Group 13
Hemant Gupta
Madhvendra Singh
Samarth Anand

MLPR Presentation

Try Pitch

Problem Statement

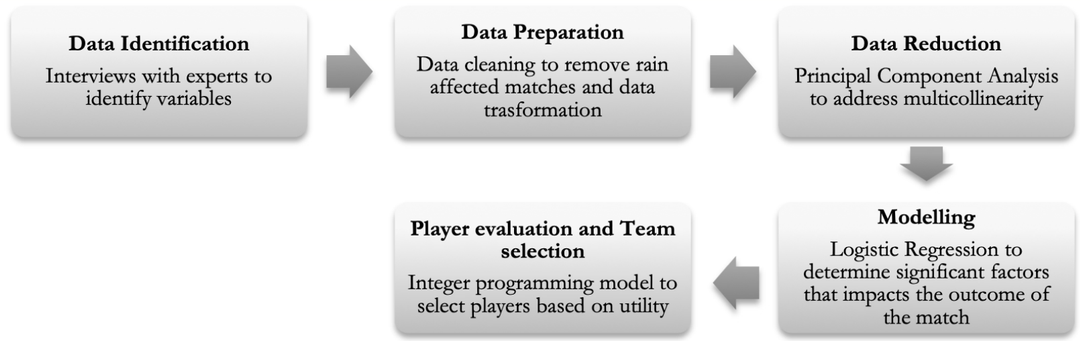
To develop a machine learning model that can analyze historical player performance data from cricket matches and predict the optimal fantasy team composition for an upcoming match.

The ultimate goal is to maximize the total fantasy points scored by the team selected by the participant, thereby increasing the chances of winning in the fantasy cricket competition.

- We are currently building this for the Indian Premier League (IPL), which will allow us to deploy and test our model as their upcoming season begins on the 22nd of March 2024.
- Data for the IPL is also readily & easily available, and more importantly, is consistent Year-on-Year, and of the same match format (T20), making it easier to work with.

Try Pitch

Literature Survey



- Scatter plots and correlation analysis explore data distribution and variable relationships with match outcome.
- Variables having weak correlation are removed from further modelling.

Table 2 Batting variables under study

Variable	Top	Middle	Lowermiddle
Runs	batopnar	batmidar	Batlmar
Strike Rate	batopnsr	batmidsr	Batlmsr
Dotball %	batopndbp	batmiddbp	Batlmdbp
Boundary %	batopnbp	batmidbp	Batlmbp
Boundary frequency	batopnbf	batmidbf	Batlmbbf
RSS	batopnrss	batmidrss	Batlmbrss
uncomfortables	batopuc	batmiduc	Batlmbuc

Table 3 Bowling variables under study

Variable	Fast	Spin
Economy	Fsteco	spneco
Average runs conceded	fstblavg	spnblavg
Bowling Strike rate	fstblsr	spnblsr
Dotball %	fstbldb	spnbldb
Boundary %	Fsblbp	spnblbp
Boundary frequency	Fsblbf	spnblbf

- Due to potential multicollinearity among the many derived batting and bowling variables, Principal Component Analysis (PCA) was used to transform the original correlated variables into a new set of uncorrelated principal components.
- 10 principal components were retained, explaining 85.98% of the total variance. Variables were loaded onto these components using varimax rotation, providing insights into their correlations within batting and bowling roles.

