
One-Day Internationals Predictor

Prepared by: Anhad Singh Jai

8 March 2025

Netaji Subhas University of Technology 27'

EXECUTIVE SUMMARY

Objective

This project aims to either predict the win percentage for the team batting second or estimate the probable target that the team batting first might set.

Goals

1st innings - To predict the total runs made by the team that bats first.

2nd innings - To predict whether a team wins or not, along with their win/loosing probabilities.

Solution

1st innings - With just 3 inputs-the current score, the current wickets left and the number of overs elapsed-we can predict the final target score.

2nd innings - With an additional input, the target score, to the 1st innings model, we can estimate whether the chasing team will win or not along with the winning/loosing chance.

Sources/Technologies used

- Used python as the base programming language to create the machine learning model.
 - Used libraries like pandas, numpy, os, sklearn and matplotlib.
 - All ODI related data-sets were imported from cricsheet website (<https://cricsheet.org/downloads/>).
 - For testing the model, ESPNcricinfo was used to tally the winning probabilities and graphs.
-

WORK-FLOW

- 1) Imported a data-set in csv format from cricsheet, the zip folder provided had about 1400 csv files, each containing the ball by ball data of each one-day international matche.
- 2) Data Cleaning was a cumbersome task, the contained a lot of NaN values, either there was no result of the match (those matches were removed).
- 3) The first innings data was extracted from the combined data, all the runs on each ball including the wides and extras were summed to get the final target for the team batting second.
- 4) Now, in the second innings data, only the useful information was retained, the rest was dropped.
- 5) I used the most basic probabilistic model in machine learning, Naive Bayes. The model took very less time to train, but had an overall accuracy score of 71%.
- 6) 71 seemed a little less, so I implemented the project using Logistic Regression, with got the accuracy to around 82%.
- 7) In order to improve the scores of the model, I used decision tree classifier, which indeed increased the accuracy to a whopping 95%, even the precision, recall, f1-score, for the win and loose classes were 95% (as shown in the image below).
- 8) Upon testing the model on the 2025 champions trophy games, I quickly realised that decision tree was not a good idea.

Model Accuracy: 0.9532					
Classification Report:					
		precision	recall	f1-score	support
	0	0.95	0.96	0.95	58676
	1	0.95	0.95	0.95	58310
	accuracy			0.95	116986
	macro avg	0.95	0.95	0.95	116986
	weighted avg	0.95	0.95	0.95	116986

- 9) The model was over-learning the training data, it would give a fair probability about the existing data in the dataset, but for new matches (The matches not present in the dataset) it would give probabilities which were way-off the values predicted the ESPNcricinfo, even it would sometimes predict the winning class wrong.
 - 10) Then I even tried the Support Vector Machine algorithm, but it was really slow for such large and complex datasets.
 - 11) I decided to stick with Logistic Regression as the primary algorithm for my model, as it was neither very slow to train, nor it was over-learning the data, hence maintaining a balance between accuracy values and the relevance of the output.
 - 12) I then shifted to neural Networks, for better predictions.
-

13) **FIRST INNINGS LSTM BASED NEURAL NETWORK:**

—>The model is trained on 3 features: curr_runs, curr_wickets, overs. The data is reshaped into (samples,1,3) for LSTM processing.

—> Architecture: A 2-Layer LSTM network with dropout(dropout rate = 0.3 to prevent overfitting) and batch normalisation

—>Optimised with Adam + MSE loss.

—> Training: Epochs size used was 100, Batch size was 64, Callbacks used : ReduceLROnPlateau(reduces learning rate if validation loss doesn't improve for 5 epochs) AND EarlyStopping(stops training if validation loss doesn't improve for 10 epochs).

LAYER	TYPE	UNITS	ACTIVATION	OTHER DETAILS
1	LSTM	64	Tanh	return_sequences=True (to stack another LSTM layer)
2	Dropout	-	-	Dropout rate = 0.3 (to prevent overfitting)
3	LSTM	32	Tanh	return_sequences=False(final LSTM layer)
4	Batch Normalisation	-	-	Normalises activations for stability
5	Dense	16	ReLU	Fully connected Layer
6	Dense (Output)	1	Linear	Predicts the final target Score

14) **SECOND INNINGS NEURAL NETWORK:**

—>The model is trained on 4 features: total_runs_y(Target_score), curr_runs, curr_wickets, overs. Shape of input data is (samples, 4), where each sample represents the match situation at a given ball.

