RunProphet: Harnessing Machine Learning to Forecast ODI Cricket Scores

[1st] Dr. Mohammad Rezwanul Huq
Department of Computer Science & Engineering
East West University
Dhaka, Bamgladesh.
mrhuq@ewubd.edu

[2nd] Md. Arham Islam Khan
Department of Computer Science & Engineering
East West University
Dhaka, Bangladesh.
2020-1-60-116@std.ewubd.edu

[3rd] Md. Israkul Islam
Department of Computer Science & Engineering
East West University
Dhaka, Bangladesh.
2020-1-60-110@std.ewubd.edu

[4th] Md. Farhad Billah
Department of Computer Science & Engineering
East West University
Dhaka, Bangladesh
2020-2-60-213@std.ewubd.edu

Abstract— The sport of cricket presents a complex and dynamic set of factors that affect how a match turns out, especially in the form of One Day Internationals (ODIs). Machine learning has become a potent method for forecasting runs scored in ODI matches thanks to technological advancements and the accessibility of substantial cricket data. The methodologies and techniques used in ODI run prediction using machine learning are summarized in this paper. One-day international (ODI) cricket matches attract millions of dollars in wagers, the majority of which are placed after the contest has started. Home field advantage, previous performances, match experience, performance at the specific location, performance against the specific opposition, experience at the specific location, and current form are some examples of these characteristics. The study uses historical cricket data, which includes player statistics, team performance indicators, pitch characteristics, and weather predictions. Utilizing this data, machine learning techniques are used to develop predictive models that calculate the likelihood that a team will score a certain number of runs during an ODI match. The predicted accuracy of various machine learning techniques, such as regression models, ensemble methods, and deep learning, is investigated and compared. The crucial steps in developing reliable prediction models are clarified by detailed discussions of important factors like feature engineering, data preparation, and model evaluation measures. The limitations and difficulties of ODI run prediction are also discussed in the research, along with the effects of player form, match dynamics, and outside variables. The ultimate goal of this study is to show how machine learning may revolutionize the sport of cricket by offering insightful information to bettors, team strategists, and cricket fans alike. By harnessing the power of data-driven predictions, stakeholders in the world of cricket can make more informed decisions, leading to a deeper appreciation and understanding of this beloved sport.

Keywords—component, formatting, style, styling, insert (*key words****)***

I. INTRODUCTION

At the Melbourne Cricket Ground, Australia and England engaged in the first ever one-day international (ODI) game in 1971. Cricket began to gain popularity all across the world starting in the 13th century. Cricket was one of the earliest sports to employ statistics as a tool for comparison and presentation. Currently, cricket is recognized as a very lucrative and well-liked sport. The game's regulations are intricate. But because cricket is the second most watched sport in the world (after football) and has a lucrative economy.[1] One Day International (ODI) cricket match prediction is a difficult and entertaining assignment that has attracted a lot of interest in the field of sports analytics. It is now possible to create predictive models that can calculate the number of runs a side will score in an ODI match because to the development of machine learning techniques and the accessibility of enormous amounts of historical cricket data. Machine learning algorithms have altered the way we think about predicting cricket matches in this era of data-driven decision-making. In order to predict the likely outcome of an ODI match, these models take into account a wide range of variables, including team performance, player statistics, pitch conditions, weather forecasts, and previous match outcomes. For cricket lovers, spectators, and even bookies trying to make wise judgments, this predictive research can be of great use. Machine learning-based ODI run prediction can provide a data-driven viewpoint to improve your decision-making, whether you're an avid cricket fan looking for an edge in fantasy cricket leagues or a team manager planning for an upcoming match. In this prediction, we will examine the approaches, data sources, and crucial elements involved in applying machine learning to predict runs in ODI matches. We'll discuss in-depth on the algorithms, data pretreatment methods, and evaluation criteria that are employed in the construction of precise prediction models. By the end of this investigation, you'll have a better knowledge of how machine learning is changing how we think about predicting ODI matches and

opening up new opportunities for stakeholders and cricket fans alike.