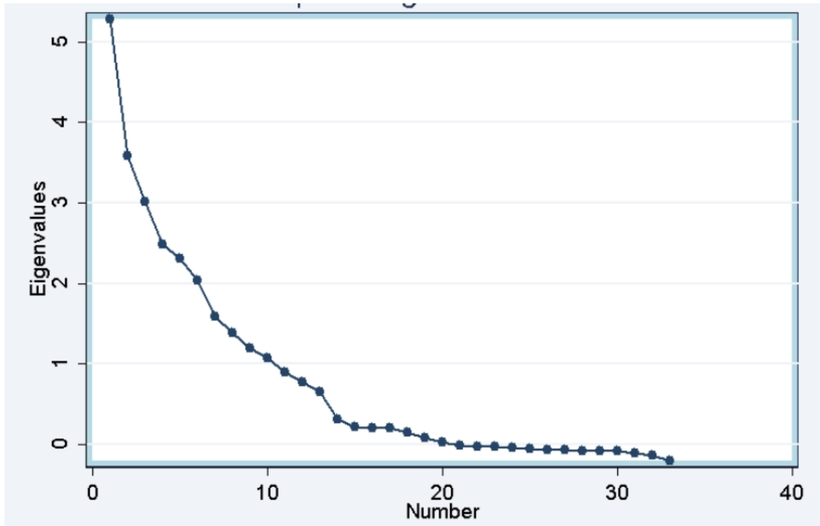
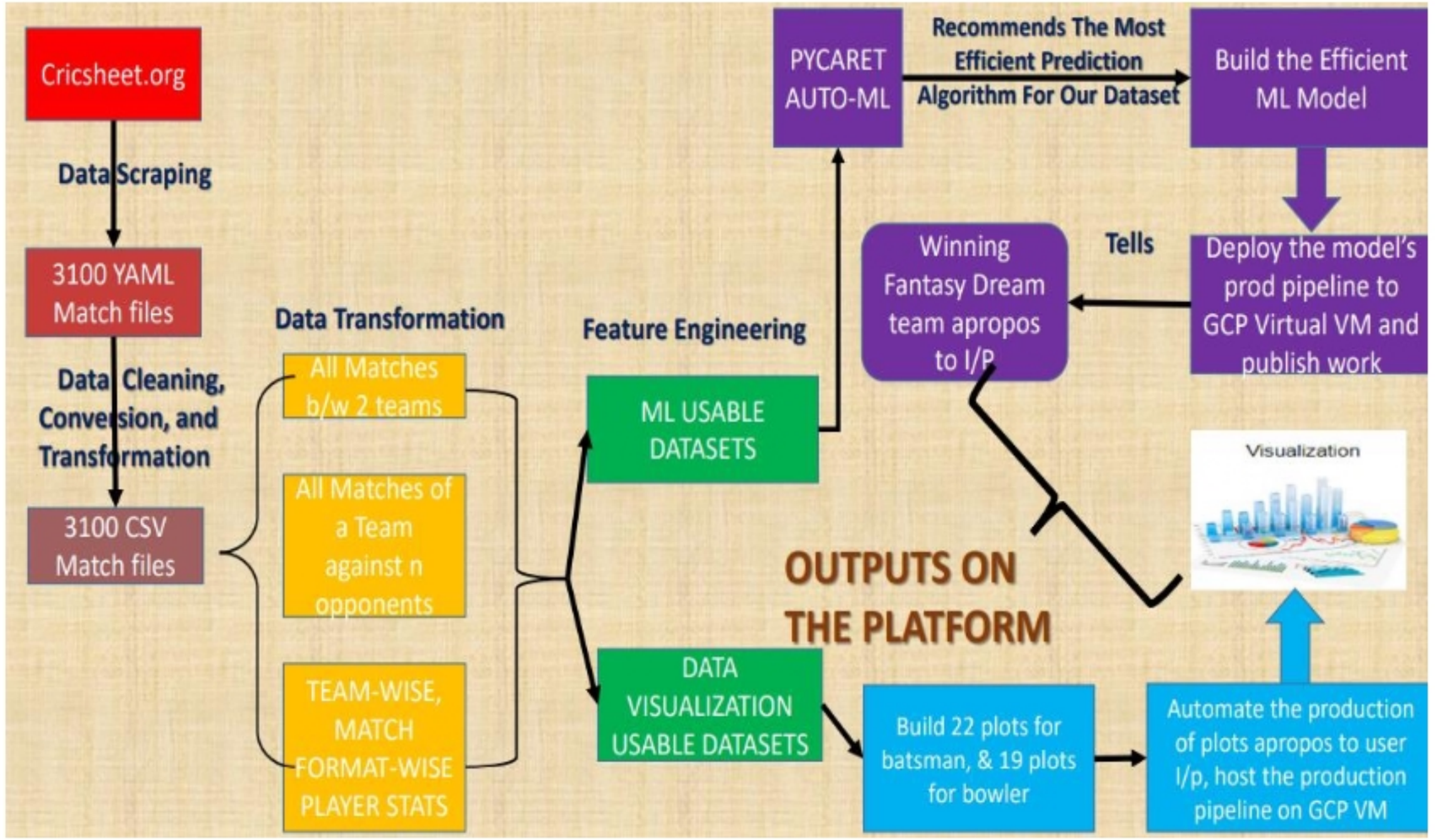


