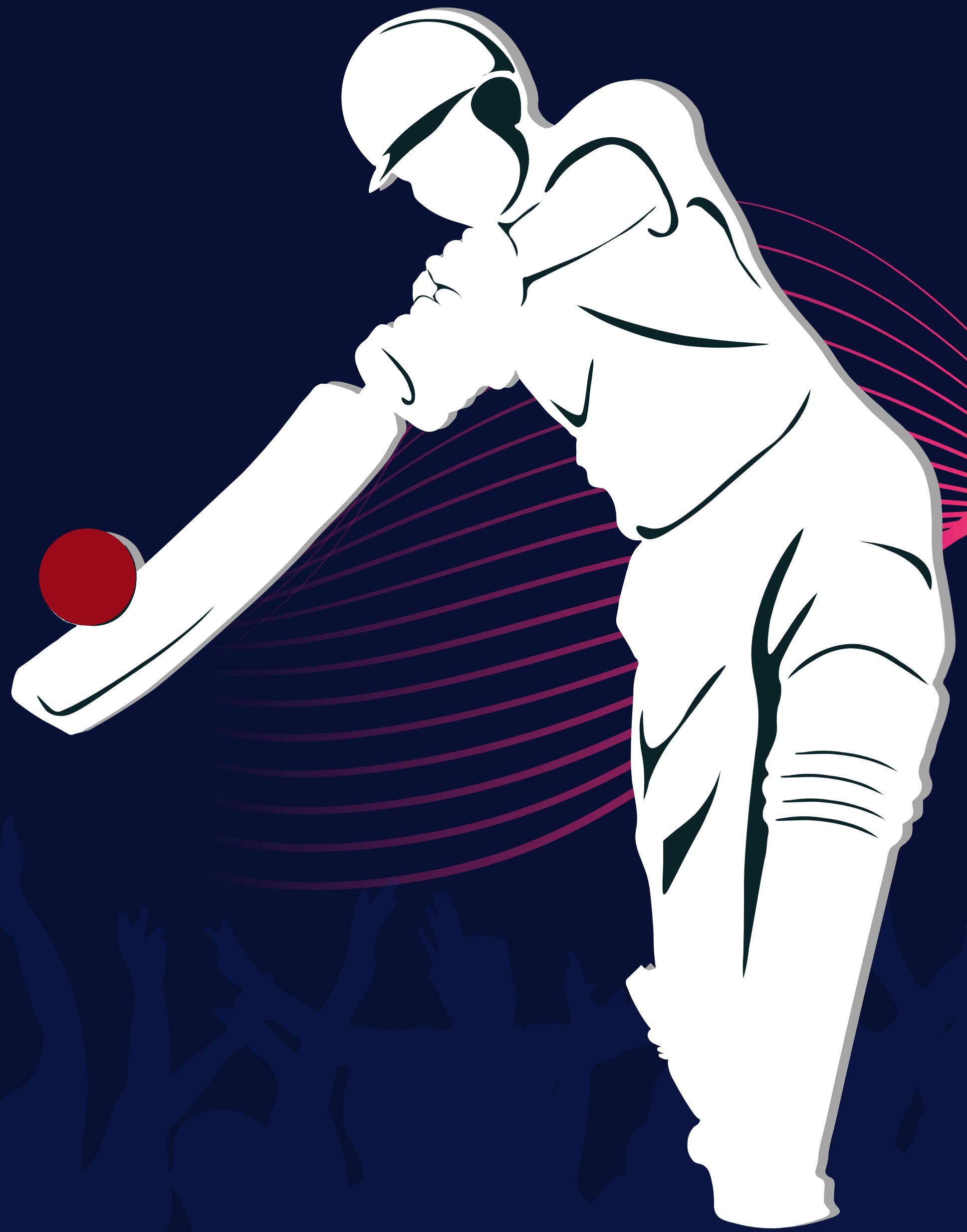
Cricket Final Score Prediction

Arnav Kapoor, Avantika Bansal, Suraj Dayma



Problem Statement

Cricket Score Prediction

CricInfo Predictor

Blackbox

Bad Predictions

Cricket is Complex !

Problem Statement

Context Aware Cricket Score Prediction
in ODI Matches

+ Weather
+ Pitch
+ Historical Playoffs
+ Bowler Batter Matchups

Problem Statement

Context Aware Cricket Score Prediction

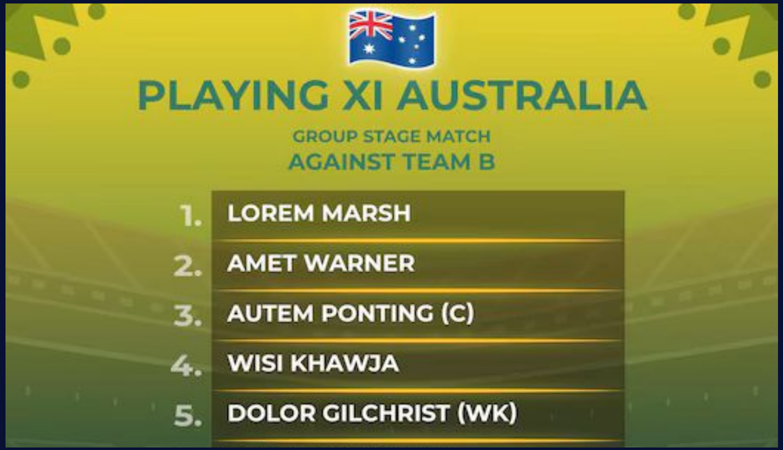
in ODI Matches

Broadcasts

Betting

APPLICATIONS

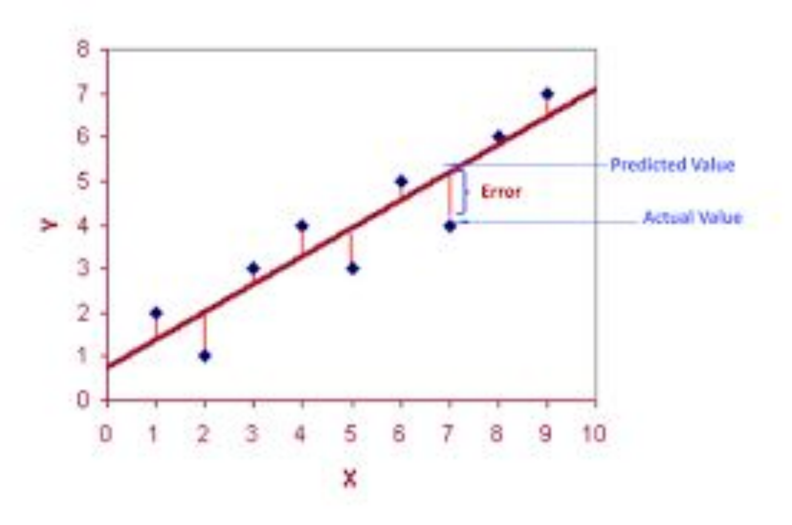
Accessible to Smaller Teams



Deciding Lineups

Performance Metric

- Used RMSE (Root Mean Squared Error) as the performance metric
- Measured average of errors between actual and predicted score
- On average how many runs off was the model



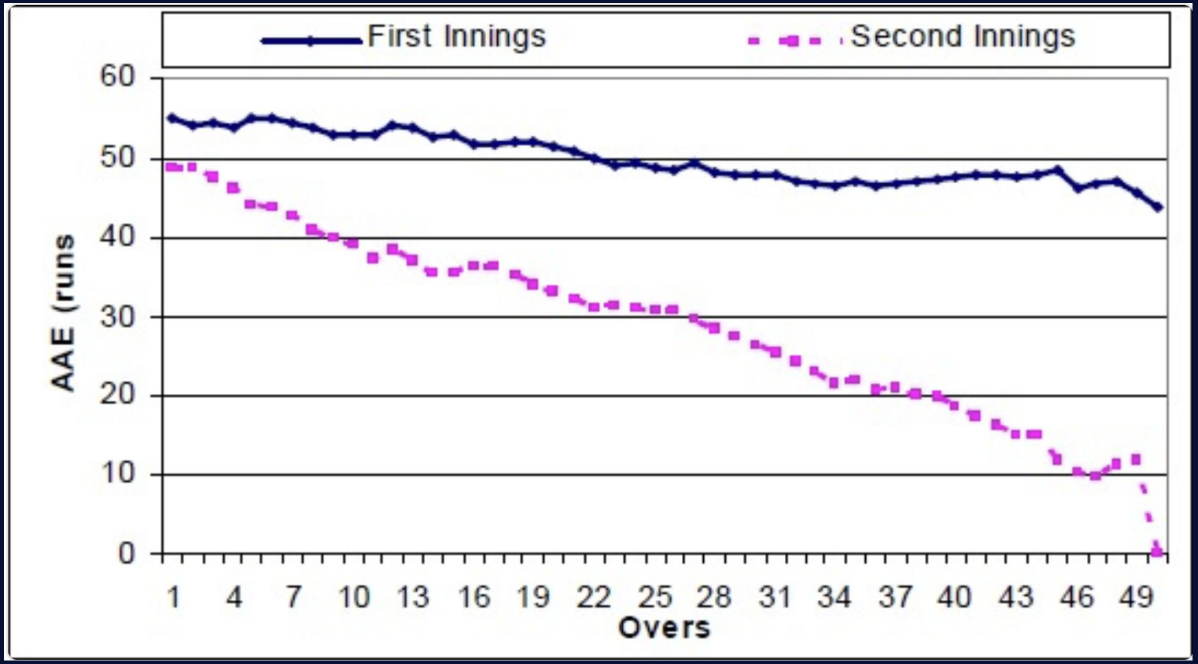
$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}$$

Literature Review

Cricket Score Prediction

Aim: Predict match score for betting.

