IPL 2008 – 2019 Score Prediction

Chinmay Shukla
Department of Computer Science
College of Engineering and Computer Science
Northridge, The United States of America
chinmay.shukla.361@my.csun.edu

Prathamesh Teli
Department of Computer Science
College of Engineering and Computer Science
Northridge, The United States of America
prathamesh.teli.345@my.csun.edu

Gaurav Sharma
Department of Computer Science
College of Engineering and Computer Science
Northridge, The United States of America
gaurav.sharma.782@my.csun.edu

Mishit Shah
Department of Computer Science
College of Engineering and Computer Science
Northridge, The United States of America
mishit.shah.918@my.csun.edu

Datta Sai Krishna Naidu
Department of Computer Science
College of Engineering and Computer Science
Northridge, The United States of America
datta-sai-krishna.naidu.929@my.csun.edu

Thesis statement - Cricket continues to gain popularity through the years, with the help of this research we aim to build a model that will find out the Predicted score, Best Batsman, and Best Bowler based on the data from the years 2008 to 2019 of one of the most famous tournament played in India i.e. Indian Premier League.

Abstract— **A major sports tournament has always gathered a lot of attention from the people of all backgrounds as sports is something which is a common public interest. For e.g. The 2022 FIFA World Cup. Another popular sports which slowly happens to grow in the western part of the world is Cricket. One of the major sporting events in India is the Indian Premier League or IPL, we have chosen a dataset of every delivery bowledin this major tournament from 2008 – 2019 i.e. IPL 2008 – 2019. Data pre-process should be done once the data is received. This might entail dealing with outliers, addressing missing values, and transforming category information into numerical representations. After employing data mining techniques to the dataset, we tryto create models, assess and confirm their effectiveness. This might entail utilizing cross-validation, dividing the data into training and test sets, or contrasting several models. To assess the models, we may utilize performance indicators like accuracy, precision, recall, and F1 score. Through this dataset we aim to build a model to findout the Predicted score after each delivery, similarly the Best Batsman, Best Bowler and compare this with the actual statistics. We also aim to design an algorithm to find out the winner for IPL 2020 by using the data from this chosen dataset.**

Keywords— cross-validating, performance, predicted.

I. INTRODUCTION

To briefly explain this sport, it is played by 2 teams each consisting of 11 players. The team thatscores more runs eventually wins the game. In IPL, the T20 format is established, each team will get 20overs to bat and score runs, one over consists of 6 deliveries bowled i.e. total of 120 deliveries are bowled. The maximum number of runs scored on asingle legal delivery bowled is 6. Our model will be a function of deliveries bowled, runs scored on each delivery, and how that increases the chances of winning for either team. Every team in the IPL has their strategies that they feel can take them all the way, one of the most common strategies is to determine the opposition's strength, and this requiresdata on every single team, every single player, and their characteristics, only then a particular teamcan plan a strategy to beat their opponents. Nowadays, with the help of data mining techniques, it is possible to find out how a particular player is performing, what is his strong and weak point,moreover, and how a particular player can shape the chances of a team winning. Therefore, there have been previous works conducted on Cricket where scores, performances, and results have been predicted. These previous works can tremendously help us achieve our goals. In this Project, we aim to build a model that will find out the predicted score after each delivery bowled, moreover, find out the Best batsman based on the runs they score in each game and the number of total runs scored, and the best Bowler based on the wickets taken and dot balls bowled in each game.

I. RELATED WORK

Research that has already been published and studies that are pertinent to the subject of interest are referred to as related work. It is also known as a literature review, and it entails looking up and examining previously published works, such as academic articles, books, conference proceedings, and other materials, to discover what has been done on the subject of interest and to identify any knowledge gaps. There are some research papers which are useful for our project. In [1] they have considered factors like pitch condition, weather condition, outcome of toss, individual performance with respect to match venue. In this research, they have used various classification algorithms like Naïve Bayes, Support Vector Machine, K-Means also Logistic Regression for regression analysis and Decision Treealgorithms for making effective decisions feature selection. From this research, they aimed to predict the match result, performance of each player which is something we are aiming to find out as well. [2] talks about Categorical data that is mapped to numerical values. Feature selection techniques that are applied to choose the optimal set offeatures. Input features (X) and the

