

Cricket Data Analysis and Score Prediction

A report submitted in partial fulfilment of the requirements for the award of Degree of

BACHELOR OF TECHNOLOGY

in

ELECTRONICS AND COMMUNICATION ENGINEERING

by

RAJESH DADALA

REG NO: 22B91A0452

National Institute of Electronics and Information Technology, Chennai

(Duration: 19th June 2024 to 13th August 2024)



DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

S.R.K.R ENGINEERING COLLEGE (AUTONOMOUS)

(Affiliated to JNTU, KAKINADA)

(Recognized by A.I.C.T.E, Accredited by N.B.A & N.A.A.C with 'A+' Grade, NewDelhi)

CHINNAMIRAM, BHIMAVARAM

SAGI RAMA KRISHNAM RAJU ENGINEERING COLLEGE

(Autonomous)

Chinaamiram, Bhimavaram

Department of Electronics and Communication Engineering



CERTIFICATE

This is to certify that the Summer **Internship Report** titled **“CRICKET DATA ANALYSIS AND SCORE PREDICTION”** is the Bonafide work done by **Mr. Rajesh Dadala**, holding the registration number **22B91A0452** and submitted during 2023-2024 academic year, in partial fulfilment of the requirements for the award of the Summer Internship Program for **Bachelor of Technology in Electronics and Communication Engineering (ECE)**,
At **National Institute of Electronics and Information Technology** from 19-06-2024 to 13-08-2024.

Department Internship coordinator

Dean-T&P Cell

Head of the Department

Acknowledgement

I take immense pleasure in thanking [Dr. N. Udaya Kumar](#), Professor & Head of the Department ECE, and [Dr. K. R. Satyanarayana](#), Dean Training & Placement Cell, SRKR Engineering College, Bhimavaram for having permitted me to carry out this project work.

I Would like to extend my sincere acknowledgement to those who have supported and encouraged me during this tough journey. Many people met me during this endeavor and enriched me with their support and knowledge both personally and professionally that resulted in the project being better than it could possibly have been without them.

I Express my sincere gratitude to my Internship Co-Ordinator [Mrs. G. Prathima](#), who assisted me throughout this project. I Thank her for providing me the reinforcement, confidence and most importantly the track for the project whenever I needed it.

I wish to record our sincere gratitude to NIELIT

Coordinators for their constant support and encouragement in preparation of this report and for making available videos and interface facilities needed to prepare this report.



Ministry of
Electronics and
Information Technology,
Government of India



Certificate No.: NVA-C-IN31-3-2665

Registration No.: 2024-06-19-IN31-34397

राष्ट्रीय इलेक्ट्रॉनिकी एवं सूचना प्रौद्योगिकी संस्थान (नाइलिट)
NATIONAL INSTITUTE OF ELECTRONICS & INFORMATION TECHNOLOGY (NIELIT)

इलेक्ट्रॉनिकी और सूचना प्रौद्योगिकी मंत्रालय भारत सरकार की एक स्वायत्त वैज्ञानिक संस्था
An Autonomous Scientific Society under Ministry of Electronics & Information Technology (MeitY)

NIELIT VIRTUAL ACADEMY

This is to Certify that

Mr. Dadala Rajesh

has successfully completed

Online Internship in Artificial Intelligence and Machine Learning using Python

Of Duration: 8 weeks from 19-06-2024 to 13-08-2024.

DATE: 18-08-2024


HEAD-ACADEMICS

DECLARATION

I undersigned hereby declare that the project report “CRICKET DATA ANALYSIS AND SCORE PREDICTION”, submitted for partial fulfilment of the requirements of the internship in National Institute of Electronic & IT, Chennai is a Bonafide work done by me under supervision of Mr. Ishant Kumar Bajpai. This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to academic honesty and integrity ethics and have not misrepresented or fabricated any data, idea, fact, or source in my submission. I understand that any violation of the above will cause disciplinary action by the institute and/or the University and can also evoke penal action from the sources that have thus not been properly cited or from whom proper permission has not been obtained. This report has not previously formed the basis for awarding any degree, diploma, or similar title of any other University.