- Linear regression with DLS method.
- 35 years of ODI matches in dataset.
- Also used pre-match features that included home team, toss winner and historical team strength.



Model: Linear Regression

Inference: Initial use of ML.

Bailey, M., & Clarke, S. R. (2006). Predicting the Match Outcome in One Day International Cricket Matches, while the Game is in Progress. *Journal of sports science & medicine*, *5*(4), 480–487.

Literature Review

Cricket Score Prediction

Aim: Predict match score and outcome.

- Used features of current score, over, target, venue, wickets fallen.
- Did not provide standard performance metrics.

Model: Linear Regression, Cricinfo Data.

Inference: Cannot be used as benchmark

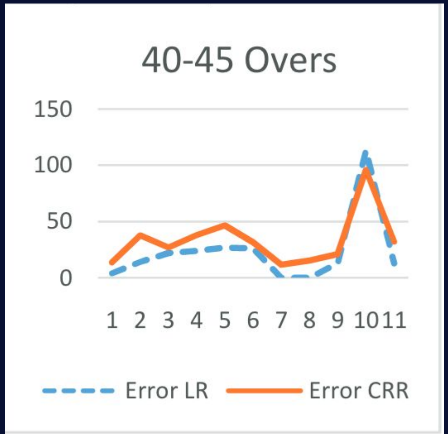


Fig. 1.4. Accuracy of 40-45 overs.

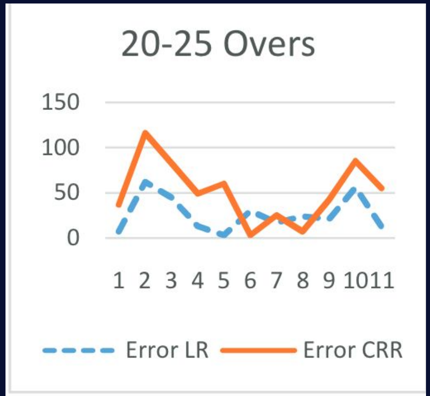


Fig. 1.2. Accuracy of 20-25 overs.

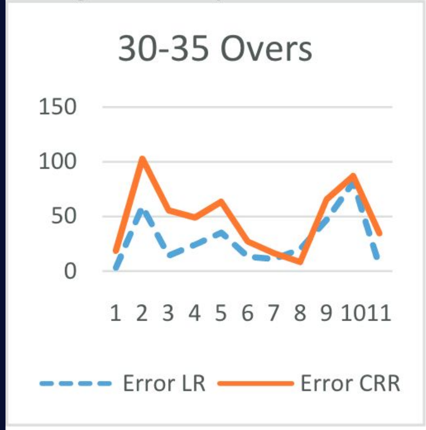


Fig. 1.3. Accuracy of 30-35 overs.

Singh, T., Singla, V., & Bhatia, P. (2015). Score and winning prediction in cricket through data mining. *2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI)*, 60–66. https://doi.org/10.1109/ICSCTI.2015.7489605

Literature Review

Cricket Analysis

Aim: Survey of papers in the thematic of cricket analysis from 2001 to 2021.

- 59 papers were screened.
- Only 5% of the papers focused on score prediction.

Model: N/A

Inference: Limited research in cricket analysis.

Table 4
Used ML techniques.

ML technique	Applications	% of studies
SVM	Umpire's gesture modeling Fast bowler's action modeling Game's outcome prediction Player-performances classification	45
RF	Predicting the winner of the tournament Predicting the winner of the game Classification of all-rounders	42
NB	Predicting the winner of the game Prediction the Batting and bowling performance Pitch behavior prediction	36
Regression	Identification of game's influential factors Innings' score prediction	31
DT	Identification of game's influential factors Winner prediction Batting and bowling performance prediction	26
NN	Game outcome prediction Cricket-shot classification Game's score prediction	21
kNN	Classification of all-rounders Player-performances classification	17
Other	Predicting the player-ranking Cricket and social media	14
XGBoot	Bowler's workload prediction	5
K-Means	Batsmen classification	3

Wickramasinghe, I. (2022). Applications of Machine Learning in cricket: A systematic review. *Machine Learning with Applications*, *10*, 100435.
https://doi.org/10.1016/j.mlwa.2022.100435

Literature Review

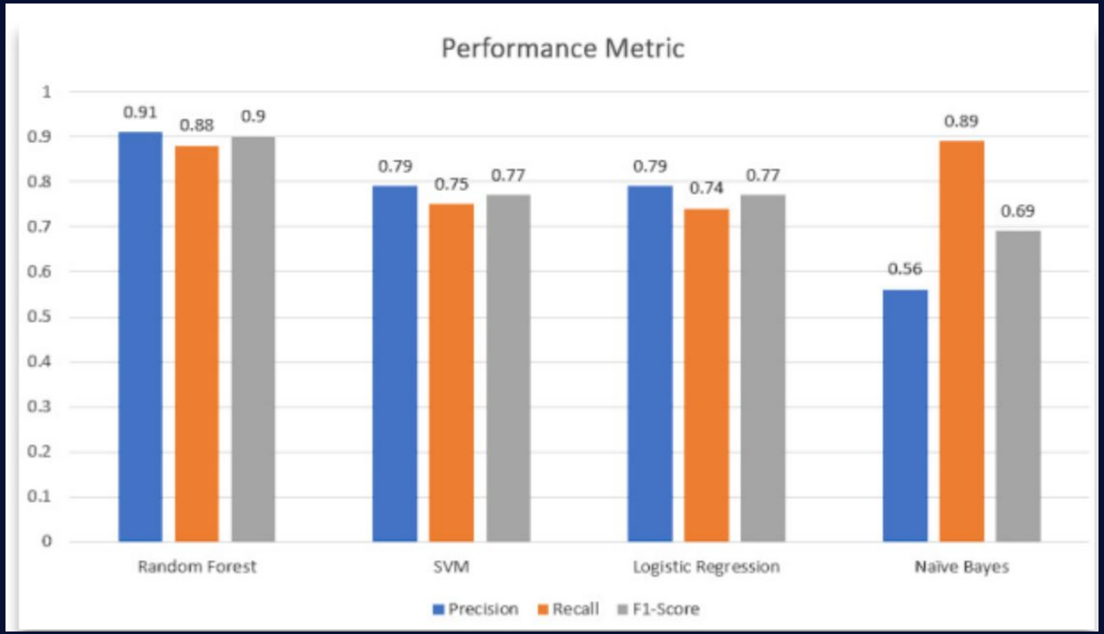
Cricket Score Prediction

Aim: Predict match score and outcome.

- Linear regression with weighted sum of features.
- How did they get the weights? Not Mentioned.
- Score prediction metrics? Not Mentioned.
- Created ‘player consistency’ metric.

Model: Linear Regression, cricsheets data.

Inference: Cannot be used as benchmark



Match Outcome Prediction

Dalal, P., Shah, H., Kanjariya, T., & Joshi, D. (2024). *Cricket match analytics and prediction using machine learning. International Journal of Computer Applications, 186*(26), 27–33. https://doi.org/10.5120/ijca2024923744

Literature Review

Cricket Score Prediction

Aim: Predict match score.

- Random Forest was giving overfit model.
- Got 26.27 RMSE with Linear Regression.