—> Architecture: 3 fully connected layers with ReLU activation, dropout layers(0.2 To prevent overfitting.

—> Optimised with Adam + Binary Cross-Entropy loss.

—> Training: Epochs size is 100, Batch size is 32.

—> Output is in form of Binary class either 0 or 1, along with their probabilities.

—> 0 : means the chasing team will lose.

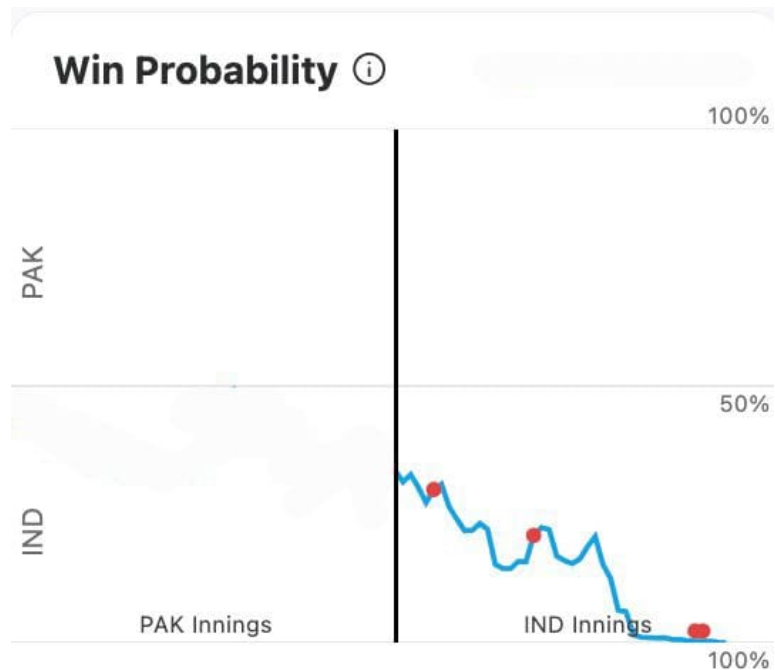
—> 1: means the chasing team wins.

LAYER	TYPE	UNITS	ACTIVATION	OTHER DETAILS
1	Dense	32	ReLU	Input Layer (4 features)
2	Dropout	-	-	Dropout rate = 0.2 (Prevents Overfitting)
3	Dense	16	ReLU	Hidden Layer 1
4	Dropout	-	-	Dropout rate = 0.2
5	Dense	8	ReLU	Hidden Layer 2
6	Dense	1	Sigmoid	Output Layer (Binary Classification: 1-> Chase Successful, 0-> Chase Failed.

TESTING THE 2ND INNINGS MODEL ON REAL LIFE SITUATIONS

1) INDIA VS PAKISTAN (CHAMPIONS TROPHY MATCH) 2025

BELOW IS THE ORIGINAL GRAPH FOR WINNING PROBABILITY OF IND(CHASING TEAM)