Figure 2 Scree plot

Bharathan, RP Sundarraj and Abhijeet. "A Self-Adapting Intelligent Optimized Analytical Model for team selection using player performance utility in Cricket." (2015).

Literature Survey



Data collected using:
- Cricsheet

Packages & Modules used:
- Yorkpy
- Plotly
- PyCaret

ML Methodology used:
- Extra Trees Regressor Model (ETR)

Algorithms to select team:
- Knapsack Algorithm
- Greedy Algorithm

S, Sachin & HV, Prithvi & Nandini, C. (2022). Data Science Approach to predict the winning Fantasy Cricket Team Dream 11 Fantasy Sports. 10.48550/arXiv.2209.06999.

Try Pitch

Preprocessing

- For Batsman:
  - Runs
  - Balls
  - 4s
  - 6s
  - 50s
  - 100s
  - Duck Out
  - Strike Rate
  - Rival
  - Venue
- For Bowlers:
  - Overs
  - Runs
  - Concede
  - Maidens
  - Wickets
  - Economy Rate
  - Rival
  - Venue

	batsman	runs	balls	4s	6s	SR	bowler	fielders	kind	player_out	date	team2	winner	venue	team1	MF	50s	100s	ducks	dr11Score
0	A Flintoff	24.0	25	2	1	96.0	Harbhajan Singh	0	caught and bowled	A Flintoff	2009-04-18	Mumbai Indians	Mumbai Indians	Newlands	Chennai Super Kings	IPL	0	0	0	30
2	A Flintoff	16.0	18	1	0	89.0	A Nehra	['DA Warner (sub)']	caught	A Flintoff	2009-04-23	Delhi Daredevils	Delhi Daredevils	Kingsmead	Chennai Super Kings	IPL	0	0	0	20
3	A Mukund	0.0	1	0	0	0.0	Sohail Tanvir	0	bowled	A Mukund	2008-05-24	Rajasthan Royals	Rajasthan Royals	MA Chidambaram Stadium, Chepauk	Chennai Super Kings	IPL	0	0	1	0
5	A Nehra	1.0	4	0	0	25.0	notOut	notOut	notOut	notOut	2015-04-30	Kolkata Knight Riders	Kolkata Knight Riders	Eden Gardens	Chennai Super Kings	IPL	0	0	0	5
6	A Nehra	0.0	1	0	0	0.0	MA Starc	['KD Karthik']	run out	MM Sharma	2015-05-04	Royal Challengers Bangalore	Chennai Super Kings	MA Chidambaram Stadium, Chepauk	Chennai Super Kings	IPL	0	0	1	0

- The program extracts relevant rows from the dataset based on user inputs.
- The predictor transforms the extracted rows into a two-dimensional matrix with appropriate columns for batsmen (12 columns) and bowlers (11 columns). It then fits the data matrix to the respective prediction models and provides the user with the average of all predictions. For all-rounders, separate predictions are made for batting and bowling, and their Dream11 scores are summed.