Model: Linear Regression

Inference: Benchmark of RMSE 26.27

Hossain, Md. S., Rumpa, U. K., Hossain, Md. S., Shovon, S. I., Huque, Md. T., & Hasan, M. (2024). One Day International Cricket Match Score Prediction Using Machine Learning Approaches. *2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE)*, 1–6. https://doi.org/10.1109/ICSPCRE62303.2024.10674897

Literature Review

Cricket Score Prediction

Model	RMSE
XGBoost	5.58

Satwani, A., Coutinho, K., John, N., Arul Jothi, J.A.: Live cricket predictions for runs and win using machine learning. In: 2024 IEEE 12th International Conference on Intelligent Systems (IS). pp. 1–6 (2024). https://doi.org/10.1109/IS61756.2024.10705216

Model	RMSE
Random Forest	Close to 0
Linear Regression	Just over 6

Suguna, R., Kumar, Y., Prakash, J., P S, D.N., Kiran, S.: Utilizing machine learning for sport data analytics in cricket: Score prediction and player categorization. pp. 1–6 (2023). https://doi.org/10.1109/MysuruCon59703.2023.10396955

Model	RMSE
Random Forest	5.95

Rahman Mahin, M.P., Ara, E., Islam, R.U.: Cricket analytics: Odi match score prediction and resource metric modeling using machine learning model. In: 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS). pp. 1–6 (2024). https://doi.org/10.1109/COMPAS60761.2024.10796074

Literature Review

Cricket Score Prediction

Our Aim: Replicate paper.

- Features: ball-by-ball features + weight related features.
- Tested on 2025 matches => RMSE of over 50.

Model: Random Forest.

Inference: Overfit Model; Cannot be used as benchmark

TABLE III
TRAINING RESULT

Model vs Metrics	MAE	MSE	RMSE	R^2
Random Forest Regression	**0.7270**	**4.8551**	**2.2034**	**0.9986**
XGBoost Regression	2.9267	21.3623	4.6219	0.9939
Linear Regression	26.7671	1278.2333	35.7523	0.6409
Elastic net Regression	28.2933	1465.1789	38.2776	0.5884
Polynomial Regression	8.8432	181.7004	13.4796	0.9446

TABLE IV
TESTING RESULT

Model vs Metrics	MAE	MSE	RMSE	R^2
Random Forest Regression	**1.9482**	**35.5167**	**5.9595**	**0.9899**
XGBoost Regression	4.1746	50.5891	7.1406	0.9855
Linear Regression	27.0091	1301.5886	36.0775	0.6309
Elastic net Regression	28.4408	1485.0190	38.5359	0.5789
Polynomial Regression	9.2147	194.8618	13.9592	0.9387

TABLE II
WEIGHT RELATED FEATURES

remaining_overs	Overs remaining
remaining_wickets	Wickets remaining
weight_overs	Over remaining/49.6
weight_wicket	Wickets remaining/10
merge_weight	(weight_overs × remaining_overs)+ (weight_wicket × remaining_wickets)

Rahman Mahin, M.P., Ara, E., Islam, R.U.: Cricket analytics: Odi match score prediction and resource metric modeling using machine learning model. In: 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS). pp. 1–6 (2024). https://doi.org/10.1109/COMPAS60761.2024.10796074

Literature Review

Cricket Outcome Prediction

Aim: Predict match outcome at intervals in the game.

Model: LSTM

- Uses features like batsman, bowler, runs, extras, wickets, innings number, over, and ball number.
- Using an LSTM, they took ball by ball data and calculated win probabilities after every ball.

Inference: LSTMs can be used for cricket analysis.

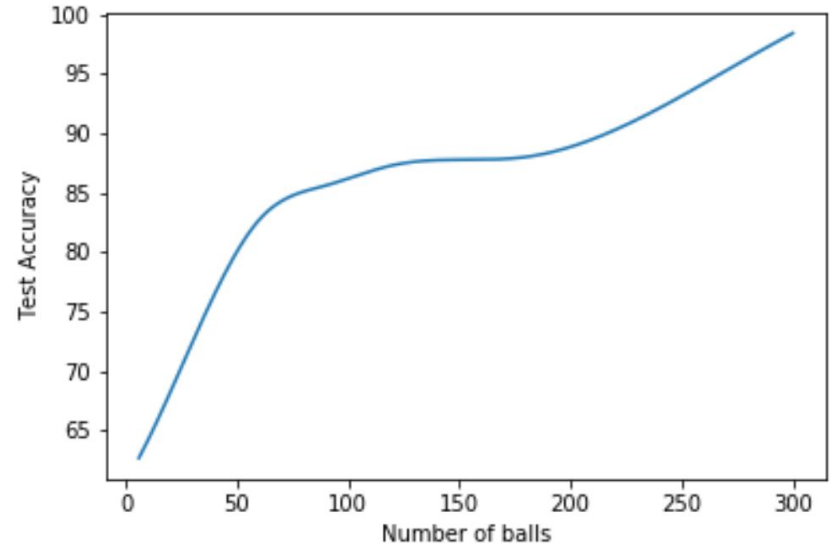


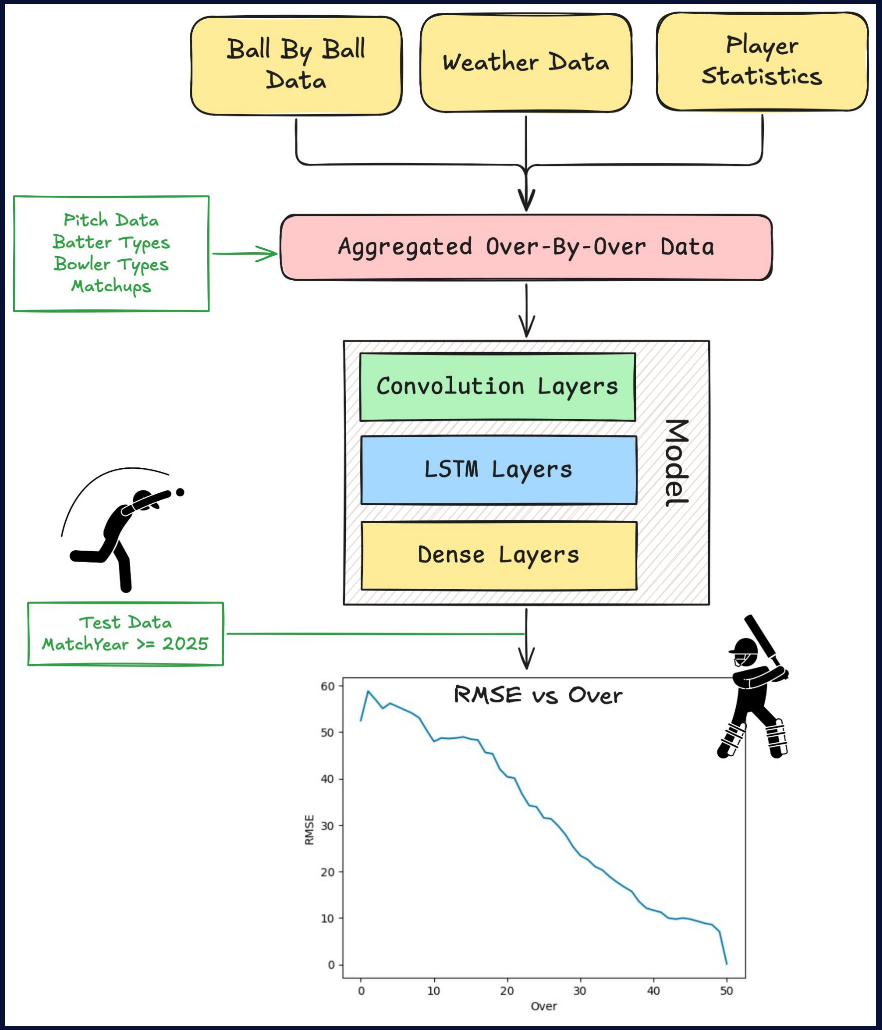
Fig. 3. Test Accuracy variation of our best model after the number of balls played in the match.

Chakwate, R., & A, M. R. (2020). Analysing Long Short Term Memory Models for Cricket Match Outcome Prediction. *International Journal for Research in Applied Science and Engineering Technology*, *8*(11), 1–8. https://doi.org/10.22214/ijraset.2020.28203

Literature Review

Gap and Solution

- No neural network based models used for final score prediction.
- Features like weather, or pitch conditions not taken into account.
- Overfit / Non generalized models.
- Women-specific models.