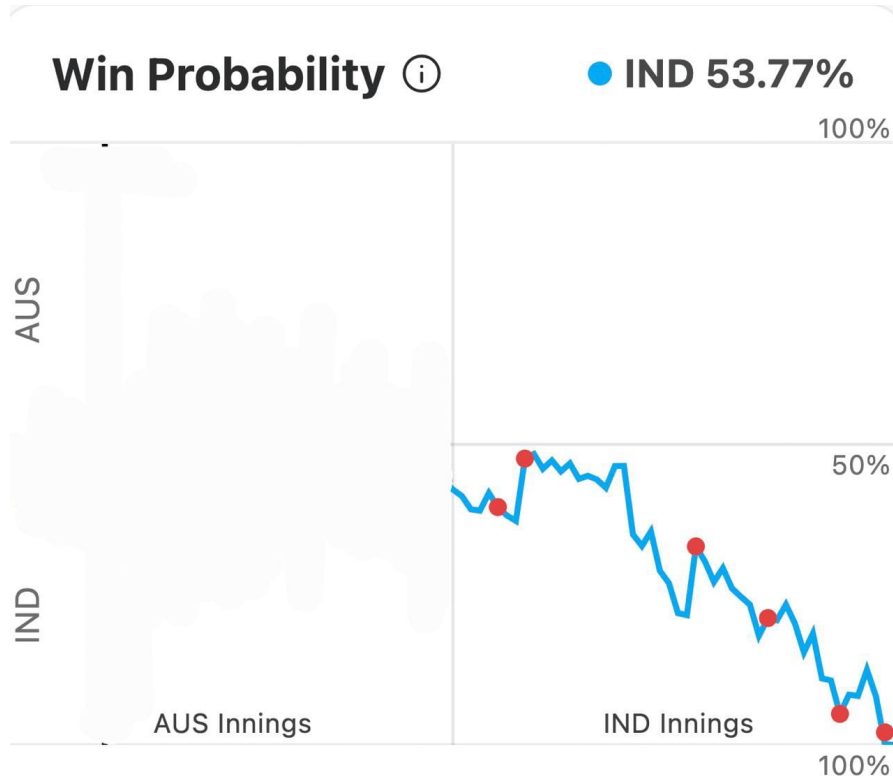


BELOW IS THE PREDICTED GRAPH FOR THE WINNING PROBABILITY OF THE IND (CHASING TEAM)

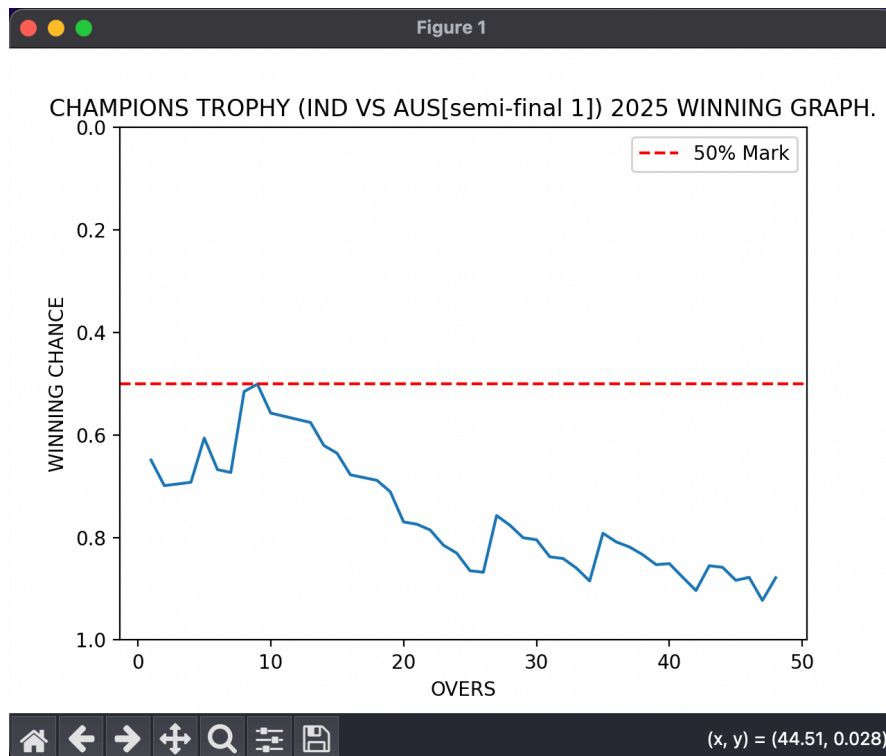


2) INDIA VS AUSTRALIA (CHAMPIONS TROPHY semi-final 1) 2025

BELOW IS THE ORIGINAL GRAPH FOR WINNING PROBABILITY OF IND(CHASING TEAM)

