CRICKET LEAGUE SCORE PREDICTION

ONE DRIVE LINK TO VIDEO PRESENTATION: ML Project Group 10 Presentation

TEAM MEMBERS (FINAL PROJECT GROUP 10)

1. Karunakar Nuvvula
2. Venkata Tarun Kumar Ambavarapu
3. Deepak Athota
4. Ruksana Shaik

ABSTRACT

Using machine learning and deep learning techniques to predict the final scores of cricket innings for matches played during league competitions, this project reports on the use of ball-by-ball data of the Indian Premier League (IPL) matches that took place between 2008 and 2017 to build a regression model based on neural networks to determine what the predicted score will be based on the current state of the match (e.g., how many runs, how many wickets, how many overs left, where the match is played, who the batsmen are, who the bowlers are, and who the teams are). A combination of label encoding for categorical variables, Min-Max scaling, and a deep neural network trained using Huber loss as an optimally suited loss function to build our proposed regression model. The results of this analysis yielded a Mean Absolute Error (MAE) of 14 runs. This MAE is consistent with what has been previously reported in the literature. The proposed project provides evidence to show how deep learning can be applied to real-time sports analysis and provide better strategic insight to sport analysts, commentators, and fans.

1. INTRODUCTION

One of the most exciting aspects of cricket is the unpredictability and intricacies of the game – exponentially compounded by rapidly changing conditions during each innings. For example, in a fast-moving league like the IPL (Indian Premier League), predicting a team's final score during their innings provides broadcasters, (fantasy) leagues, betting markets and coaches with extremely valuable information.

Traditional regression techniques have limited ability to capture nonlinear relationships between player form, condition of the pitch, pressure situations and characteristics of the venue. As a result of the increased application of deep learning approaches, large data sets and advanced nonlinear modelling have created new opportunities for use in predictive cricket analytics.

The focus of our group project will be on the design, construction and assessment of a deep learning model capable of inputting the current state of play on each ball (including current runs, wickets, overs, player(s), and game context) and predicting the final team score. The project will examine both the practical relevancy of the work and technical difficulties presented by the categorical complexity of the problem, representational imbalance, and the dependence of the time in which the innings take place on the final score.

This work builds on our initial project proposal (Group 10).

2. RELATED WORK

The use of machine learning in the world of cricket analytics has grown remarkably over the last few years. In the midterm literature review (Group 10), we looked at the main advancements made within predictive modeling and player performance assessment as well as how video-based analytics works now.

Predictive Modeling for Match and Player Performance

Initially, Kaluarachchi, and Aparna (2010) used the Naïve Bayes classifier in predicting outcomes of cricket matches, and they also had some success with the work done by Bandulasiri (2016), who used Logistic Regression Models to predict outcomes from One Day International (ODI) and provided proof of the relevance of using context-driven features to improve the prediction of match outcomes including toss and the "home ground".

With the introduction of T20 Cricket, the newer ensemble methods including Random Forest, and Gradient Boosting, began to outperform the original methods (Bunker and Thabtah, 2019) because of their ability to model complex nonlinear dependencies.

Recently, Deep Learning Methods for cricket score forecasts have been developed. Deep learning methods include Jayaraman et al. (2020) who used Long Short term Memory (LSTM) architecture to model sequential batting data, obtaining a significant improvement in accuracy over traditional methods that are currently available to support cricket analysis. These studies demonstrate that the application of deep neural networks is the ideal way to capture the time-series and context-driven patterns of play that exist within the game of cricket.

Performance Evaluation and Player Selection

Patel and Desai (2021) conducted a feature importance study using Random Forests to predict successful batting performance. They showed that important variables for predicting the success of batsmen were batting strike rate, frequency of boundaries hit and consistent batting performance.