Our Dataset

Cricsheet

OpenMeteo

Cricinfo

Engineered

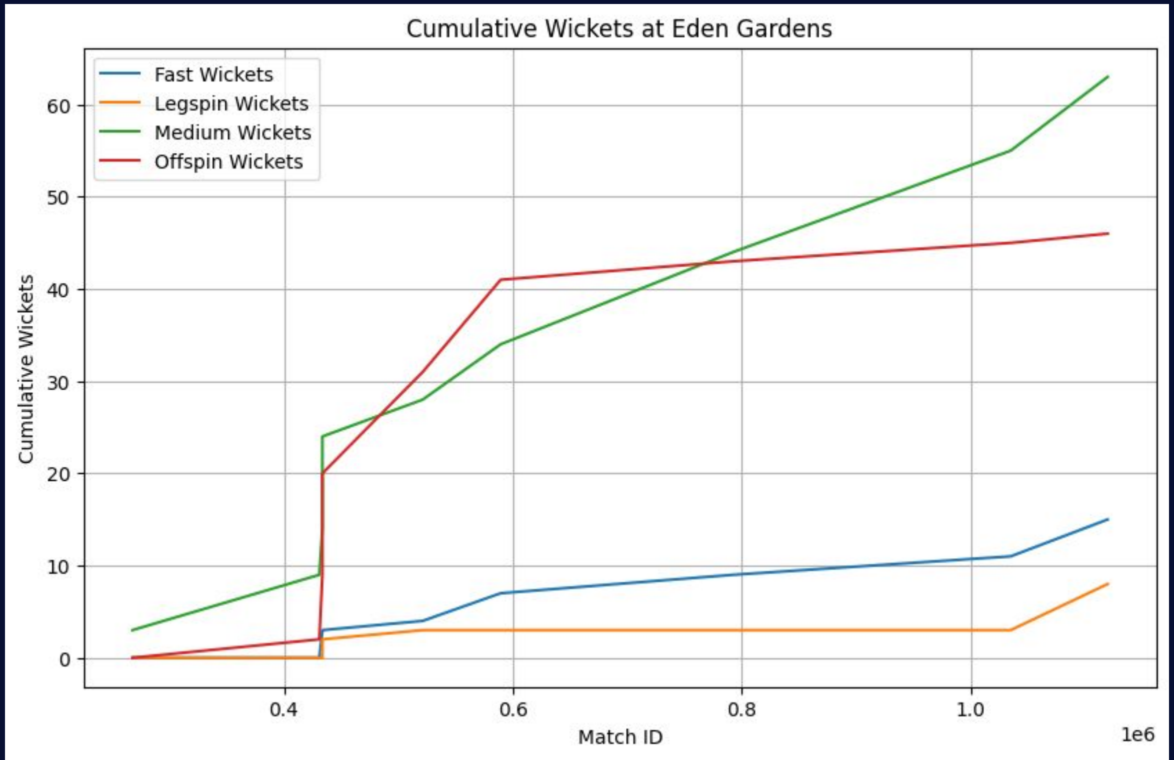
Table 1: List of Features Used in the Model

Category	Features
Ball-by-ball Data	match_id, innings, over, ball, runs_batter, runs_extras, runs_total, batter_total_runs, batter_balls_faced, bowler_total_runs, bowler_balls_bowled, team_total_runs, wickets_taken, rr, remaining_balls
Team Stats	win_rate_last_5, avg_runs_last_5, win_rate_last_5_bowl, avg_wkts_last_5
Weather Data	temperature_2m, relative_humidity_2m, cloud_cover, wind_speed_10m, dew_point_2m
Pitch Stats	fast_wickets, legspin_wickets, medium_wickets, offspin_wickets, slow_wickets, unknown_wickets, runs_left_hand, runs_right_hand, sum_fast_wickets, recent_fast_wickets, sum_legspin_wickets, recent_legspin_wickets, sum_medium_wickets, recent_medium_wickets, sum_offspin_wickets, recent_offspin_wickets, sum_slow_wickets, recent_slow_wickets, sum_unknown_wickets, recent_unknown_wickets, sum_runs_left_hand, recent_runs_left_hand, sum_runs_right_hand, recent_runs_right_hand, year
Individual Batter Stats	bat_team_player_1_runs, bat_team_player_1_balls, bat_team_player_1_dismissals, bat_team_player_1_avg, bat_team_player_1_sr, bat_team_player_1_runs_vs_spin, bat_team_player_1_balls_vs_spin, bat_team_player_1_dismissals_vs_spin, bat_team_player_1_runs_vs_pace, bat_team_player_1_balls_vs_pace, bat_team_player_1_dismissals_vs_pace, bat_team_player_1_avg_vs_spin, bat_team_player_1_sr_vs_spin, bat_team_player_1_avg_vs_pace, bat_team_player_1_sr_vs_pace, bat_team_player_1_is_left, bat_team_player_1_is_right
Individual Bowler Stats	bowl_team_top_bowler_1_runs_conceded, bowl_team_top_bowler_1_balls_bowled, bowl_team_top_bowler_1_wickets, bowl_team_top_bowler_1_bowling_avg, bowl_team_top_bowler_1_bowling_sr, bowl_team_top_bowler_1_economy, bowl_team_top_bowler_1_runs_vs_right, bowl_team_top_bowler_1_balls_vs_right, bowl_team_top_bowler_1_wickets_vs_right, bowl_team_top_bowler_1_runs_vs_left, bowl_team_top_bowler_1_balls_vs_left, bowl_team_top_bowler_1_wickets_vs_left, bowl_team_top_bowler_1_bowling_avg_vs_right, bowl_team_top_bowler_1_bowling_sr_vs_right, bowl_team_top_bowler_1_economy_vs_right, bowl_team_top_bowler_1_bowling_avg_vs_left, bowl_team_top_bowler_1_bowling_sr_vs_left, bowl_team_top_bowler_1_economy_vs_left, bowl_team_top_bowler_1_is_spin, bowl_team_top_bowler_1_is_pace
Prediction Targets	innings_total_score, over_total_score, next_over_total_score

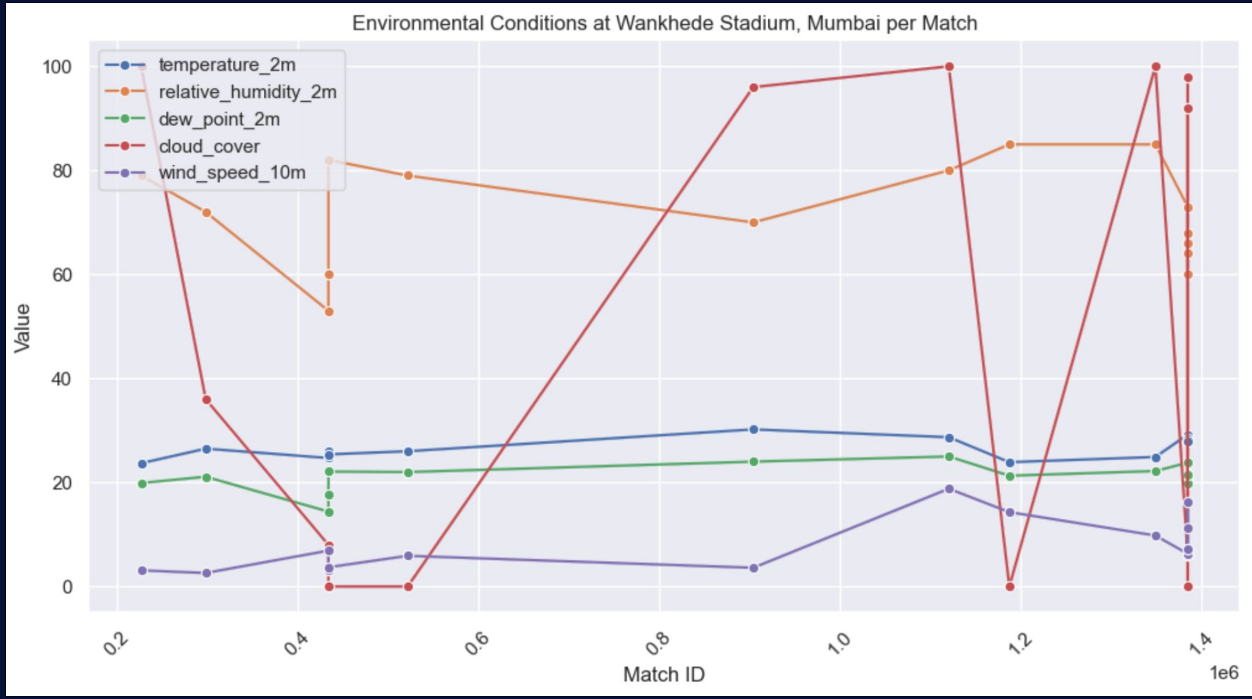
Data Processing Pipeline

- 65 features for ODI matches from 2005 - 2025 for top 9 teams
- Ball by Ball data from cricksheets in JSON form, converted into CSV format
- Extracted data for game state like number of balls, runs and wickets, run rate, required run rate
- Collected or derived pitch, weather and team rolling stats
- Web scraped data from espn cricinfo for player details like their batting and bowling style
- Combined into a final CSV file

Data Processing Pipeline



Pitch Data (Bowling)