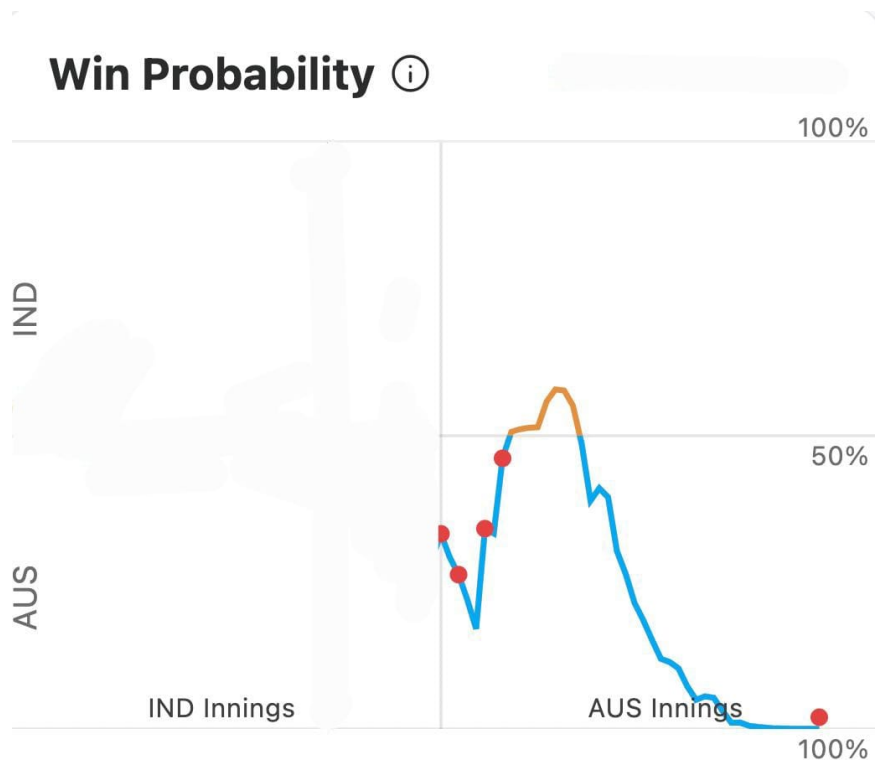


BELOW IS THE PREDICTED GRAPH FOR THE WINNING PROBABILITY OF THE IND (CHASING TEAM)



3) INDIA VS AUSTRALIA (ODI WORLD CUP 2023 FINALS) AT AHMEDABAD

BELOW IS THE ORIGINAL GRAPH FOR WINNING PROBABILITY OF IND (DEFENDING THEIR TARGET)



BELOW IS THE PREDICTED GRAPH FOR THE WINNING PROBABILITY OF THE IND (CHASING TEAM)



TESTING THE 1ST INNINGS MODEL ON REAL LIFE SITUATIONS



TESTING PROCESS- We will take 4 test cases for each match, we will input the current runs, current wickets and the overs elapsed till now. We will take the output at 14th, 23rd, 33rd, 45th overs. NOTE- These overs chosen for testing are random. Our aim is to compare our machine learning output with 2 datapoints, the ESPNcricinfo score prediction at that instant and the actual scorecard of the match at the end of the innings.

1) INDIA VS NEW ZEALAND(CHAMPIONS TROPHY 2025 FINALS) AT DUBAI.

THE ACTUAL SCORE-CARD

RESULT

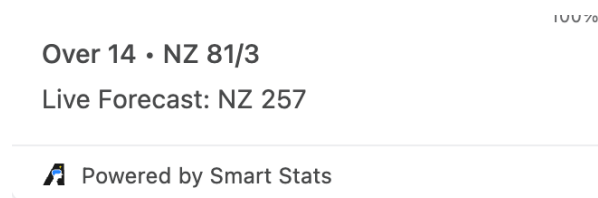
Final (D/N), Dubai (DICS), March 09, 2025, ICC Champions Trophy

 New Zealand	251/7
 India	(49/50 ov, T:252) 254/6
India won by 4 wickets (with 6 balls remaining)	

So as we can see The final score that New Zealand had put up on the board was 251, hence a target of **252** was set.

1) 14th Over:

Predicted by ESPN



Predicted by Model

```
Enter the current runs: 81
Enter the current wickets lost: 3
Enter the over going on right now: 14
Predicted Final Score: 247.0
```

So we observe here, our prediction is fairly close to the actual target, but is 10 behind what ESPN predicted.