S, Sachin & HV, Prithvi & Nandini, C. (2022). Data Science Approach to predict the winning Fantasy Cricket Team Dream 11 Fantasy Sports. 10.48550/arXiv.2209.06999.

Try Pitch

Literature Survey

Data collected using:

- GitHub Database

Data Preprocessing:

- Total Runs Scored
- Total Balls Faced
- 50 Scored
- 100 Scored
- No 4s
- No 6s
- Runs Conceded
- Overs Bowled
- No Wickets Taken
- Average Runs
- Strike Rates
- Bowling Average
- Economy Rate
- Opposition Team
- Match Venue
- Pitch Type

Models Used:

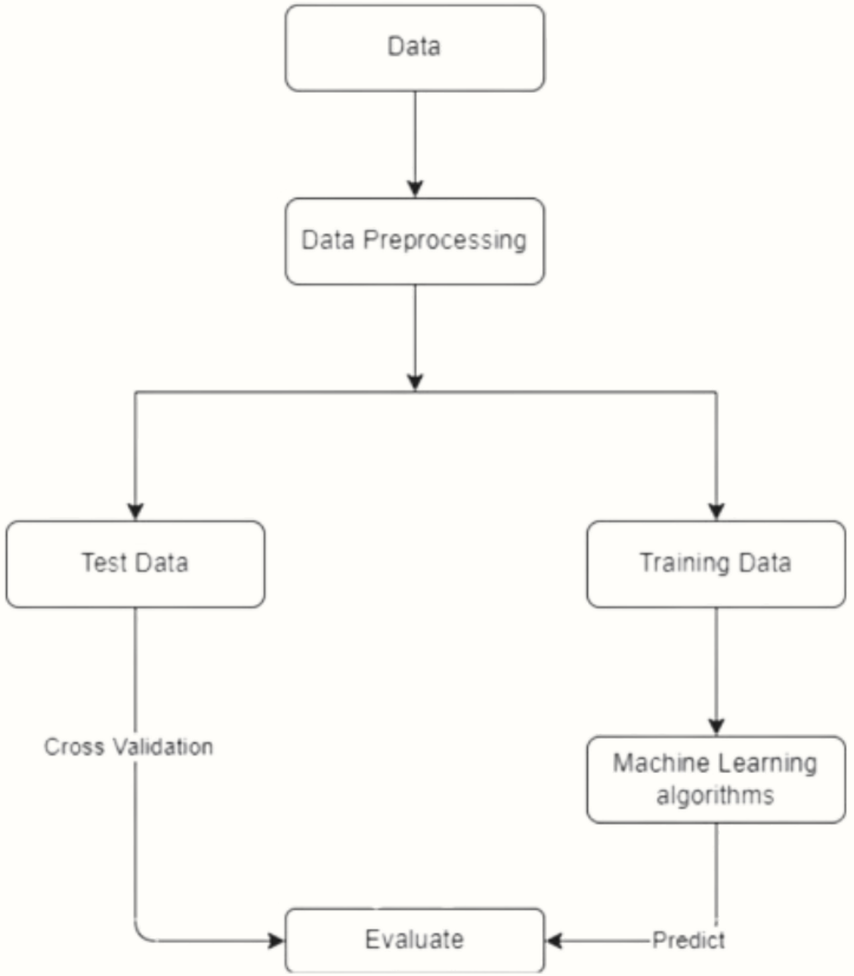
- XGBoost:

$$L(f) = \sum_{i=1}^{n} L(\hat{y}_i, y_i) + \sum_{m=1}^{M} \Omega(\delta_m)$$

$$\Omega(\delta) = \alpha \mid \delta \mid + 0.5\,\beta \mid\mid w \mid\mid^2$$

- Catboost:
  - For categorical features with more categories than a specified threshold, CatBoost applies a three-step process:
    - Randomly divides the data into subsets.
    - Converts labels to integers and categorical features to numerical values.
    - Calculates a score to determine the best split point.