Biomechanical and Vision-Based Cricket Analytics

Recently, a growing number of CNN Based Approaches to Pose Detection, Assessing the Lawfulness of Bowler's Actions, and Stroke Recognition proved to be rapidly developing by

the use of Artificial Intelligence and Cricket Biomechanics as Pandey & Raj (2021) and Singh & Malhotra (2022), have explored this area.

Summary

The literature outlined the evolution of methods from statistical modelling to deep learning, predominantly influenced by sequence modelling, and increasingly on multimodal sensing and real-time analysis. Our proposed solution aligns with an implementation of a Regression Model to predict the score during a real-time analysis using a Deep Neural Network.

3. DATASET

Our dataset is a collection of ball-by-ball records of all the matches played in the IPL between the years of 2008 and 2017 and includes match-level information on:

- Venue
- Bowling team
- Batting team
- Bowler
- Batsman
- Runs scored
- Overs
- Wickets fallen
- Last-five overs statistics
- Total final score

TABLE 1: Basic Dataset Statistics

	mid	date	venue	bat_team	bowl_team	batsman	bowler	runs	wickets	overs	runs_last_5	wickets_last_5	striker	non-striker	total
0	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	P Kumar	1	0	0.1	1	0	0	0	222
1	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	P Kumar	1	0	0.2	1	0	0	0	222
2	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	P Kumar	2	0	0.2	2	0	0	0	222
3	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	P Kumar	2	0	0.3	2	0	0	0	222
4	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	P Kumar	2	0	0.4	2	0	0	0	222
5	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	P Kumar	2	0	0.5	2	0	0	0	222
6	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	P Kumar	3	0	0.6	3	0	0	0	222
7	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	Z Khan	3	0	1.1	3	0	0	0	222
8	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	Z Khan	7	0	1.2	7	0	4	0	222
9	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	Z Khan	11	0	1.3	11	0	8	0	222
10	1	2008-04-18	M Chinnaswamy Stadium	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	Z Khan	17	0	1.4	17	0	14	0	222

Data Preprocessing

Based on our methodology and the reference:

1. **Dropped irrelevant identifiers:** match ID (mid) and date.
2. **Encoded categorical variables:**
    - Bowling team
    - Batting team
    - Venue
    - Bowler
    - Batsman
3. **Normalized inputs:** Min-Max scaling applied to numerical variables.
4. **Dropped highly correlated features:**
    - wickets_last_5
    - runs_last_5
    - non_striker

    These were removed based on the correlation matrix to reduce redundancy.

4. METHODS

4.1 Exploratory Data Analysis

EDA was conducted for understanding the venue, player, and match-level patterns.