2) 23rd Over:

Predicted by ESPN

Over 23 • NZ 107/3

Live Forecast: NZ 260

Predicted by Model

```
Enter the current runs: 107
Enter the current wickets lost: 3
Enter the over going on right now: 23
Predicted Final Score: 256.0
```

So, as we observe here, my prediction lies between what the real target was and what ESPN predicted.

3) 33rd Over:

Predicted by ESPN

Over 33 • NZ 147/4

Live Forecast: NZ 258

Predicted by Model

```
Enter the current runs: 147
Enter the current wickets lost: 4
Enter the over going on right now: 33
Predicted Final Score: 253.0
```

4) 45th Over:

Predicted by ESPN

Over 45 • NZ 201/5



Live Forecast: NZ 240

Predicted by Model

```
Enter the current runs: 201
Enter the current wickets lost: 5
Enter the over going on right now: 45
Predicted Final Score: 248.0
```

2) INDIA VS PAKISTAN(CHAMPIONS TROPHY 2025)

THE ACTUAL SCORE CARD:

 Pakistan	241
 India	(42.3/50 ov, T:242) 244/4
India won by 6 wickets (with 45 balls remaining)	

1) 14th Over:

Predicted by ESPN

Over 14 • PAK 61/2
Live Forecast: PAK 241

Predicted by Model

```
Enter the current runs: 61
Enter the current wickets lost: 2
Enter the over going on right now: 14
Predicted Final Score: 243.0
```

2) 23rd Over:

Predicted by ESPN

Over 23 • PAK 90/2
Live Forecast: PAK 246

Predicted by Model

```
Enter the current runs: 90
Enter the current wickets lost: 2
Enter the over going on right now: 23
Predicted Final Score: 244.0
```

3) 33rd Over:

Predicted by ESPN

Over 33 • PAK 150/2

Live Forecast: PAK 270

Predicted by Model

```
Enter the current runs: 150
Enter the current wickets lost: 2
Enter the over going on right now: 33
Predicted Final Score: 273.0
```

Here we observe that both ESPN and my model's prediction deviates from the actual target, this might be because at that stage it seemed Pakistan could make a score in the 270's.

4) 45th Over:

Predicted by ESPN

Over 45 • PAK 212/7

Live Forecast: PAK 242

Predicted by Model

```
Enter the current runs: 212
Enter the current wickets lost: 7
Enter the over going on right now: 45
Predicted Final Score: 244.0
```

3) ENGLAND VS AUSTRALIA(CHAMPIONS TROPHY 2025)

THE ACTUAL SCORE CARD:

 **England**

351/8

 **Australia**

(47.3/50 ov, T:352) **356/5**

Australia won by 5 wickets (with 15 balls remaining)

1) 14th Over:

Predicted by ESPN

Over 14 • ENG 102/2

Live Forecast: ENG 302

Predicted by Model

```
Enter the current runs: 102
Enter the current wickets lost: 2
Enter the over going on right now: 14
Predicted Final Score: 298.0
```

2) 23rd Over:

Predicted by ESPN

Over 23 • ENG 154/2

Live Forecast: ENG 331

Predicted by Model

```
Enter the current runs: 154
Enter the current wickets lost: 2
Enter the over going on right now: 23
Predicted Final Score: 333.0
```

3) 33rd Over:

Predicted by ESPN

Over 33 • ENG 216/3

Live Forecast: ENG 346

Predicted by Model

```
Enter the current runs: 216  
Enter the current wickets lost: 3  
Enter the over going on right now: 33  
Predicted Final Score: 342.0
```

4) 45th Over:

Predicted by ESPN

Over 45 • ENG 303/5

Live Forecast: ENG 343

Predicted by Model

```
Enter the current runs: 303  
Enter the current wickets lost: 5  
Enter the over going on right now: 45  
Predicted Final Score: 349.0
```

A Clear observation that we can see, or which might be interpreted as common sense is that as the overs progress, and we get closer to the end of the first innings, the delta between the predicted score and the score made by the team batting first gets minimised.