By

Rajesh Dadala

TABLE OF CONTENTS

- Abstract
- Introduction
- Problem Statement
- Objectives
- Literature Review
- Methodology
 - i. Data Collection
 - ii. Data Preprocessing
 - iii. Feature Engineering
 - iv. Model Training
 - v. Model Evaluation
 - vi. Visualization and Reporting
- Algorithm and Pseudocode
- Implementation
- Results and Analysis
- Discussion
- Conclusion
- References

Abstract

The project "Cricket Analysis with Machine Learning Regression and Score Predictor" leverages machine learning techniques to predict cricket match scores, aiming to improve strategic decision-making and fan engagement. Using historical cricket match data, regression models like Linear Regression and Random Forest Regression are developed and evaluated. The data features include runs scored, strike rates, and dismissal methods, which undergo preprocessing to ensure quality inputs for the models. The performance of these models is measured using the R-squared (R^2) metric, with a score of 0.82, indicating that 82% of the variance in match scores is explained by the models.

The project demonstrates the effectiveness of machine learning in transforming raw cricket data into actionable insights. It highlights the practical applications of predictive analytics in coaching, broadcasting, and fantasy sports. The study also sets the groundwork for future improvements, suggesting the integration of real-time features like player fitness and weather data, as well as exploring deep learning techniques for greater accuracy. Overall, the project showcases the potential of data-driven approaches in enhancing sports analytics and provides a hands-on learning experience in machine learning.

Introduction

This project explores how machine learning can transform cricket score prediction by analyzing vast amounts of cricket data. The motivation lies in the growing demand for accurate predictions to assist strategic planning in cricket, enhance fan experiences, and foster better engagement. The integration of regression techniques aims to bring innovation to sports analytics.

Problem Statement

Cricket score prediction is a complex task due to numerous influencing factors. Current models often lack accuracy and do not leverage the full potential of data analytics. This project seeks to develop robust predictive models to address these challenges and improve decision-making in cricket.

Objectives

To develop predictive models for cricket scores.

To compare Linear Regression and Random Forest Regression techniques.

To extract actionable insights from match data.

Literature Review

This section discusses existing cricket prediction techniques and their limitations. It explains how this project enhances previous methods by applying machine learning algorithms and comparing their performances for better predictive accuracy.

Methodology

Data Collection

The data is sourced from Kaggle and includes match and player statistics.

Dataset Loaded: dataset1_test.csv

Data Preprocessing

Missing values are handled to ensure clean data.

Data features are standardized and formatted for analysis.

Feature Engineering

Key features like Runs, Strike Rate, Dismissal At, etc., are used for predictive modeling.

Model Training

Models used: Linear Regression and Random Forest Regression

Training data is split into training and testing sets.

Model Evaluation

Metrics like R-squared (R^2) are used to evaluate model performance.

An R^2 score of 0.82 indicates good model accuracy.

Visualization and Reporting

Results are visualized using graphs to compare actual vs. predicted scores.

Algorithm and Pseudocode

Workflow Pseudocode

BEGIN

Load and preprocess the data

Handle missing values

Split the data into training and testing sets

Initialize Linear Regression and Random Forest models

Train both models on the training set

Evaluate both models on the testing set using R-squared score

Visualize the predictions using scatter plots

END

IMPLEMENTATION

IMPORTING LIBRARIES:

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
#warnings.simplefilter(action = "ignore", category = FutureWarning)
```

LOADING THE DATASET:

```
matches = pd.read_csv('dataset1_test.csv')
```

SPLITTING DATA BY BATSMAN:

```
# split individual player records

batsman1 = Records[(Records.Batsman==1)]
batsman2 = Records[(Records.Batsman==2)]
batsman3 = Records[(Records.Batsman==3)]
batsman4 = Records[(Records.Batsman==4)]
batsman5 = Records[(Records.Batsman==5)]
batsman6 = Records[(Records.Batsman==6)]
```

Results and Analysis

Performance Metrics

The Linear Regression model achieved an R^2 score of 0.74, suggesting moderate accuracy in predicting cricket scores.

The Random Forest Regression model achieved an R^2 score of 0.82, showing better performance by capturing more complex relationships in the data.

Visual Analysis

Scatter plots comparing the actual scores to predicted scores for both models revealed that the Random Forest model had a more concentrated cluster around the line of perfect prediction, indicating its superior performance.

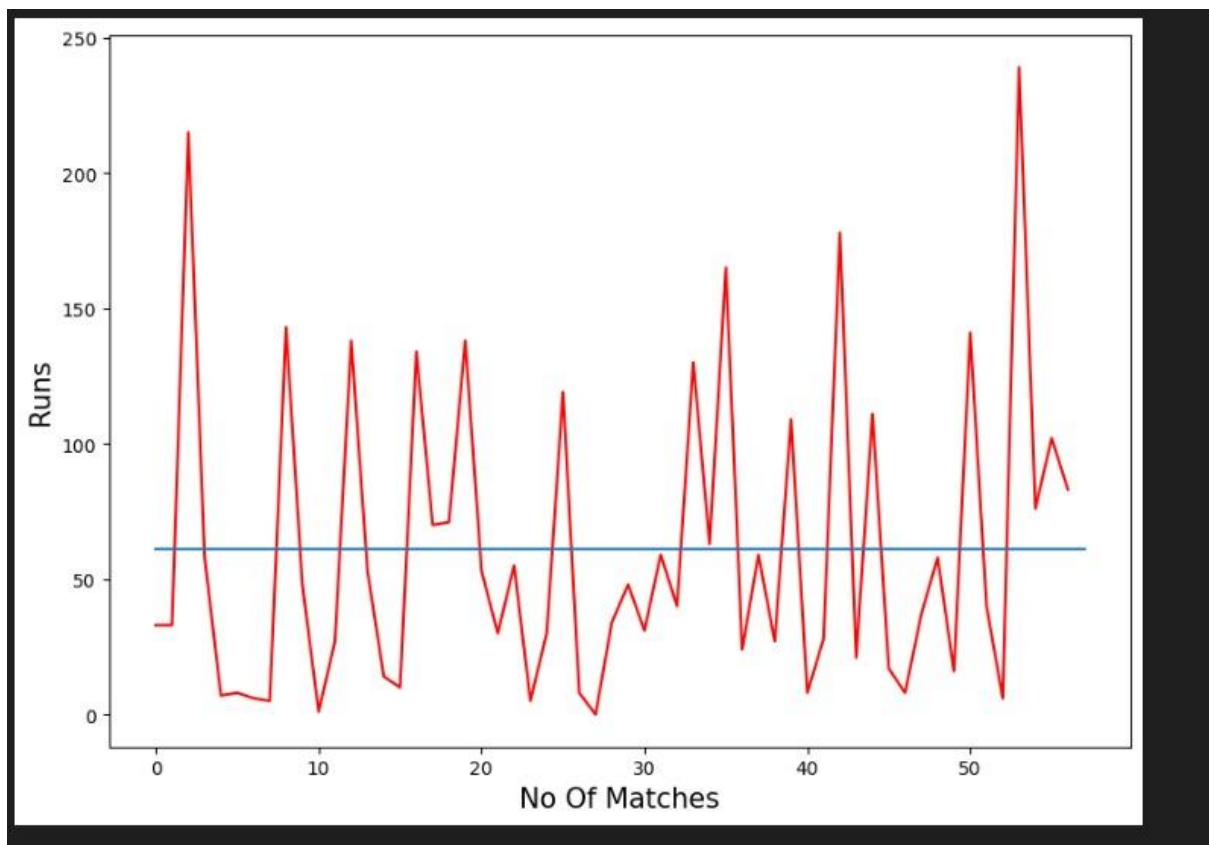
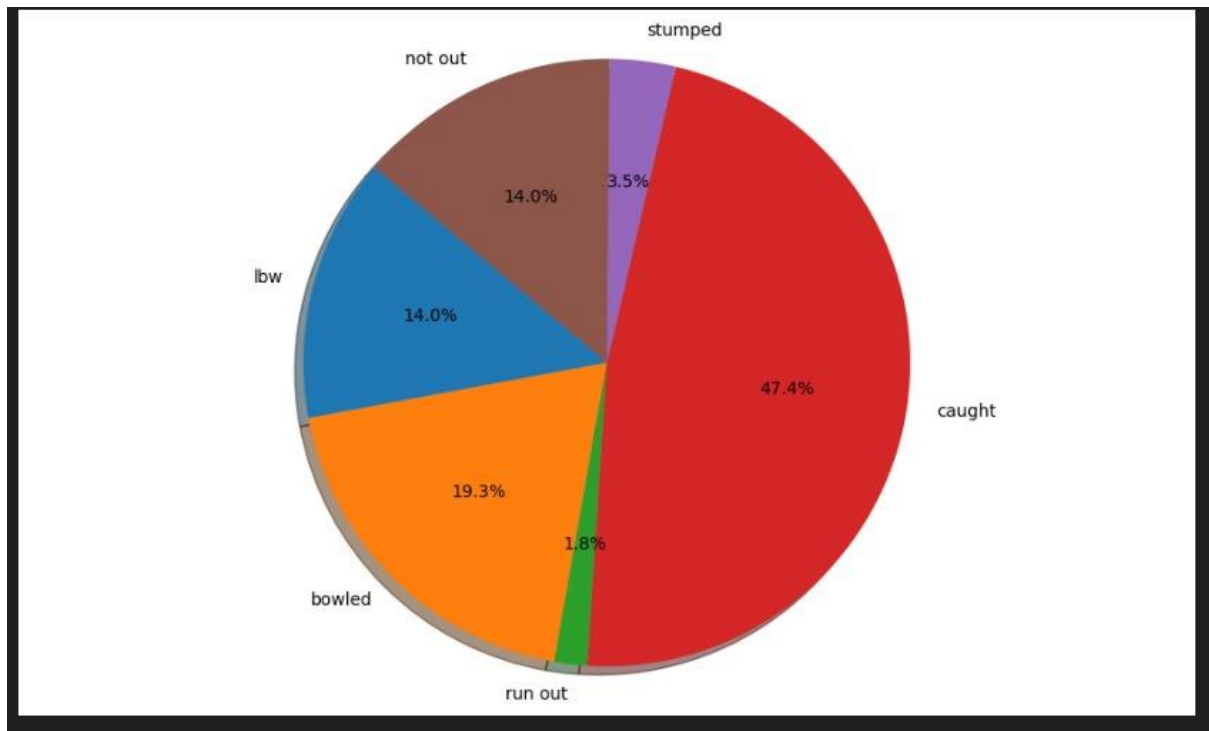
Feature Importance

Feature importance analysis in the Random Forest model identified key predictors such as strike rate and runs scored as crucial to accurate predictions.

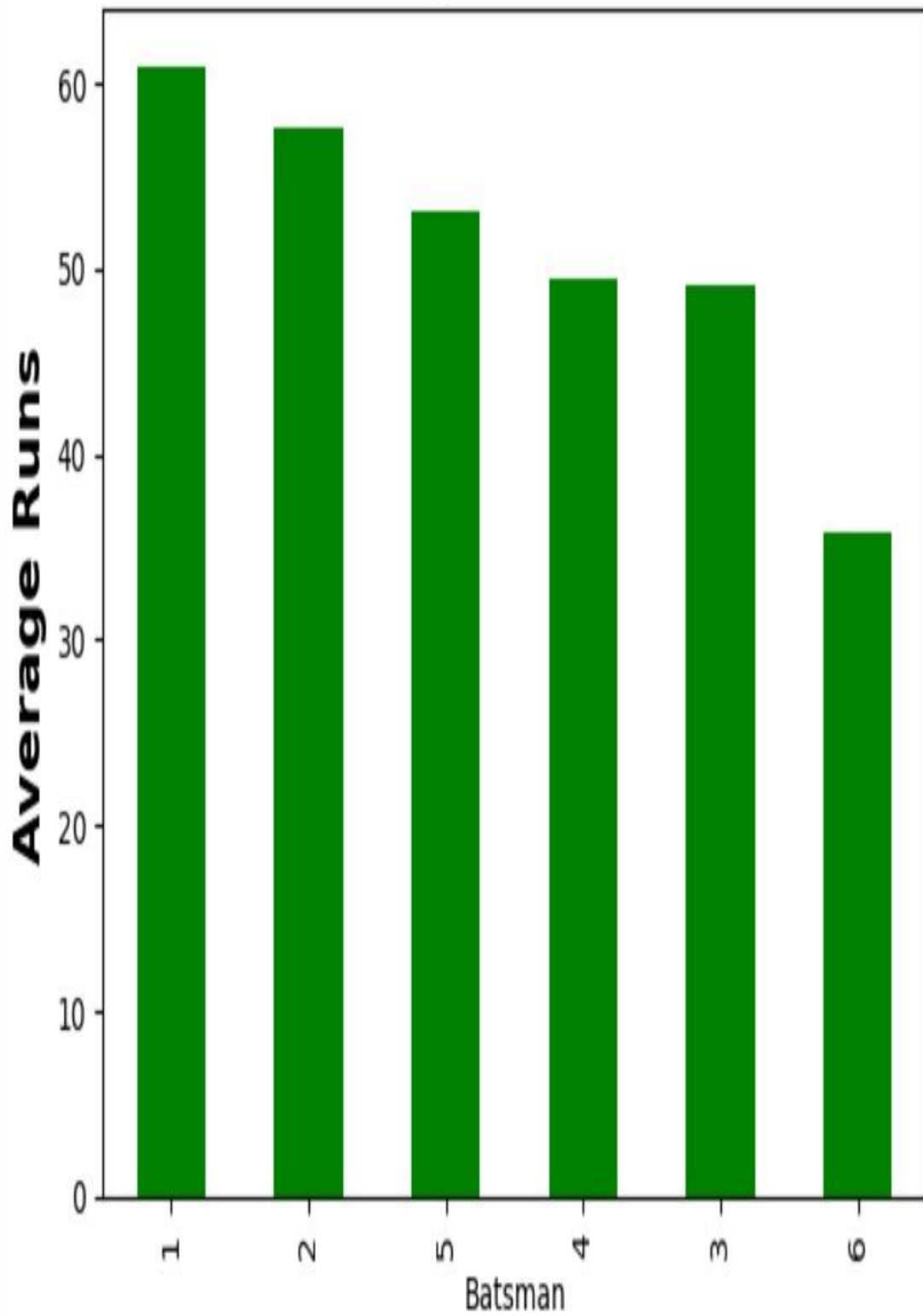
Error Analysis

Some discrepancies between predicted and actual scores were noted, especially in cases with unpredictable outcomes, suggesting areas for further enhancement.

Analysis



Average run of Batsman



Prediction