II. BACKGROUND

One-Day International (ODI) cricket matches are played over 50 overs per side, and the team with the most runs at the end of the match wins. Cricket score prediction is the task of forecasting the final score of an ODI match before it is played. This can be a challenging task, as there are many factors that can affect the outcome of a match, such as the weather conditions, the pitch conditions, and the performance of the individual players.
There are a number of different machine learning algorithms that can be used for cricket score prediction. Some of the most popular algorithms include:

- Linear regression
- Lasso regression
- Ridge regression
- Random forest regressor
- Decision tree regressor

These algorithms can be used to train a model on a dataset of historical ODI matches. The model can then be used to predict the score of a new ODI match by providing it with the relevant input features, such as the teams playing, the weather conditions, and the pitch conditions.
The accuracy of cricket score prediction models has improved significantly in recent years, due to advances in machine learning and the availability of large datasets of historical ODI matches. However, there are still some challenges that need to be addressed, such as the need to improve the models' ability to handle unexpected events, such as player injuries and weather disruptions.

III. OBJECTIVE

The project's goals are to:

- Develop a model that can correctly forecast the total runs scored in an ODI cricket match.
- Identify the important variables that affect the total runs scored in an ODI cricket match.
- Future ODI cricket matches can be predicted using the algorithm.

The model can be applied to many different situations, including:

- To make betting and cricket fans' decisions more informed.
- To assist cricket teams in creating plans for maximizing their run-scoring ability.

Different machine learning algorithms, including support vector machines, random forests, and linear regression, can be used to create the model. Many other types of data, including historical match data, weather data, and player performance data, can be used to train the model. The model's accuracy will vary depending on a number of variables, including the caliber of the training data and the difficulty of the method used. Machine learning algorithms, however, have proven to be extremely successful at forecasting the number of runs scored in ODI cricket matches. Overall, the initiative to forecast runs in cricket matches using ODI has the potential to have a big impact on the game. The project can enhance the spectator experience, assist cricket teams in formulating better plans, and assist cricket boards and broadcasters in providing better coverage of the sport by creating a model that can precisely forecast how many runs will be scored during an ODI cricket match.

IV. METHODOLOGY

ALGORITHMS

1. **Linear regression:**

   A dependent variable and one or more independent variables are modelled using the statistical technique of linear regression by fitting a linear equation to the data. It has a long and rich history and is one of the most used statistical techniques in machine learning and data science. Simply put, linear regression is a method for determining the line that best fits a given set of data points. To determine causal connections, forecast future values, and comprehend complicated systems, linear regression can be utilised. It is a strong instrument that may be applied to a variety of issues.

2. **Lasso regression:**

   Lasso regression is a statistical modelling method that carries out both regularisation and variable selection. It operates by including a penalty term in the cost function of the linear regression model. This term pushes the model to choose a smaller subset of variables and to reduce the coefficients of the selected variables towards zero. Lasso regression is very beneficial for datasets with lots of variables or multicollinearity. Prediction, variable selection, and model interpretation are just a few of the many applications it is employed in.

3. Ridge regression:

Ridge regression is a statistical technique that carries out regularisation to raise the predictability and understandability of machine learning models. It functions by adding a penalty term to the linear regression model's cost function, which encourages the model to reduce the variable coefficients toward zero. For datasets with a lot of variables, multicollinearity, or noisy data, ridge regression is very helpful. Prediction, variable selection, and model interpretation are just a few of the many applications it is employed in.

4. Random forest regressor:

The supervised machine learning approach known as the random forest regressor uses ensemble learning to create a model for forecasting continuous values. It is a strong and adaptable method that may be used to address a variety of regression issues. It works especially effectively when dealing with issues involving high-dimensional data and non-linear correlations between the features and the target variable.

5. Decision tree regressor:

The decision tree regressor is a supervised machine learning technique that forecasts continuous values using a tree-like structure. It is an effective and flexible technique that can be used to handle a variety of regression issues. High-dimensional data and non-linear correlations between features and the target variable make it especially well-suited for these issues.

Data analysis:

In our project our dataset have 15 columns and 350899 rows. The columns names are :mid = Main Id,date = Date ,venue = Where the match is held , Bat_team = Batting Team ,Bowl Team = Bowling Team , Batsman = A person who hits the ball , Bowler = A person who through the ball , runs = Runs scored , wicket = Wicket , overs = Overs , run_last_5= Runs scored in last 5 Overs , wicket_last_5= Wickets in last 5 Overs , striker= Batsman playing as main 1 ,non_striker= Batsman playing as runner up (Not main 0), total= Total Score(Target Variable).