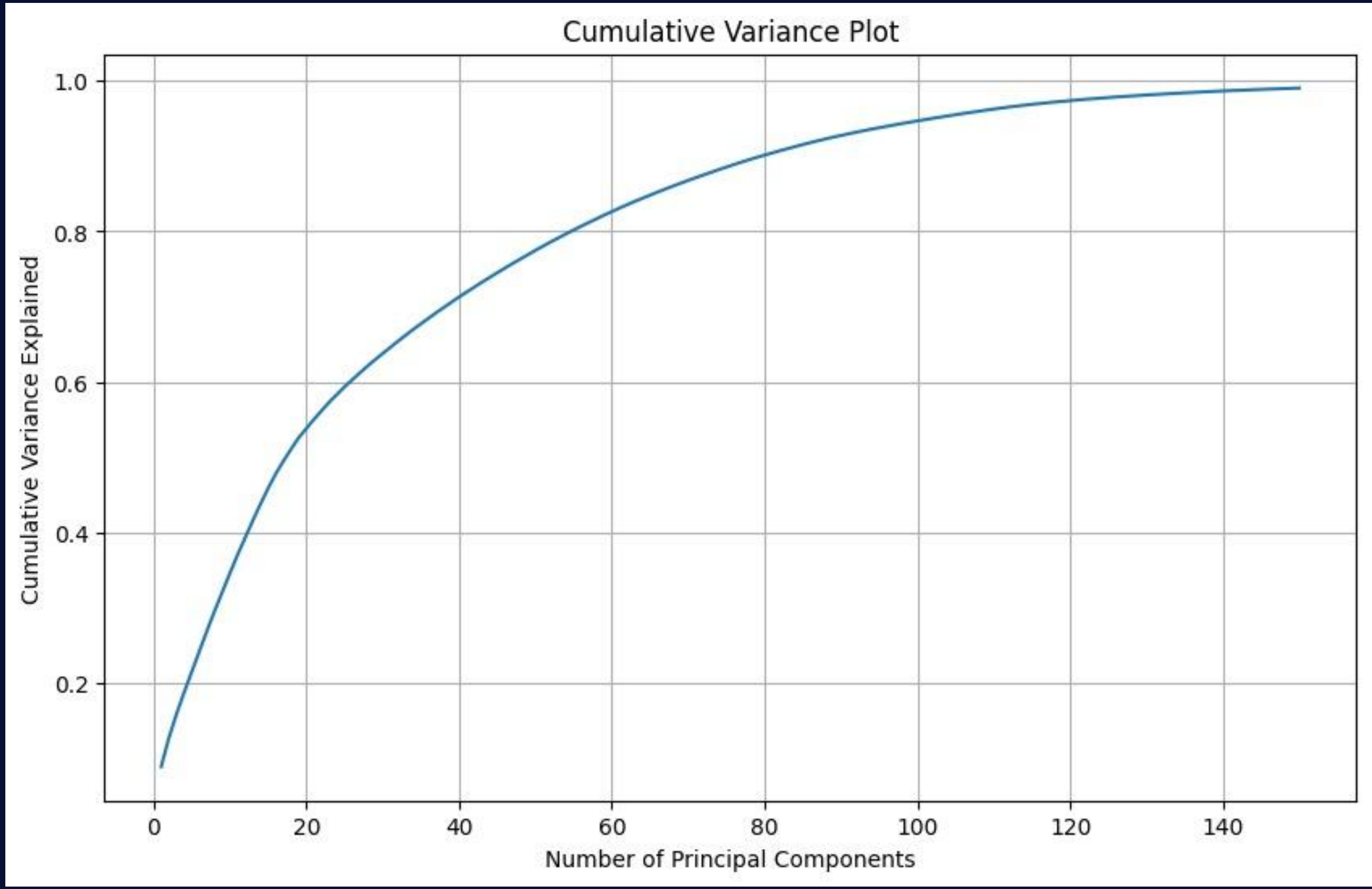


Weather Data

Methodology

Feature Selection

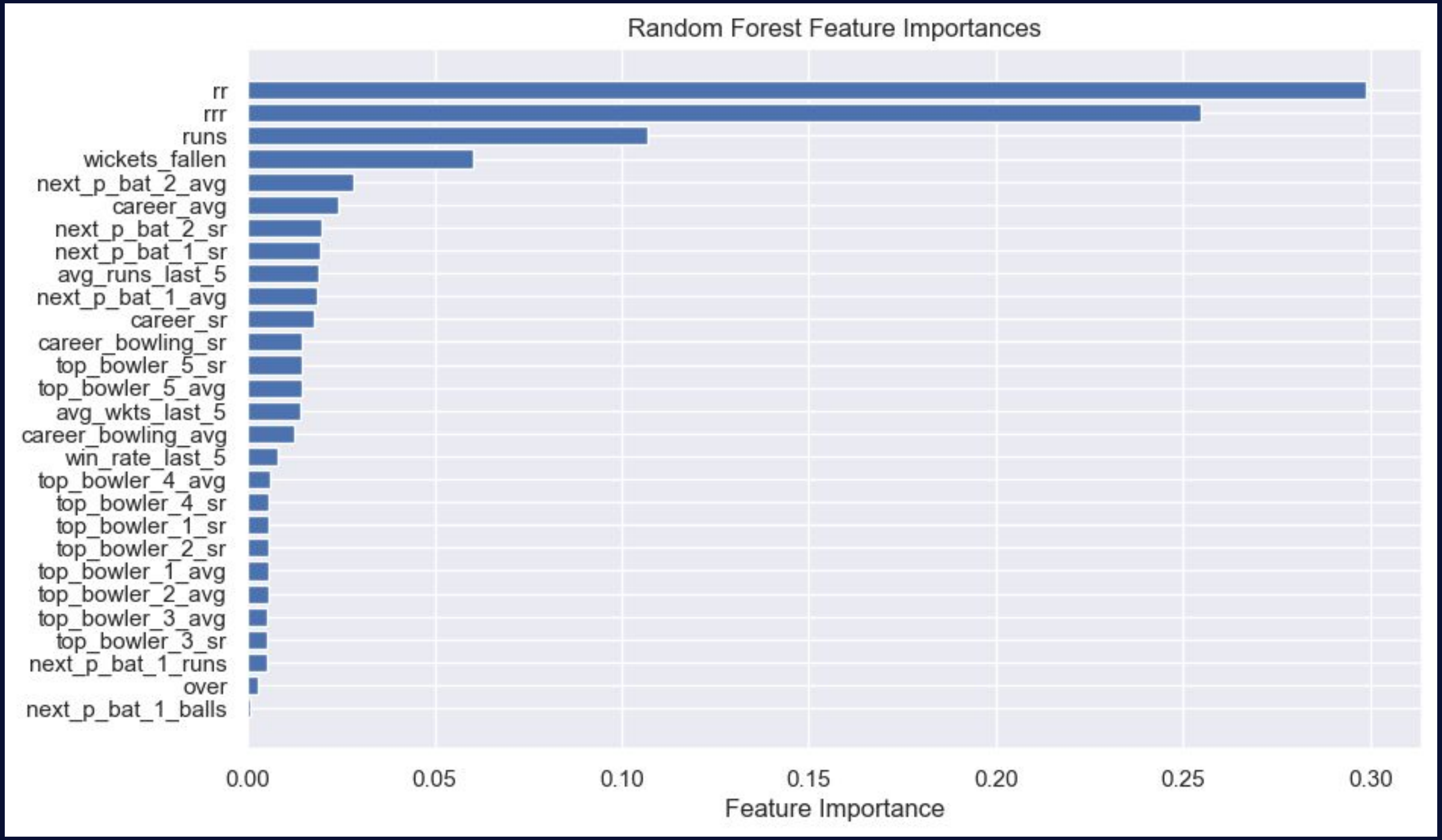
- Standardisation and normalisation of feature values
- Principal Component Analysis to identify the most relevant features



Methodology

Feature Selection

- Random Forest Regression with 32.8 RMSE
- Data points sampled every 10 overs.