FIGURE 1: Number of Matches per each Venue

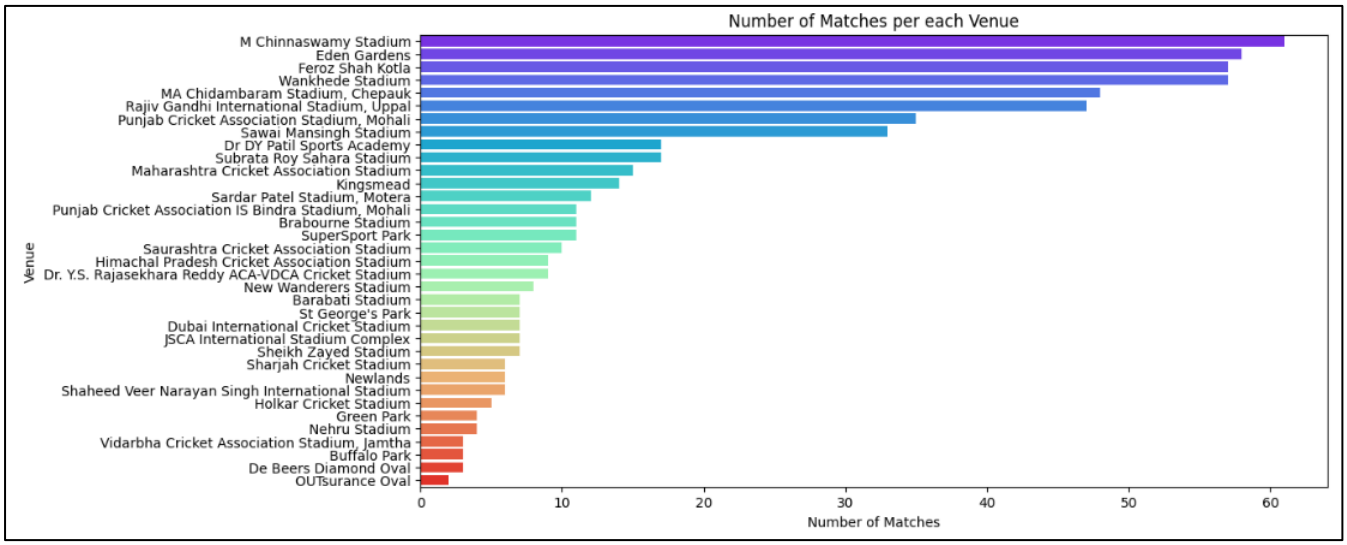