We have check the null value of dataset. We also check the data type of our project:

Integer(mid ,runs ,wickets ,runs_last_5 ,wickets_last_5 ,striker ,non-striker ,total)
Object(date ,venue,bat_team ,bowl_team ,batsman,bowler)
Float(overs)

Data Visualization:

Total Runs by Each Batting Team

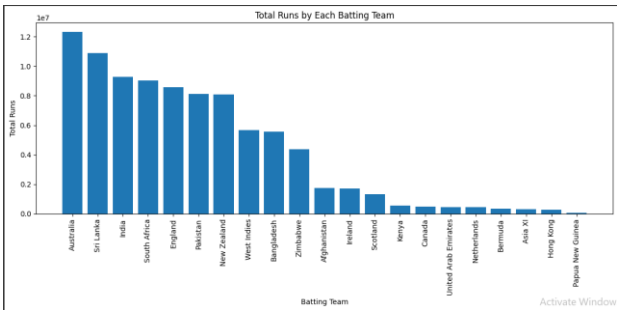

Figure 1. Total Runs by Each Batting Team.

Average Runs Scored by Each Batting Team

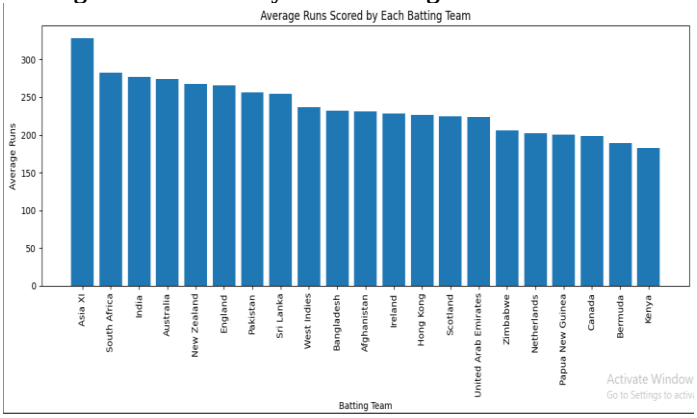

Figure 2. Average Runs Scored by Each Batting Team.

Total Wickets Taken by Each Bowling Team:

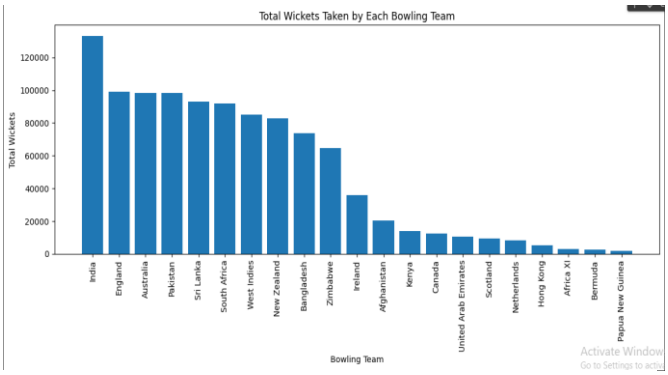

Figure 3. Total Wickets Taken by Each Bowling Team.

Average Wickets Taken by Each Bowling Team:

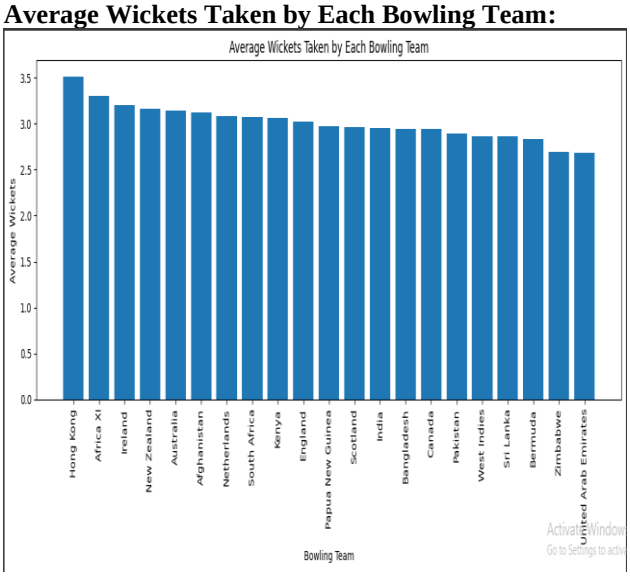

Figure 4. Average Wickets Taken by Each Bowling Team.