- Random Forest:
  - Random Forests conduct bootstrap aggregating of decision trees, with random subsets of features considered at each split to reduce overfitting.
  - This randomness and ensemble approach make Random Forests more robust than single decision trees while avoiding correlations between trees.



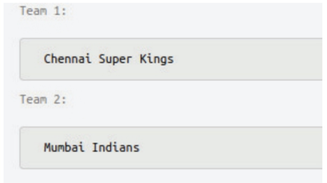
S. Choudhari, N. Wagholikar, A. Swaminathan and S. Kurhade, "Dream11 IPL Team Recommendation using Machine Learning and Skill-Based Ranking of Players," 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, 2022, pp. 1-6, doi: 10.1109/ICONAT53423.2022.9725819.

Try Pitch

Literature Survey

Team Selection:

- Python PuLP library
- Offers greater customisability & accounts for greater constrains in team selection.



$$Error\_Rate = \frac{Actual\_Points - Predicted\_Points}{Actual\_points}$$

Table 1. Error percentages

Model	Error%
XGBoost	38
CatBoost	34
Random Forest	45

$$Error\_Rate = \frac{A\_total\_Points - P\_total\_Points}{A\_total\_points}$$

Table 3. Error percentages

Lowest Error	Highest Error	Average Error%
12.0%	18.6%	15.3%

S. Choudhari, N. Wagholikar, A. Swaminathan and S. Kurhade, "Dream11 IPL Team Recommendation using Machine Learning and Skill-Based Ranking of Players," 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, 2022, pp. 1-6, doi: 10.1109/ICONAT53423.2022.9725819.

Try Pitch

Akshit Behera (published by Analytic Vidhya)

Literature Survey

A Scientific Method to Select Your Fantasy Sports Team

Using machine learning and optimization techniques to select your team on the fantasy sports platform — Dream11.com

https://medium.com/analytics-vidhya/a-scientific-method-to-select-your-fantasy-sports-team-b23726136256

Try Pitch

Dataset

Ball_by_Ball (csv)

- ID: int64
- innings: int64
- overs: int64
- ballnumber: int64
- batter: object
- bowler: object
- non-striker: object
- extra_type: object
- batsman_run: int64
- extras_run: int64
- total_run: int64
- non_boundary: int64
- isWicketDelivery: int64
- player_out: object
- kind: object
- fielders_involved: object
- BattingTeam: object

Matches (csv)

- ID: int64
- match_id: int64 (corresponds to ID of Ball_by_Ball.csv)
- City: object
- Date: object
- Season: object
- Team1: object
- Team2: object
- SuperOver: object
- WinningTeam: object
- Method: object
- Player_of_Match: object
- Team1Players: object
- Team2Players: object
- Umpire1: object
- Umpire2: object
- Venue: object

Scraped from: https://cricsheet.org/

Dataset

- The dataset consists of 2 CSVs linked by a key 'ID'
- The dataset that we created offers a ball by ball data of all IPL matches from 2008-2024 which allows us to create a variety of features.
- The data was created by scraping cricsheet.org, where the data is freely & openly available, doing away with any ethical concerns.

Ball_by_Ball (csv)

ID	match_id	innings	overs	ballnumber	batter	bowler	non-striker	extra_type	batsman_run	extras_run	total_run	non_boundary	isWicketDelivery	player_out	kind	fielders_involved	BattingTeam
3359821001	335982	1	0	1	SC Ganguly	P Kumar	BB McCullum	legbyes	0	1	1	0	0	NA	NA	NA	Kolkata Knight Riders
3359821002	335982	1	0	2	BB McCullum	P Kumar	SC Ganguly	NA	0	0	0	0	0	NA	NA	NA	Kolkata Knight Riders
3359821003	335982	1	0	3	BB McCullum	P Kumar	SC Ganguly	wides	0	1	1	0	0	NA	NA	NA	Kolkata Knight Riders