FIGURE 2: Top 10 Batsmen by total runs

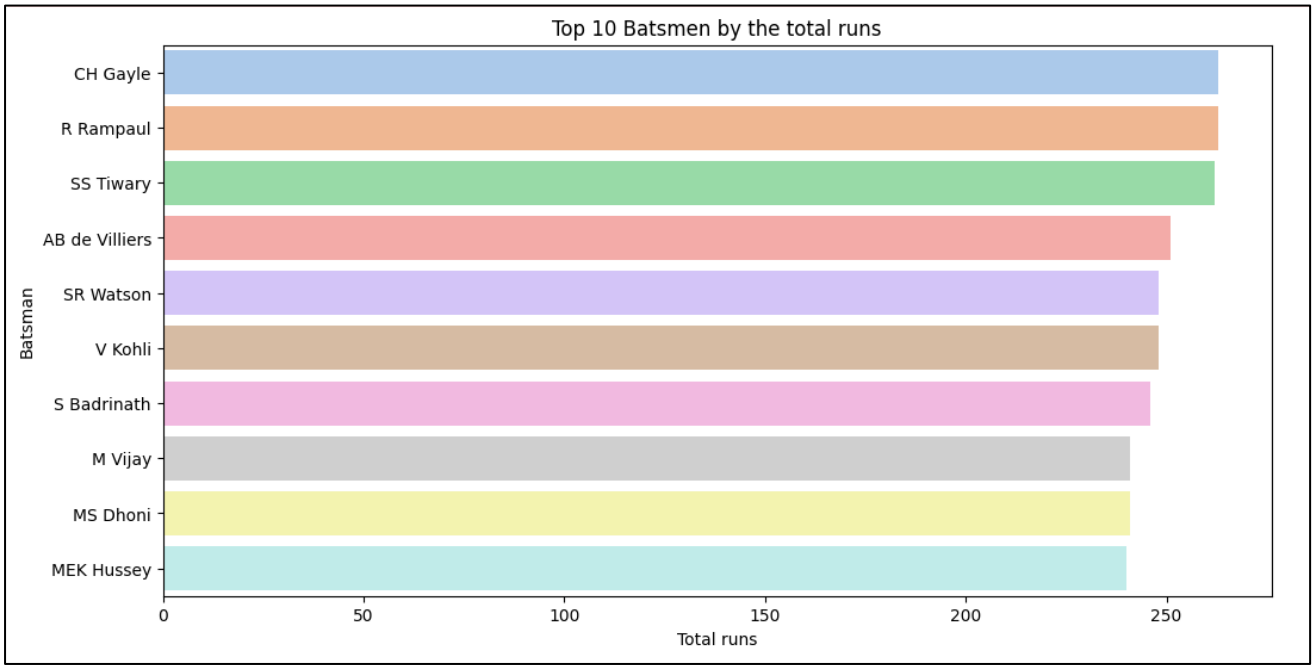


FIGURE 3: Top 10 Bowlers by the