Batting Team Contribution to Total Runs:

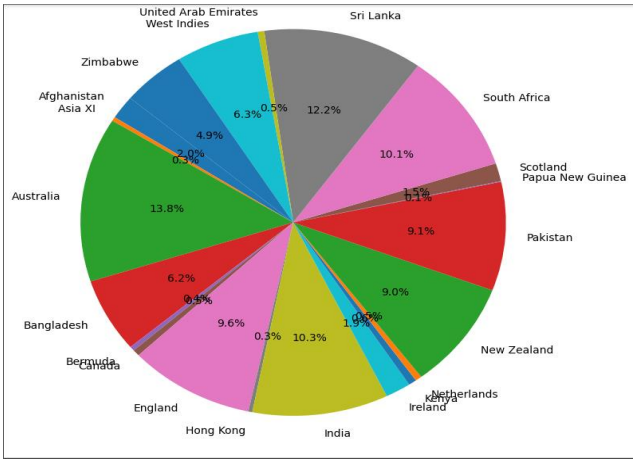

Figure 5. Batting Team Contribution to Total Runs.

Bowling Team Contribution to Total Wickets Taken:

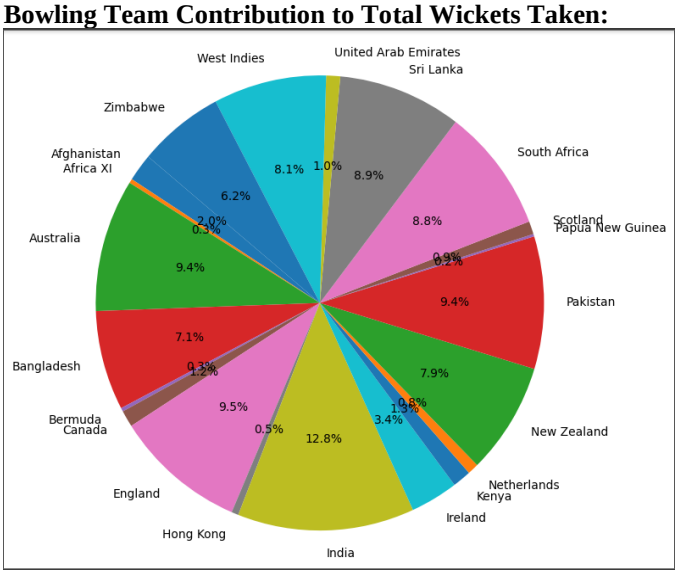

Figure 6. Bowling Team Contribution to Total Wickets Taken.

We have drop some columns like:('mid' , 'venue' , 'batsman', 'bowler', 'striker', 'non-striker')

ICC Oneday ranking batting team bowling team ranking:('New Zealand', 'Sri Lanka', 'South Africa', 'Australia', 'India', 'West Indies', 'Bangladesh', 'Pakistan', 'England', 'Zimbabwe', 'Ireland', 'Scotland', 'Netherlands', 'United Arab Emirates', 'Afghanistan')

There are many teams in our project who did not reach the last 5 overs we have filtered them. We have converted the object datatype to datetime64[ns].
For converting the object value into categorical value, we have done Feature Engineering in our project.
WE have found data after doing all this stuff about 298182. Where the Training data 76% (226922) and the Testing data 24% (71260).

V.Results

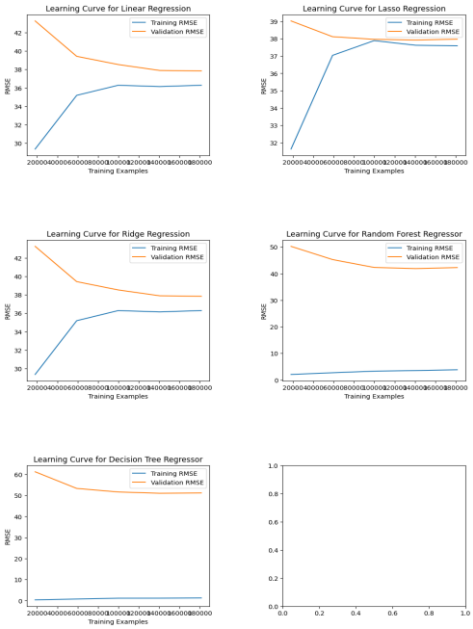

Figure 7. Learning Curves.