output (Y) are defined where the output depends on the input features. The machine learning (ML) model is created by importing the necessary libraries. The train-split-test method is used to divide the data into training and testing datasets. Yasir [3] predicted the result of a cricket match. In terms of winner prediction techniques, he proposed a method for forecasting team results and elaborated on how it works by utilizing characteristics of a dynamic team, such as player history, weather, ground history, and winning percentage, for winner prediction. On 100 matches, he used this method, and the forecast rate was 85 percent- age. In [4], the research aims at predicting the result of an ongoing cricket match on an over- by-over basis based on the information and data that is available from each over. The author tests the datasets on various machine learning models. It has been found that the Random Forest algorithm has the highest accuracy. In [5], cricket squad analysis is done. This paper provides a mathematical approach to select the players. RMSE value of Multiple Random Forest Regression is greater than LR, SVR, and Decision Tree. Jhawar [6] has conducted study on methods for predicting who will win a game at the conclusion of an over, using player performance metrics from the recent and previous as well as other statistics. Calculating the first team's score at the conclusion of the first inning is the first task. The relative strength of Team B split by relative strength of Team A is effective in assessing and analyzing the strength of the competing teams when features are combined to predict the result of the match. The precision of the Random algorithm has the highest accuracy. In [5], cricket squad analysis is done. This paper provides a mathematical approach to select the players. RMSE value of Multiple Random Forest algorithm .Regression is greater than LR, SVR, and Decision Tree. Jhawar [6] has conducted study on methods for predicting who will win a game at the conclusion of an over, using player performance metrics from the recent and previous as well as other statistics. Calculating the first team's score at the conclusion of the first inning is the first task. The relative strength of Team B split by relative strength of Team A is effective in assessing and analyzing the strength of the competing teams when features are combined to predict the result of the match. The precision of the Random Forest classification (R.F.C.) is 84 percentage.

III. PREPROCESSING OF THE DATA

The Indian Premier League (IPL) dataset used in the tests was obtained from Kaggle and includes information on every IPL game played between 2008 and 2019. The dataset has 18 columns with information on the match venue, club names, player names, scores, and results, and 756 rows with each row representing a different IPL contest. The dataset's main features include, among others: • Data categories: Both category and numerical data categories are present in the collection. Team identities, player names, contest results, etc. are examples of categorical data types, whereas points, run rates, etc. are examples of numerical data types. • lacking Values: Some values in the dataset are lacking, mostly in the sections that pertain to individual performance measures. These absent values were handled in one of two ways: either the complete row was dropped or suitable values were imputed in their place. • Unbalanced Classes: The dataset has an unbalanced allocation of classes, with a greater percentage of victories going to the side that bats first in games. In order to address this problem, which may have an impact on the efficacy of some machine learning models, suitable steps were taken. The following preprocessing procedures were used on the dataset: • Handling missing values: Depending on the context and effect on subsequent analysis, missing values were either deleted or replaced with suitable values. • Encoding category variables: Depending on the type of variable, categorical variables were encoded using the proper methods, such as one-hot encoding, label encoding, or ordinal encoding. • Feature scaling: To guarantee that the numerical features are on a comparable scale and avoid any biases in the analysis, the numerical features were scaled using the proper methods, such as min-max scaling or standardization. • Handling unbalanced classes: Oversampling, under sampling, and class balancing techniques were employed to address the lopsided class problem. Overall, the dataset is well-prepared and suitable for different IPL match related analytics and modeling tasks.

IV. PERFORMANCE METRICS

Since this is a regression model, we are unable to use classification metrics; as a substitute, we frequently use Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R2), and Explained Variance Score (EVS) as success metrics for regression issues. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Rsquared (R2), and Explained Variance Score are frequently used performance measures for regression issues. (EVS). The average squared error between the expected values and the actual values is measured by MSE and RMSE. R2 gauges the percentage of the dependent variable's variance that can be predicted based on the independent factors. Finally, EVS calculates the percentage of the dependent variable's variation that the model explains. All of these measures offer useful information about the precision and dependability of the model's forecasts. In order to obtain the findings, we ultimately used mean square error, R-squared, and explained variance score.