Methodology

Dataset Variants

Bare Bones Dataset

- Game state
- Best bowler data (bowling average)
- Best batsman data (batting average)
- Ball by ball data

With Player Stats

- Game state
- Best bowler stats (bowling average)
- Best batsman stats (batting average)
- Ball by ball data
- Player stats

With Pitch/Weather

- Game state
- Best bowler stats (bowling average)
- Best batsman stats (batting average)
- Ball by ball data
- Pitch stats
- Weather data

Methodology

Tested Models

1. Classical ML
   - XGBoost
   - Linear Regression

2. Deep Learning
   - LSTM

Model	LSTM	XGBoost	Linear Regression
Bare bones model	36.516	54.877	44.643
With player stats	37.363	42.393	45.098
With pitch/weather	39.271	40.854	43.487

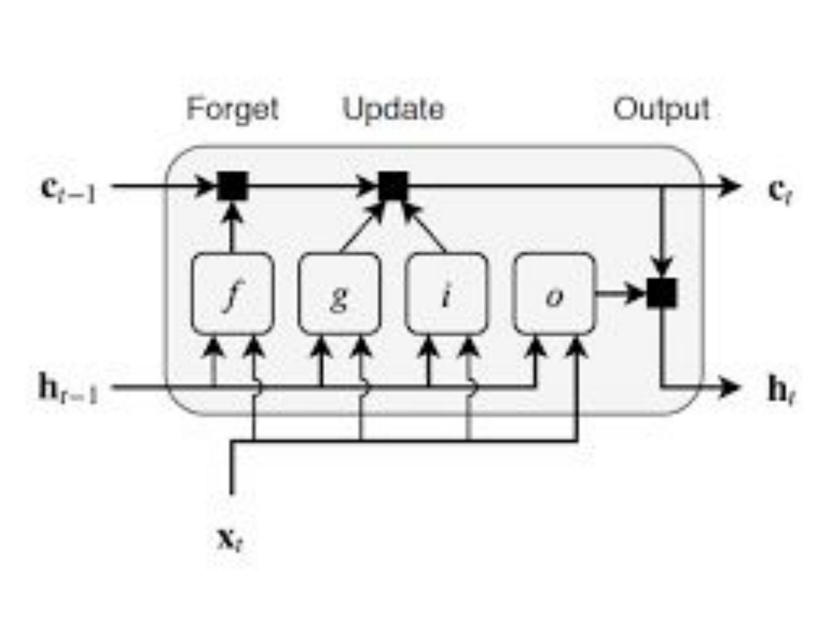
Table 1: Model Performance with Different Feature Sets

Performance Metric is RMSE

Methodology

LSTM

- Trained the LSTM with trial and error
- Used a variety of features with each different dataset
- Best performance for Bare Bones dataset



Methodology

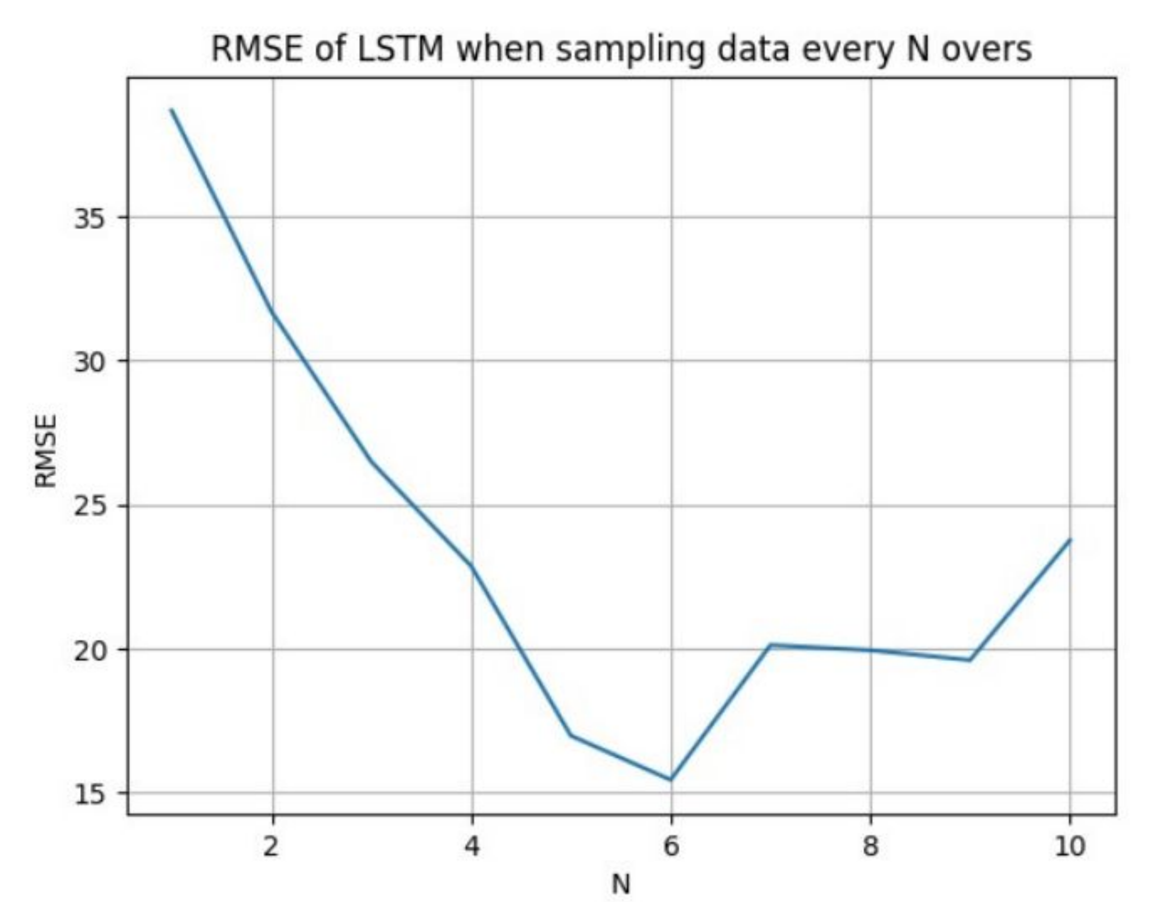
LSTM

```
model = Sequential()
model.add(Input(shape=(len(allowed_overs), X_train_seq.shape[2])))
model.add(Masking(mask_value=-1))

model.add(Conv1D(filters=256, kernel_size=7, activation='relu', padding='same'))
model.add(Conv1D(filters=128, kernel_size=7, activation='relu', padding='same'))
model.add(Conv1D(filters=64, kernel_size=5, activation='relu', padding='same'))

model.add(LSTM(256, return_sequences=True))
model.add(LSTM(128, return_sequences=True, dropout=0.1))
model.add(LSTM(128, return_sequences=True, dropout=0.1))

model.add(LayerNormalization())

model.add(TimeDistributed(Dense(128, activation='relu')))
model.add(TimeDistributed(Dense(64, activation='relu')))
model.add(TimeDistributed(Dense(64, activation='relu')))
model.add(TimeDistributed(Dense(64, activation='relu')))
model.add(TimeDistributed(Dense(32, activation='relu')))
model.add(TimeDistributed(Dense(16, activation='relu')))
model.add(TimeDistributed(Dense(1, activation='linear')))

optimizer = Adam(learning_rate=1e-3)
model.compile(optimizer=optimizer, loss='mean_squared_error')
```