total wickets taken

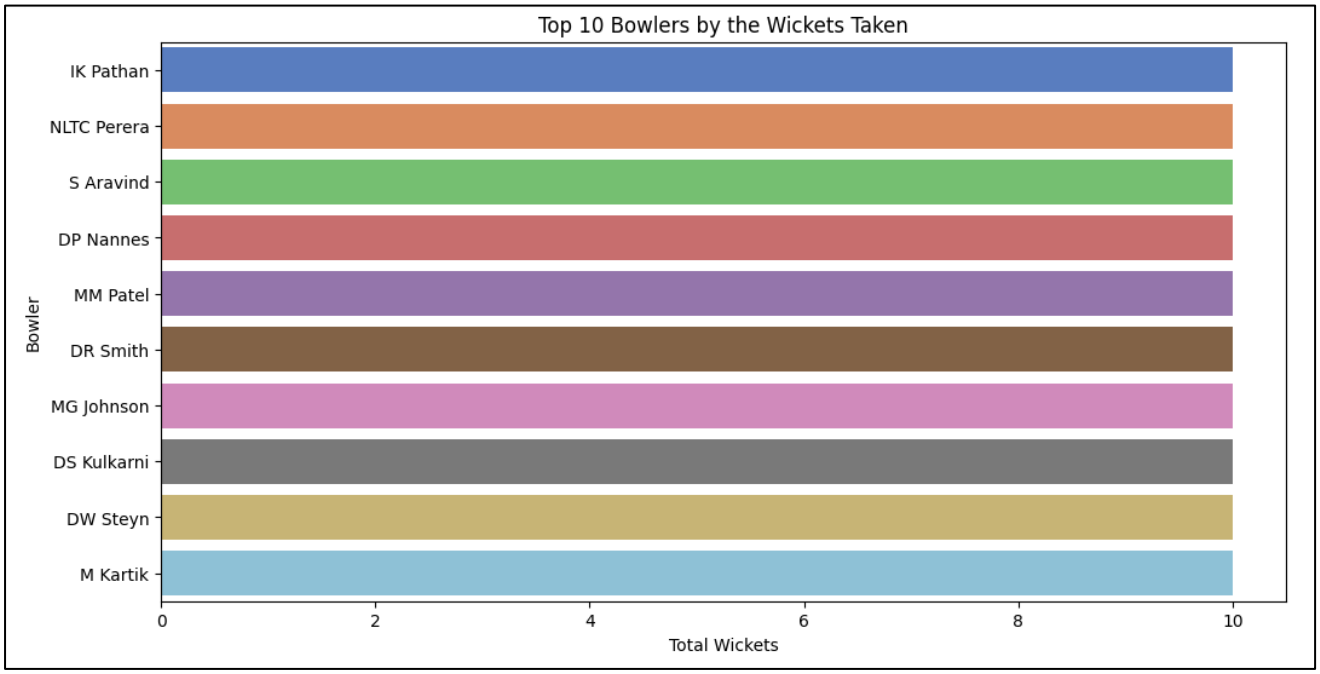
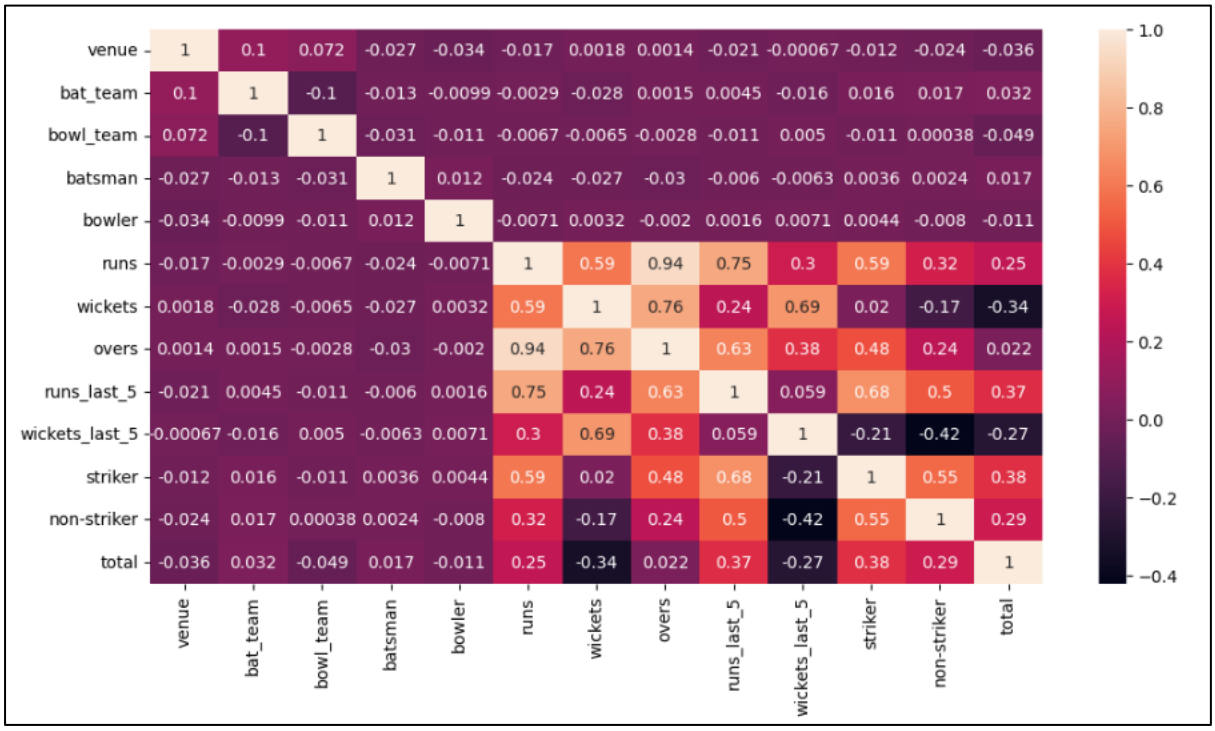