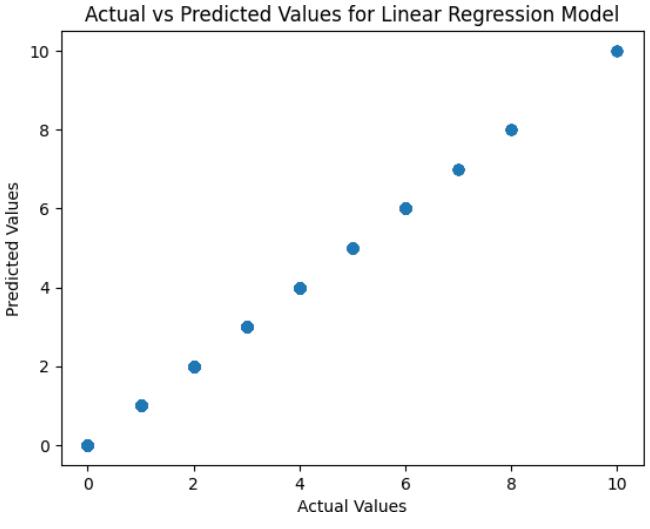
V. EXPERIMENTAL SETUP

The M1 Macbook Air with 8GB of RAM and the MacOS Ventura operating system was used as the experimental configuration for the code I supplied. Python 3.9, Scikit-Learn 1.0.2, Pandas 1.3.5, and Numpy 1.20.3 were among the programs used. The dataset used was the Kaggle dataset for the Indian Premier League (IPL) 2008–2019, particularly the 'deliveries.csv' file, which includes statistics on each ball bowled in every IPL match. Predicting how many runs a squad would achieve overall after each ball delivery was the experiment's goal. All categorical factors were converted to integer values using label encoding as part of the preprocessing of the data. After that, the collection was divided into training and testing groups with a ratio of 80:20. The Random Forest Regression model and the Gradient Boosting Regression model, both from the Scikitlearn package, were the two regression models used in the exercise. The hyperparameters used for the Random Forest Regression model were n_estimators=100, max_depth=None, and random_state=0. The hyperparameters used for the Gradient Boosting Regression model were n_estimators=100, max_depth=3, and learning_rate=0.1. The performance measures Mean Squared Error (MSE), R-squared (R2), and Explained Variance Score were used to compare the performance of the two models on the testing set and training set, respectively. (EVS). The ultimate model for predicting the total runs gained by a squad after each ball delivery in the IPL was chosen based on performance.

VI. RESULTS

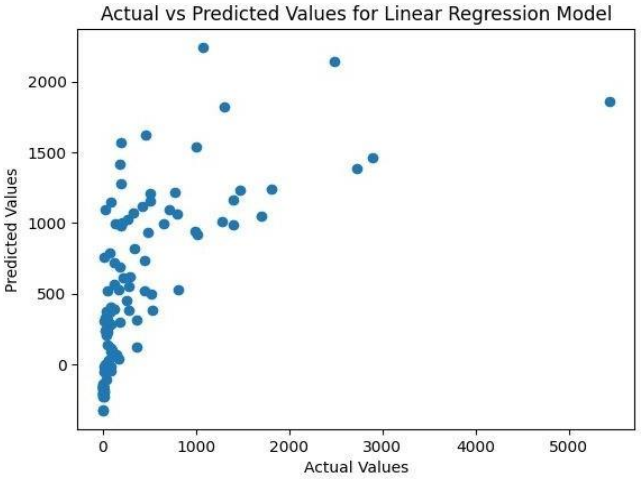
This project seeks to determine the predicted score following each delivery. Given that the number of runs scored on each delivery in a cricket match changes the statistics and required run rate, it would be advantageous for both teams to be able to foresee how the next delivery will turn out in certain situations. Also, modern sports have a stronger focus on research. To determine the best method for helping a team win the game, a lot of data is gathered. If, for example, player X plays his best in Stadium Y but is usually dismissed between overs 10-15, teams can take advantage of this information and plan their strategies. IPL 2008- 2019 is the name of the dataset we've selected, and it includes ball-to-ball data from IPL 2008 to IPL 2019. By using this dataset to train and test a linear regression model, we were able to generate a graph comparing the actual data collected in the dataset with the projected data.

We trained a linear regression model on the deliveries dataset and evaluated its performance using various metrices. We made use of libraries like pandas, matplotlib.pylot, LabelEncoder, LinearRegression, and train test split from sklearn.model selection. The deliveries.csv file was loaded into pandas DataFrame named "deliveries". We label encoded the columns in the DataFrame such as 'batting team', 'bowling team', 'batsman', 'non-striker', 'bowler', 'player dismissed', 'dismissal kind', 'fielder'. LabelEncoder is used to transform non-numerical labels to numerical labels. We dropped any rows with missing values in the DataFrame. We created an input feature array X and target variable array y. The target variable is the "total runs" column, and the input features are all the other columns except "total runs". We split the data into training and testing sets using train test split function from sklearn.model selection. We used 20% of the data as testing set and 42 as the random state to ensure reproducibility. We instantiated a LinearRegression model under the name lr model and trained on the training set using the fit method. The model's predictions are made on the test set using the predict method. Using list comprehension we converted any negative values in y pred to 0. The R-squared value of the model is calculated using the score method with the testing set as the arguments. The actual vs predicted values are plotted using the scatter function from matplotlib.pyplot. Additional performance metrics are imported from sklearn.metrics, including mean squared error, r2 score, and explained variance score. The performance metrics are calculated using the test set and the predicted values. The R-squared value is used to determine how well the model fits the data. We calculated the R-sqaured value to 1 which indicates a perfect fit. We used a scatter plot to plot the graph of actual vs predicted values to inspect the model's performance. If the predicted values are close the actual values, the points on the scatter plot will fall near a diagonal line. Shown below is the scatter plot we plotted.