Try Pitch

Features Preprocessing

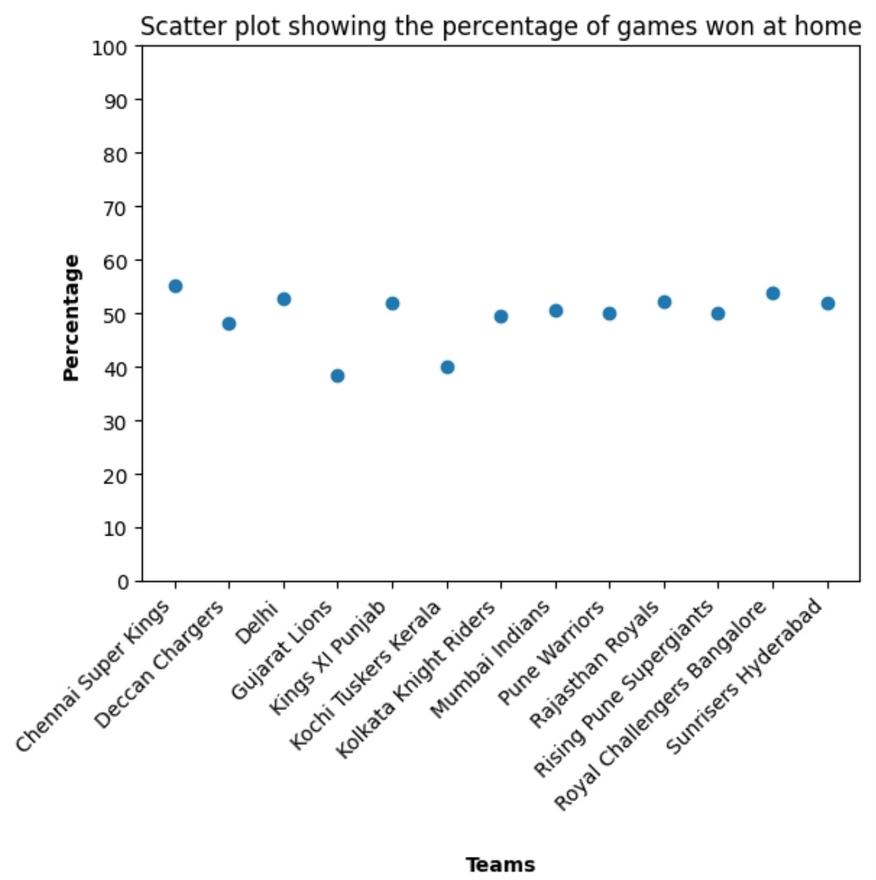
- Feature selection was done based on the Dream11 algorithm & preprocessing conducted by other researchers.
- Delhi Daredevils was rebranded to Delhi Capitals, therefore both Delhi Daredevils and Delhi Capitals are saved as Delhi.
- Rising Pune Supergiants also rebranded to Rising Pune Supergiant. That has also been fixed.
- Dropped all rain affected matches or abandoned matches.

Home-field advantage was found to be significant for many teams including India, South Africa, Sri Lanka, New Zealand, and Pakistan. Among all theteams, the South African team has the highest winning chance (72%) in home games.

K. P. Jayalath, "A machine learning approach to analyze ODI cricket predictors," Journal of Sports Analytics, doi: 10.3233/JSA-17175.