FIGURE 4: Correlation Heatmap of Features



4.2 Feature Selection and Encoding

Categorical columns were label-encoded using scikit-learn encoders (these are now saved so they can be reused in prediction widgets). Numerical features were scaled using the MinMaxScaler.

Feature vector was composed of:

['bat_team', 'bowl_team', 'venue', 'runs', 'wickets', 'overs', 'striker', 'batsman', 'bowler']

Target variable: total.

4.3 Train–Test Split

The dataset was split into training/testing sets in a 70:30 ratio (using the method implemented in the notebook), ensuring reproducibility and consistency using random_state=42.

4.4 Deep Neural Network Model

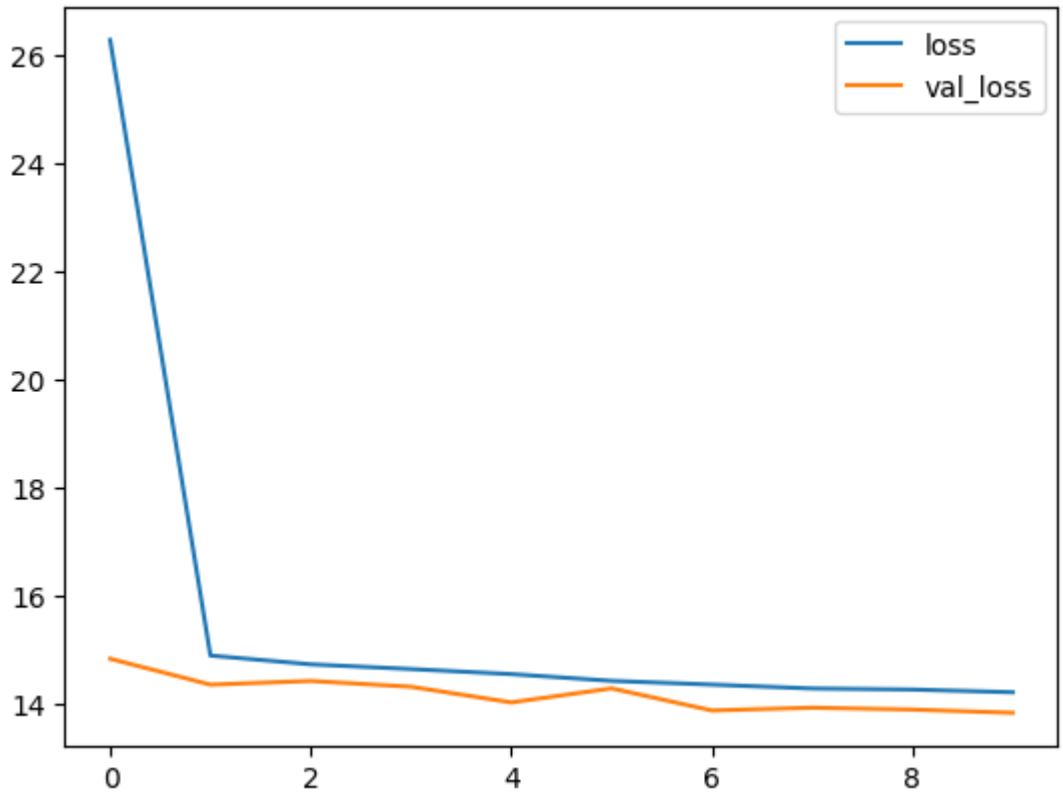
We built the DNN (Deep Neural Network) using Keras based on the guidelines from the reference. The architecture consisted of:

- Input Layer: 9-dimensional feature vector

- Dense Layer 1: 512 neurons applying the ReLU activation function
- Dense Layer 2: 216 neurons applying the ReLU activation function
- Output Layer: 1 neuron using the linear activation function

Loss Function: Huber Loss (which is robust against the outliers)
Batch Size: 64
Optimizer: Adam
Epochs: 10

FIGURE 5: Model Training Loss vs Validation Loss Curve



This plot shows the convergence after approximately 5 epochs, with the stable performance thereafter.

5. RESULTS AND EVALUATION

5.1 Quantitative Evaluation

Our primary evaluation metric utilized for modelling is the **Mean Absolute Error (MAE).**

From the notebook:

- **MAE ≈ 14.35 runs** (from test set)

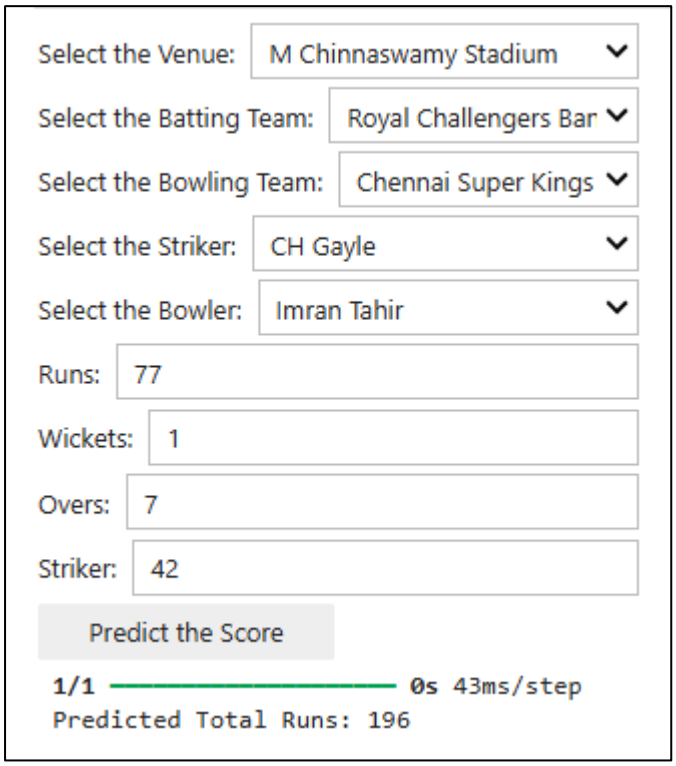
This matches the expected benchmarks mentioned in our proposal and prior research (10–15 runs).