Methodology

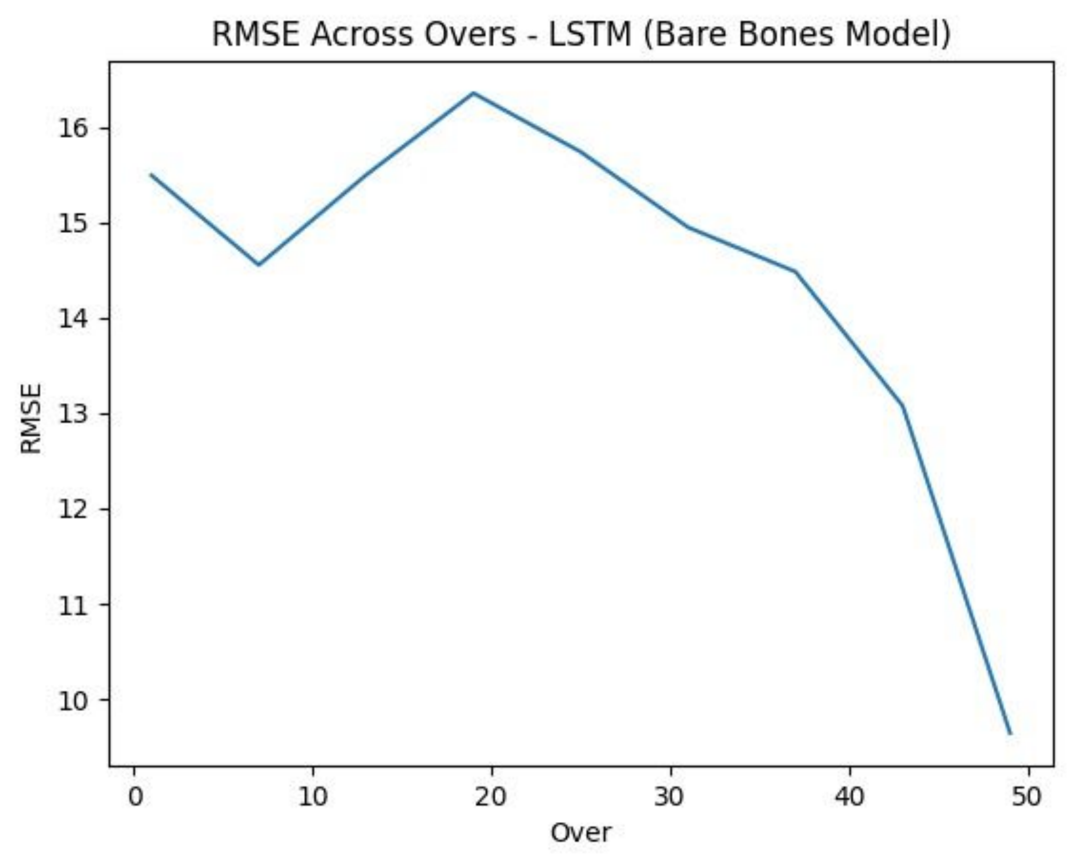
Optimal Sampling

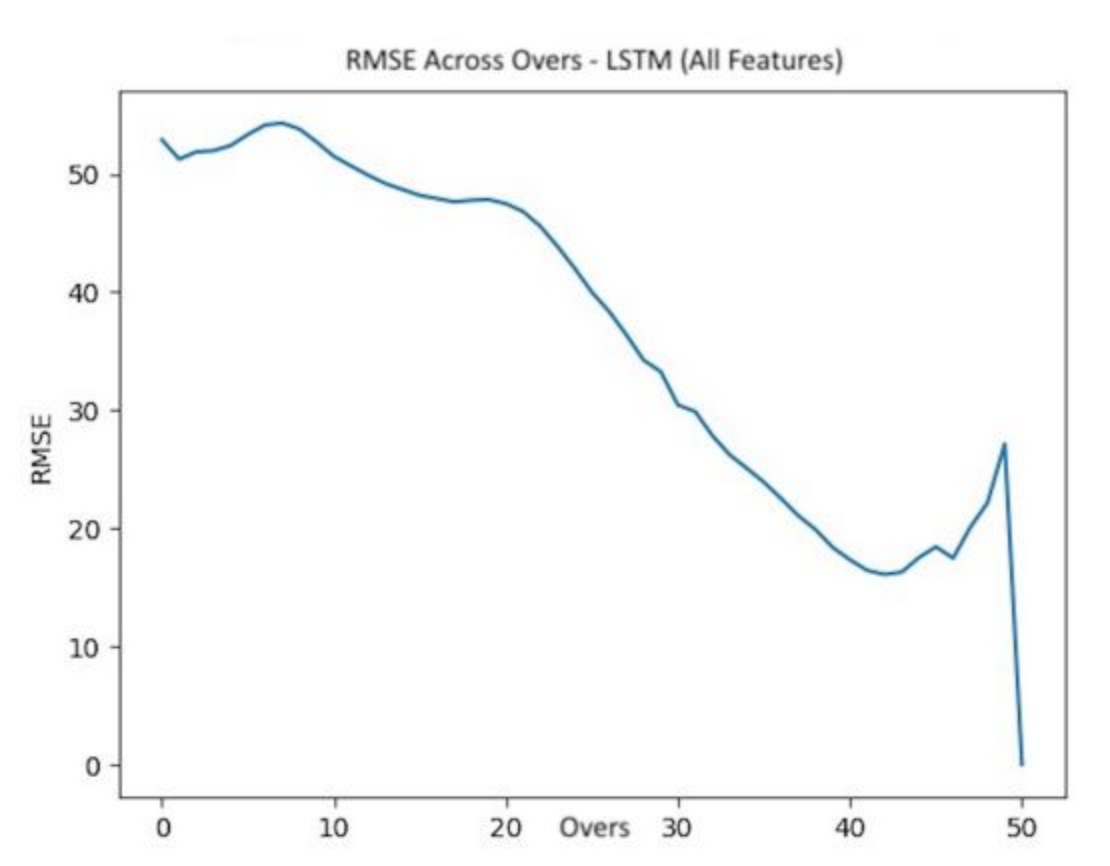
- Sampled the data at every N over to verify best RMSE
- RMSE value dropped down when sampling at every 6th over
- Made 6 models offset by an over to predict every over for the match



Methodology

LSTM - 6 Over Sampling

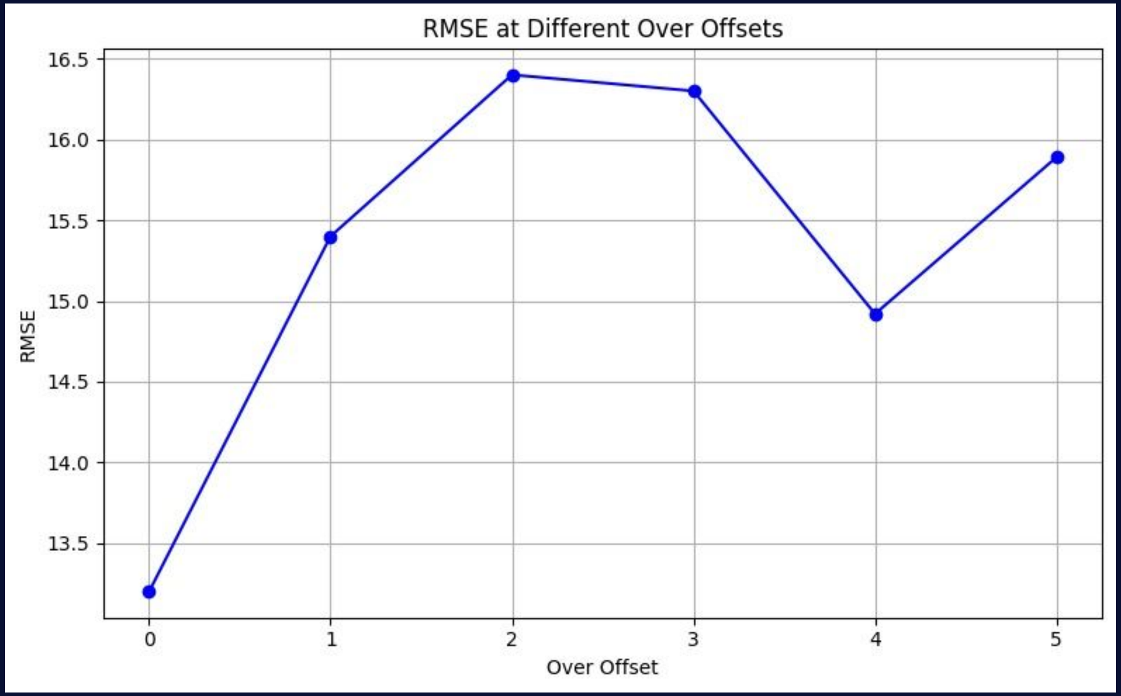




Methodology

LSTM - Multiple Offset Models

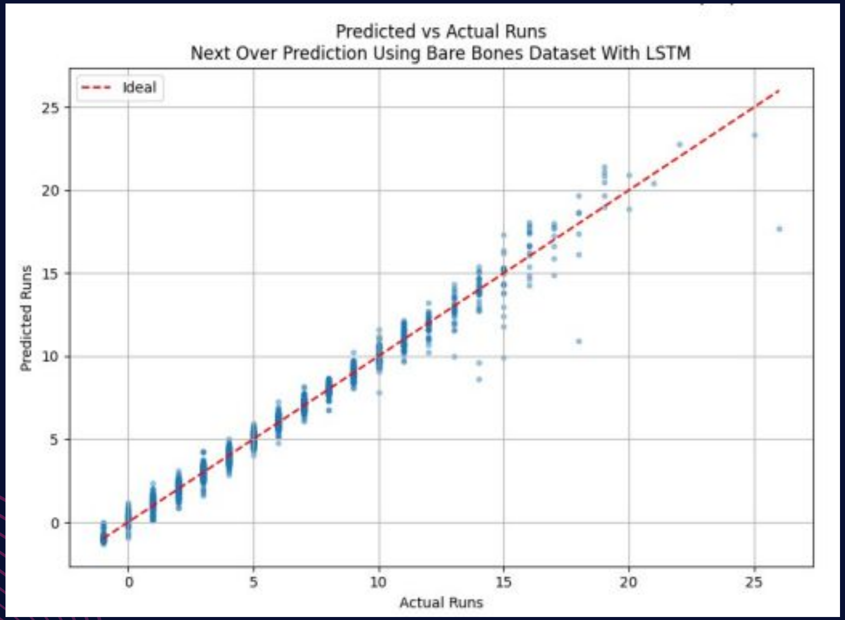
Offset	RMSE
0	13.2
1	15.4
2	16.4
3	16.3
4	14.92
5	15.89
Average	15.35

