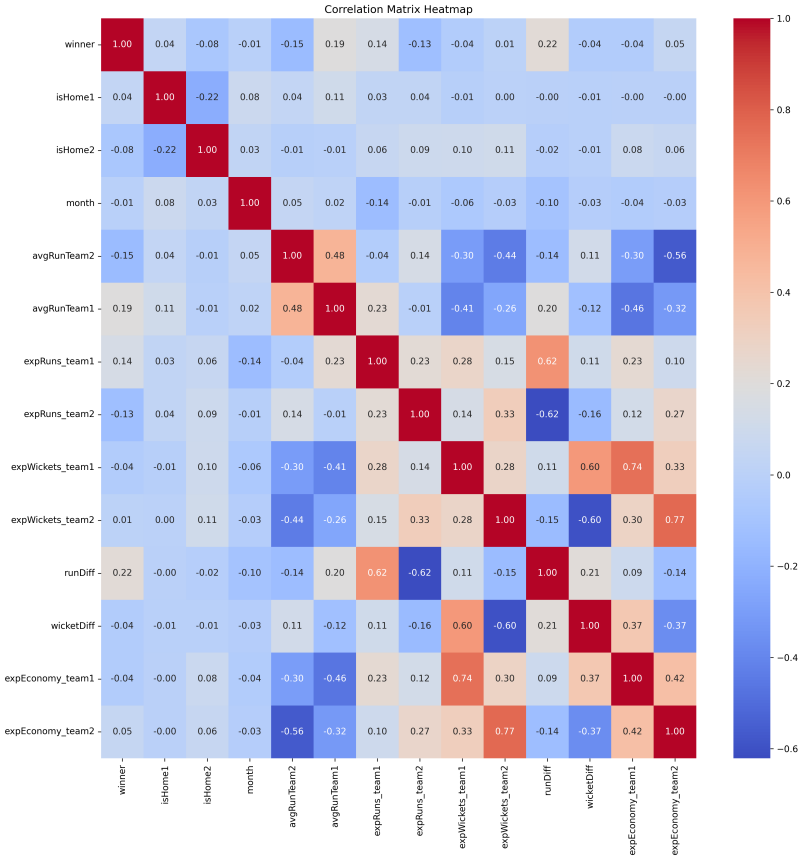

Fig. 1. Correlation Matrix

Considering a dataset with N classes, the entropy may be calculated using the formula:

$$E = -\sum_{i=1}^{N} p_i \log_2(p_i)$$

Information Gain is another essential concept in decision trees, measuring the reduction in uncertainty and entropy obtained by constructing a data set based on one particular feature. The objective is to obtain a feature that maximizes the reduction of uncertainty when using a split, leading to more precise and coherent data sets.

C. Random Forest

A Random Forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve predictive accuracy and control overfitting. The sub-sample size is controlled with the max-samples parameter if bootstrap=True (default); otherwise, the whole dataset is used to build each tree. The ”forest” it builds is an ensemble of decision trees, usually trained with the bagging method. The combination of learning models increases the overall result, according to the general idea of the bagging method.

D. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a highly efficient classification and regression analysis model based on supervised machine learning. It is used in this study to differentiate data points within their feature space. While linear regression works in lower-dimensional data classification, SVM searches for a hyperplane that maximizes the distance from one class to another, which is an essential objective of SVM. To achieve a high degree of generalization and sufficient performance, this margin is essential. Unlike other classification methods, SVM is generally robust to outliers in classification tasks, as long as they don't significantly affect the margin.

The objective is to find the hyperplane $w \cdot x + b = 0$ that maximizes the margin between the two classes. The margin is calculated by the distance between the nearest data points, supporting vectors from each class, to the hyperplane. The margin can be represented as:

$$\text{Margin} = \frac{2}{\|w\|}$$

Where $\|w\|$ is the Euclidean norm (magnitude) of the weight vector w. The Lagrange multipliers α_i are introduced to solve the constrained optimization problem. The final decision boundary can be expressed as $w \cdot x + b = 0$.

VI. Result and Analysis

A. Result

We have used four models on our feature vector, and among them, SVM and Logistic Regression provided the most consistent results with approximately 70% accuracy. Our attempt to enhance the accuracy by introducing new features resulted in some notable improvements in some cases. For example, the Support Vector Machine (SVM) model's accuracy improved from 70% to 72% when additional features were included. Random Forest model improved their accuracy by a 5% and became 70% and decision tree improved from 62% to 64%. However, Logistic Regression's accuracy remained the same at 71%. The results are summarized in table III.

B. Analysis

The features we have selected have not affected the result significantly according to the correlation matrix in figure 1. Furthermore, there seems to be no pattern between them and the target. We generate a correlation matrix for the feature and target where it shows that the expected run difference has the maximum correlation of 0.22. Average runs of the team are the second and third most correlated. However, the correlation value is very low in all of them. Features with higher correlation will enhance results further.

VII. Conclusion

Predicting the winner of a cricket match is both a high-demand and intriguing task. Our objective was to predict match outcomes based on a combination of external and internal factors related to the participating teams. Through various experiments in feature engineering, we created additional features to enhance our predictive models. Despite our efforts, the results demonstrated only minor improvements in accuracy. While there were some enhancements, the overall accuracy remains a potential area for improvement.

The correlation matrix analysis revealed a low correlation between the feature vector and the match winner, indicating the complexity of the task. For future research, incorporating weather data from match days could provide valuable insights, potentially influencing match outcomes. Additionally, exploring advanced techniques from the realm of deep learning might offer more sophisticated feature extraction methods, potentially leading to more accurate predictions.

Given our access to detailed ball-by-ball data, employing time series analysis could unearth valuable patterns, anomalies, and statistics. This deeper understanding could significantly contribute to improving the accuracy of our predictions, paving the way for more precise and reliable match outcome forecasts.

References

[1] Richter, C., O'Reilly, M., & Delahunt, E. (2021). Machine learning in sports science: challenges and opportunities. Sports Biomechanics, 1-7.

[2] Apostolou, K., & Tjortjis, C. (2019, July). Sports Analytics algorithms for performance prediction. In 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA) (pp. 1-4). IEEE.

[3] Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. International Journal of Data Science and Analytics, 5, 213-222.

[4] Passi, K., & Pandey, N. (2017). Predicting players' performance in one day international cricket matches using machine learning. Computer Science & Information Technology (CS & IT).

[5] Kapadia, K., Abdel-Jaber, H., Thabtah, F., & Hadi, W. (2020). Sport analytics for cricket game results using machine learning: An experimental study. Applied Computing and Informatics, 18(3/4), 256-266.

[6] Jhanwar, M. G., & Pudi, V. (2016, September). Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach. In MLSA@ PKDD/ECML.

[7] Pathak, N., & Wadhwa, H. (2016). Applications of modern classification techniques to predict the outcome of ODI cricket. Procedia Computer Science, 87, 55-60.

[8] Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

[9] Utkarsh Tomar. (2023). ODI Men's Cricket Match Data (2002-2023) [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/3780212

[10] Courneya, K. S., & Carron, A. V. (1992). The home advantage in sport competitions: A literature review. Journal of Sport & Exercise Psychology, 14(1).

[11] Sood, G., & Willis, D. (2016). Fairly Random: The Impact of Winning the Toss on the Probability of Winning. arXiv preprint arXiv:1605.08753.