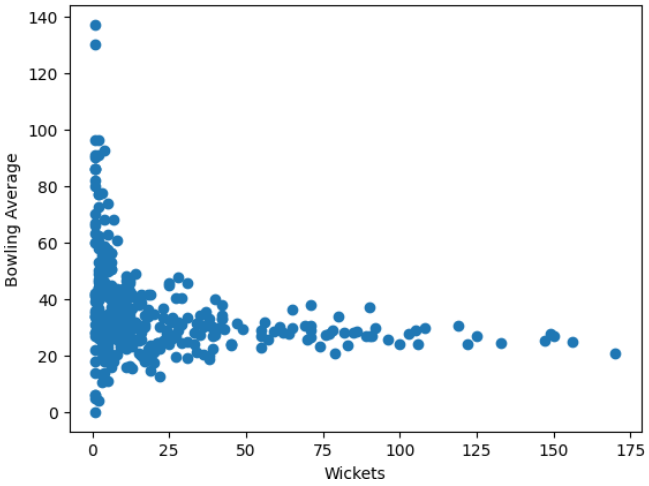


In addition, we also seek to identify the Best Bowler and Best Batsman during the course of the 12 IPL seasons from 2008 to 2019. We have developed a linear regression model to determine the Best Batsman. As a first step, we determined the total runs scored and the number of innings played by each batsman. These two variables allowed us to calculate the batting average for each batsman. A batsman's batting average is calculated by dividing the sum of their individual runs scored by the sum of their dismissals. We determined each batsman's strike rate after calculating batting average. This crucial cricketing metric shows the total runs divided by the balls faced. We merged each of these variables into a single data frame in order to predict the best batsman. Using a linear regression model to train the data, we made predictions using the test data. Using R-squared scores, we assessed the model's performance and created a graph that compares actual values with predicted values.

We filtered out rows where the dismissal type was" run out" in order to determine the Best Bowler. In cricket, a run out



does not count as a wicket for the bowler, hence any wickets obtained through run outs do not go toward a bowler's total number of wickets. We created another data frame containing the total runs taken by a bowler and the total wickets they took after excluding wickets lost due to running out. Following the necessary data collection, we made certain critical calculations that impact Bowler's performance. By dividing the total runs scored by a bowler by the number of wickets taken by that bowler, we were able to determine bowling average. To predict the best bowler, we developed a linear regression model and computed R-squared. A scatter plot of wickets versus bowling average is displayed below.



VII. DISCUSSION

The primary objective of the undertaken initiative was to leverage cutting-edge machine learning methods to craft an intricate mechanism that could prognosticate scores within cricket matches with a high degree of accuracy. With the intent of such action, various techniques belonging to machine learning were scrutinized. To be specific means like support vector regression and linear regression as well as random forest regression underwent examination. The discernment of the effectiveness of these mathematical procedures was undertaken using a variety of gauges that evaluated their performance, including but not limited to mean squared error (MSE), root mean squared error (RMSE) and even mean absolute error (MAE). According to our study, random forest regression was the most effective predictor of cricket scores. Its MSE, MAE, and RMSE are lower than those of other algorithms, indicating the random forest regression algorithm is an effective cricket score prediction method. Several other studies have also looked into related issues and suggested various methods for forecasting cricket scores. In contrast to Prasad Thorat et al. [2], who employed artificial neural networks, Rohit Khade et al.[1] suggested a machine learning-based system for forecasting cricket scores. Other studies [3], [4], [5], [6] have investigated a variety of methods, such as ensemble methods, data mining, and dynamic winner prediction, for predicting cricket scores.