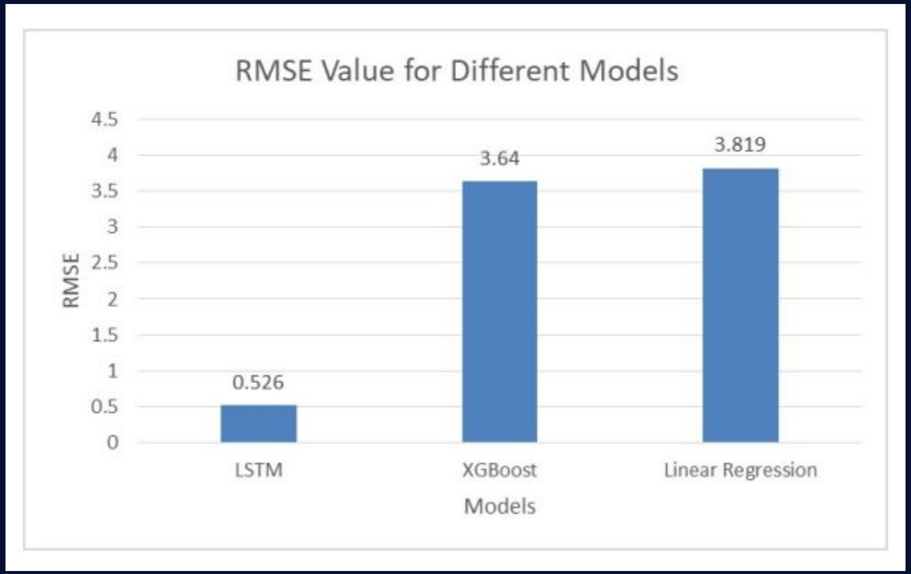


Methodology

Next Over Prediction

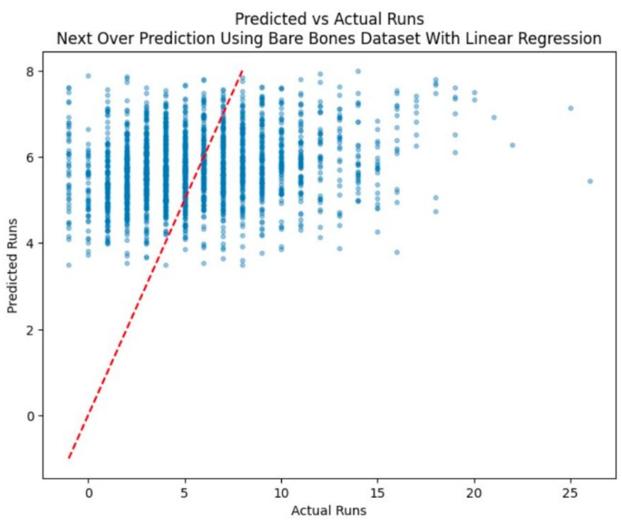
- Also did next over prediction using XGBoost, Linear Regression and LSTM.
- LSTM showed significantly better results



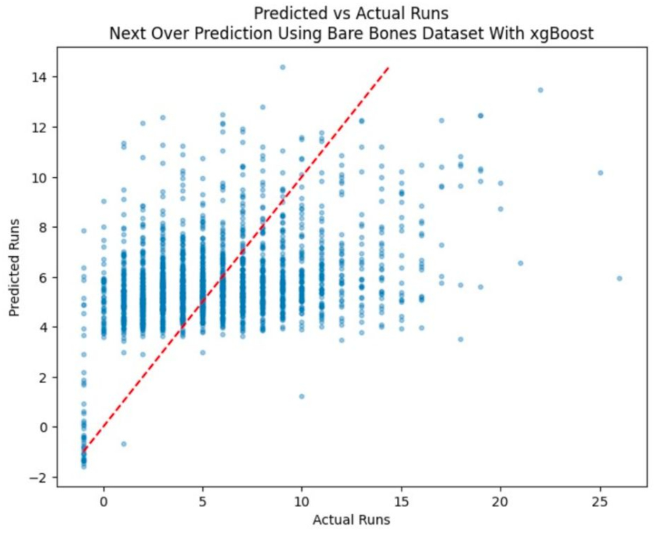


Methodology

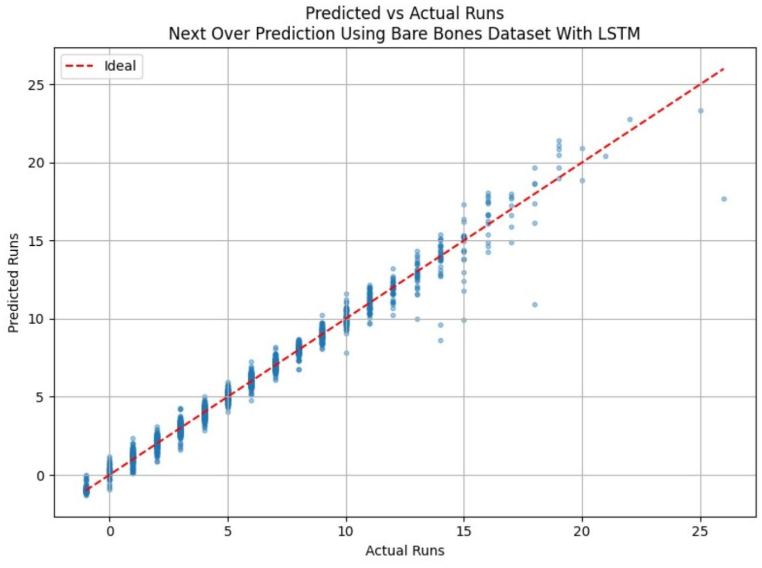
Next Over Prediction



(a) Linear Regression



(b) XGBoost



(c) LSTM

Methodology

Women Specific Model

RMSE - 31.055

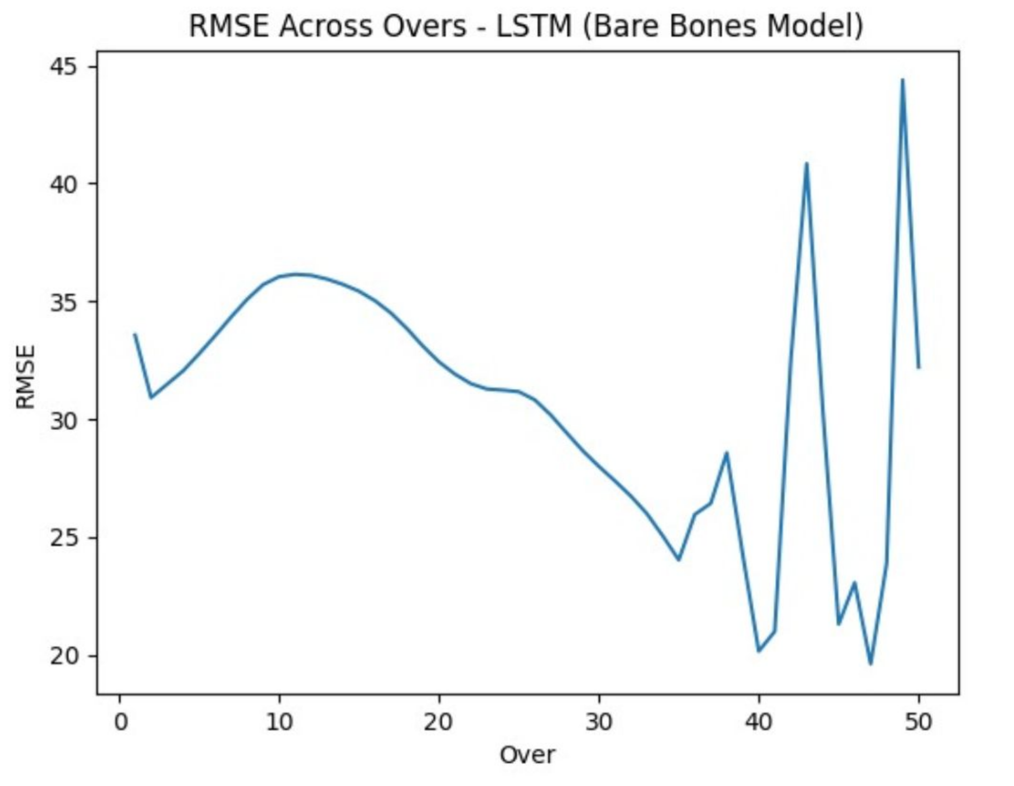


Fig. 6: Womens LSTM Bare Bones

Performance (RMSE)

	XGBoost	Linear Regression	LSTM
Initial models with all features	40.854	43.487	39.271
Models with the bare bones dataset	54.877	44.643	36.516
Tuned Models	27.67	-	15.35
Next Over Prediction	3.64	3.819	0.526

Conclusion

1. Beat benchmark model with 26.27 RMSE.
2. Next-over prediction.
3. Women-specific model.