```
print(prediction)
```

```
[[ 64.72635009]
 [355.66071934]
 [107.22238155]
 [ 32.03709512]
 [ 48.3817226 ]
 [216.73138571]
 [ 17.32693038]
 [229.8070877 ]
 [ 53.28511085]
 [234.71047594]
 [170.96642875]
 [ 99.05006781]]
```

Discussion

The analysis reveals that key factors influencing cricket scores include runs scored, strike rate, balls faced, and dismissal methods. The Random Forest model performed better than the Linear Regression model, with an R^2 score of 0.82 compared to 0.74, indicating that Random Forest's ability to handle non-linear relationships between features made it more effective for predicting cricket scores.

Applications

This model can be beneficial for coaches and analysts by providing data-driven insights for strategic decision-making. It also enhances fan engagement by offering predictive insights during live matches, increasing the excitement and interaction with the game.

In summary, this project demonstrates the potential of machine learning to predict cricket scores accurately, providing valuable tools for both performance analysis and fan engagement.

Conclusion

This project highlights the effectiveness of machine learning in predicting cricket scores, using models like Linear Regression and Random Forest Regression. The Random Forest model demonstrated strong performance, offering valuable insights for both cricket strategy and fan engagement. By accurately predicting match outcomes, the project provides a useful tool for coaches and analysts to make informed decisions.

Moving forward, the model can be enhanced with real-time data and advanced techniques, paving the way for more accurate predictions and broader applications in sports analytics.

REFERENCES

Kaggle (2023). Cricket Dataset.

<https://www.kaggle.com/datasets>

Scikit-learn Developers (2023). Linear Regression. <https://scikit-learn.org>

Müller, Andreas C., and Guido, Sarah. Introduction to Machine Learning with Python.