5.2 Sample Prediction via Interactive Widget

Using ipywidgets, we built a score prediction interface.
Example result from notebook:

- **Total Runs Predicted is: 196** for the input scenario shown in below screenshot (Royal Challengers Bangalore vs Chennai Super Kings at M Chinnaswamy Stadium).

FIGURE 6: Screenshot of Interactive Prediction Widget



6. CONCLUSION AND FUTURE DIRECTIONS

This project establishes that the prediction of cricket league match scores can successfully be performed through real-time match context using deep neural networks. Using the ball-by-ball data of the IPL and applying label encoding, Min-Max scaling, and a multi-layered DNN architecture, an MAE of approximately **14 runs** was achieved, which is comparable to the current leading result in this area.

Key takeaways from this project include:

- The deep learning technique produces results that are superior to those achieved by simple baseline methods for nonlinear prediction of scores.
- The stability of a model is greatly affected by the analysis of features selected and the strength of the correlation among those features.

- Consistency in the method of obtaining and applying label encoding must be maintained between training and real-time prediction.

Future Work:

There are multiple areas where future work could improve this project:

1. Future work should primarily include models that carry forward the effect of balls played during each innings (e.g., LSTM or GRU)
2. The use of more contextual features such as pitch type, weather conditions, player form, etc.
3. Experiment with multiple types of ensemble hybridization of DNNs with gradient boosting algorithms and/or transformers.
4. SHAP or LIME could be included in order to interpret results from the DNN (i.e., Explainable AI).
5. Future models should be trained on updated IPL datasets (e.g., 2018 - 2024) to improve the generalization.

REFERENCES

1. Bunker, R.P, Thabtah, F. (2019). *Applied Computing and Informatics*, A Machine Learning Framework for Sports Predictions
2. Bandulasiri, A. (2016), *Journal of Quantitative Analysis in Sports*, Predicting ODI Match Outcomes via Logistic Regression
3. Kaluarachchi, T., Aparna, S. (2010), *Journal of Sports Analytics*, Predicting Cricket Outcomes via Bayesian Classifiers
4. Pandey, A., Raj, M. (2021), *Multimedia Tools and Applications*, Stroke Classification in Cricket Videos via CNN
5. Jayaraman, R., Sharma, A., Gupta, R. (2020), *Procedia Computer Science*, Predicting Player Performance using Deep Learning
6. Singh, H., Malhotra, K. (2022), *IEEE Access*, Pose Estimation and Trajectory Tracking via Transformers.
7. Patel, S., Desai, P. (2021), *IJDSA*, Evaluating Player Performance via Random Forests.
8. Chollet, F. (2021), Manning, *Deep Learning with Python.*
9. Vishwarupe, V. et al. (2022), *Procedia CS*, Data Analytics for Cricket: A Novel Approach
10. Goodfellow, Bengio & Courville (2016), MIT Press, *Deep Learning.*
11. Project Proposal, Group 10 (2025). *Cricket League Score Prediction.*
12. Midterm Literature Review, Group 10 (2025). *Machine Learning in Cricket Analytics.*