Average percentage of matches won at home: 49.57
→ Home team advantage was found to be near zero.

$$\text{Home Win Percentage (Team)} = \frac{\text{Home Wins}}{\text{Total Wins}} \times 100$$



Try Pitch

Features Preprocessing

- **Batting Average**

$$Batting Average = \frac{\sum_{i=1}^{n} Runs_i}{Dismissals}$$

- **Bowling Average**

excluding run outs, wickets taken on no balls, and wickets taken on free hits

$$Bowling\ Average = \frac{\sum_{i=1}^{n} Runs\_Conceded_i}{Wickets\_Taken}$$

- **Balls per wicket**

minimum 200 balls bowled

$$Balls\ per\ Wicket = \frac{\sum_{i=1}^{n} Balls\_Bowled_i}{Wickets\_Taken}$$

- **Balls per wicket**

Strike Rates (minimum 200 balls faced, excluding leg byes, wide, and no balls

$$Strike\ Rate = \frac{\sum_{i=1}^{n} Runs\_Scored_i}{\sum_{i=1}^{n} Balls\_Faced_i} \times 100$$

- **Avg Balls faced**

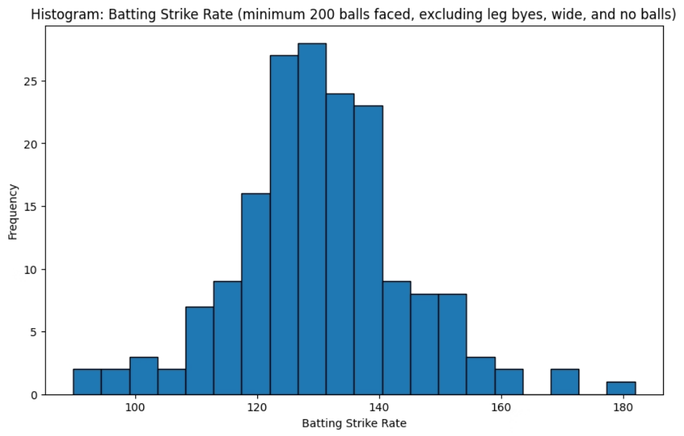
minimum 200 balls faced, excluding wide and no balls

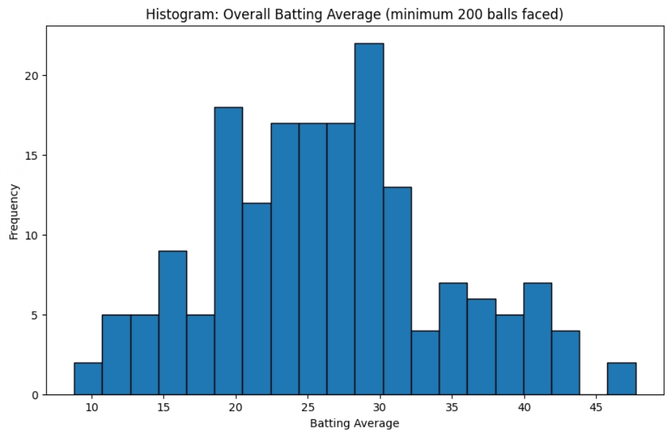
$$Average\ Balls\ Faced\ per\ Innings = \frac{\sum_{i=1}^{n} Balls\_Faced_i}{Innings\_Played}$$

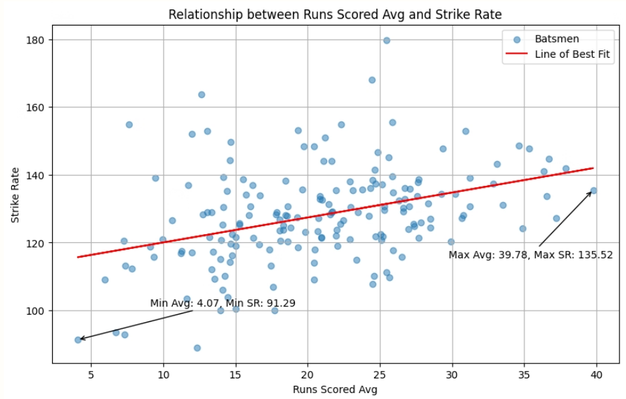
- No. of 100s
- No. of 50s
- Avg. number of 4s per inning (minimum 200 balls faced)
- Avg. number of 6s per inning (minimum 200 balls faced)