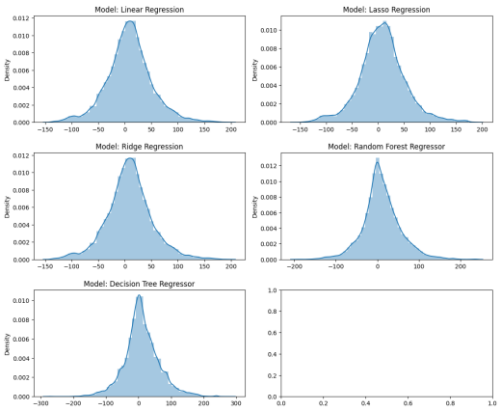

Figure 8. Iterate through models and plot their fit.

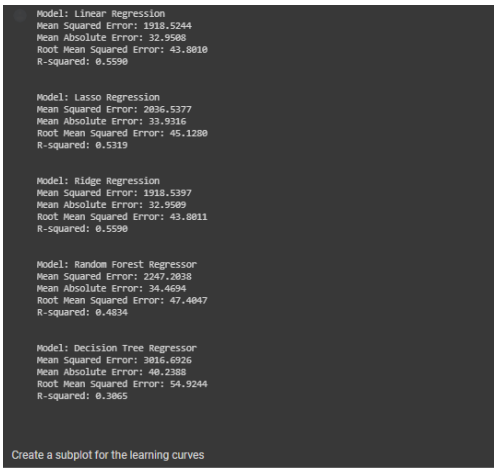

Figure 9. Each model's Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, R-squared.

VI. Conclusion

The Linear Regression model is the best because it makes predictions that are very close to the actual values, as shown by its low MSE, MAE and RMSE. It also has a high R-squared value (0.559), which means it fits the data well and explains a lot about the target variable. Overall, Linear Regression is the top-performing model among the others.

VII. References

1. Desai, K., Doshi, S., & Dholay, S. (2018). Predicting Outcome of ODI Cricket Games. International Journal of Emerging Research in Technology and Computer Science (IJERT), 3(01). https://www.ijert.org/predicting-outcome-of-odi-cricket-games?fbclid=IwAR37823XgOl0gpszOax9BPYKU9YVU6Hyg5c-m-DI30QTdXKgquAqKy-fSZ4

2. Micheal,B., Stephen,R.C.(2006). Predicting the Match Outcome in One Day International Cricket Matches, while the Game is in Progress,480–487 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3861745/?fbclid=IwAR1Kaupt_TgT_MHDK4n2MIICHlTzN7gjU-kvd26i_4jRUVaah3lM3HvUhM

3. Akarshe, S., Khade, R., Bankar, N., Khedkar, P., & Ahire, P. P. (2022). Cricket Score Prediction using Machine Learning Algorithms. International Journal of Emerging Research in Technology and Computer Science, 11(1), 50-58. https://www.researchgate.net/publication/346085808_Cricket_Prediction_using_Machine_Learning_Algorithms

4. Chowdhury, S., Islam, K. M. A., Rahman, M. M., Raisa, T. S., & Zayed, N. M. (2020). One Day International (ODI) Cricket Match Prediction In Logistic Analysis: India VS. Pakistan. Horizon Research Publishing. https://www.academia.edu/44829813/One_Day_International_ODI_Cricket_Match_Prediction_in_Logistic_Analysis_India_VS_Pakistan?fbclid=IwAR0EckBiSm74vTp3QCKbJR0iaJX-BcGoJun-6pc2BIVzmDN3zMq6zAo4cCQ

5. HK, P., R, P., Kumar, P., & Kumar, N. (2023). Cricket Score Prediction Using Machine Learning. International Journal of Engineering and Advanced Technology, 12(5), 102-107.https://ijirt.org/master/publishedpaper/IJIRT157821_PAPER.pdf

6. Manchanda, C. (2022, February 15). IPL Score Prediction Using Machine Learning. https://medium.com/mlearning-ai/ipl-score-prediction-using-machine-learning-f6587e353532