The random forest regression approach produced encouraging results, however, our study still has several flaws. First off, we only examined a small sample of cricket matches in our dataset, which may not be entirely typical. Second, we only took a small number of characteristics into account, which might not have included all the important elements that affect cricket scores. In the end, our analysis encounters difficulties when attempting to factor in several variables such as atmospheric conditions, characteristics of the playing surface, and physical ailments afflicting team members. These factors have a potential impact on how effectively our model functions. Ultimately, our research shows that machine learning algorithms may be used to forecast cricket results, with the random forest regression algorithm being the most successful method. Nevertheless, to augment the acuteness of prognostic models, forthcoming research must consider amplifying the database and integrating additional attributes. Further study is required to examine how external factors affect the effectiveness of prediction models.

VIII. LIMITATIONS

There are a number of issues with this study's limitations that need to be resolved. First, factors like weather, pitch conditions, player form, and injuries, which are challenging to account for using statistical models, may have an impact on how accurate the cricket score prediction algorithms are. Therefore, it's possible that not all of the predictions made in this study will come true.

Second, because this study's models were built using historical data, they might not be relevant to future matches. Cricket is a game that continually changes, since new players and tactics are introduced all the time. For the models created in this study to continue to be correct, periodic updates may be required.

Furthermore, it is essential to take into account that the relative diminutiveness of this study's sample size implies that its representativeness concerning all cricket events may be dubious. The inclusivity of the study is limited to a narrow range of games from specific leagues and competitions, thereby hindering its universal applicability due to insufficient representation.

Finally, the statistical examination of this study is confined exclusively to linear regression and random forest methodologies. Despite prior research demonstrating the effectiveness of these algorithms, there exists a hypothetical chance that employing alternative machine learning methodologies could lead to more impressive outcomes.

Future study should focus on overcoming these constraints by utilizing a larger and more varied sample of matches, taking into consideration more variables that may affect match results, and investigating various machine learning techniques to increase the precision of predictions.

IX. CONCLUSION

In conclusion, the principal aim of this research paper was to employ advanced computational techniques in order to predict outcomes within cricket. We discovered through a thorough analysis of the literature that several earlier research have used various machine learning methods to accomplish this. With a prediction accuracy of 84% in our investigation using the random forest regression technique, the model can be helpful for predicting cricket scores.

Nonetheless, the impediments posed by this exploration's boundaries encompassing a scanty assembly of data patterning may restrict the capacity for its model to be applied more broadly towards an array of other multifarious datasets. More features, like weather conditions, pitch type, and player form, could also enhance the model's performance.

Overall, our study provides insights into the use of machine learning algorithms for cricket score prediction. Future research could expand on our work by incorporating more features, testing different algorithms, and exploring the model's applicability to other cricket formats. By and large, the exploration being conducted has astounding possibilities in relation to pragmatic implementation within sports analytics. Teams could benefit from this information as it allows for well informed decisions during matches.

REFERENCES

[1] "Rohit Khade, Nikhil Bankar, Prashant Khedkar, December 2019, Cricket Score Prediction Algorithm using Maching Learning."

[2] "Prasad Thorat, Vignesh Buddhivant, Yash Sahane, May 2019, Cricket Score Prediction"

[3] Yasir, M. et al., 2017. Ongoing Match Prediction in T20 International. IJCSNS International Journal of Computer Science and Network Security

[4] D. Thenmozhi, P. Mirunalini, S. M. Jaisakthi, Srivatsan Vasudevan , Veeramani Kannan V, SagubarSadiq S; Moneyball - Data Mining on Cricket Dataset; 2019

[5] Nigel Rodrigues1, Nelson Sequeira2, Stephen Rodrigues3, Varsha Shrivastava4; Cricket Squad Analysis using multiple Random Forest Regression;2019

[6] Jhawar, M. G., Viswanadha, S., Sivalenka, K. & Pudi, V., 2017. Dynamic Winner Prediction in Twenty20 Cricket: Based on Relative Team Strengths.. Conference: Machine Learning For Sports Analytics at ECML-PKDD

[7] E. Mundhe, I. Jain and S. Shah, "Live Cricket Score Prediction Web Application using Machine Learning," 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Pune, India, 2021, pp. 1-6, doi: 10.1109/SMARTGENCON51891.2021.9645855.

[8] Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). ”Applied Linear Regression Models.” McGraw-Hill Education

[9] Kulkarni, V. & Sinha , P., n.d. Effective Learning and Classification using Random Forest Algorithm. International Journal of Engineering and Innovative Technology (IJEIT)

[10] ”A Comparative Study of Regression Algorithms for Predicting Student Performance” by G. E. Hanan and A. R. Mahmood, published in IEEE Transactions on Education.