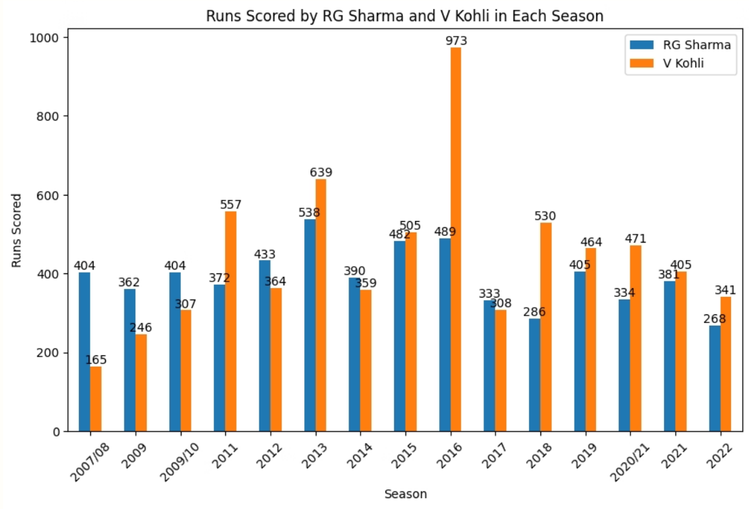
- Dot Ball %age for batsmen (min. 200 balls faced)
- Runs given per game (min. 200 balls faced)
- Wickets taken per game (min. 200 balls faced)
- Avg. overs bowled per game (min. 200 balls faced)
- Runs given per over (min. 200 balls faced)
- Dot Ball %age for bowlers (min. 200 balls faced)
- %age of 6s conceded per over (min. 200 balls faced)
- %age of 4s conceded per over (min. 200 balls faced)

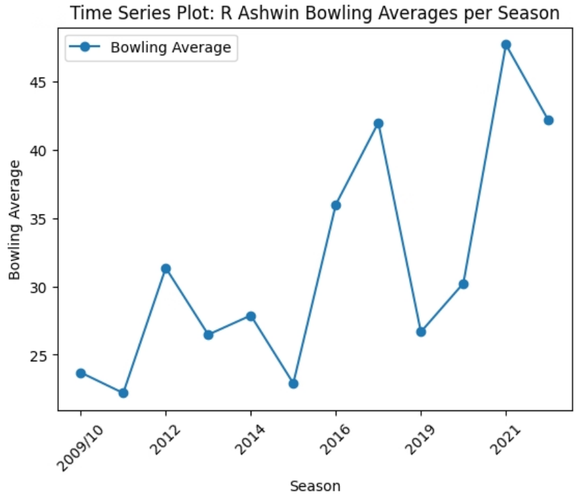
EDA

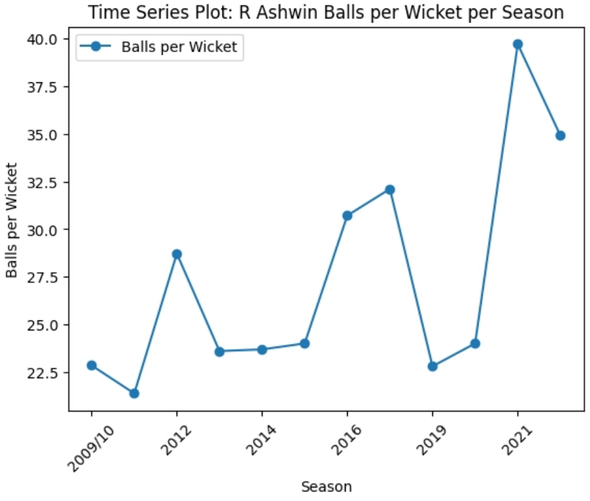

Histogram: Batting Strike Rate (minimum 200 balls faced, excluding leg byes, wide, and no balls)
Frequency
Batting Strike Rate


Histogram: Overall Batting Average (minimum 200 balls faced)
Frequency
Batting Average

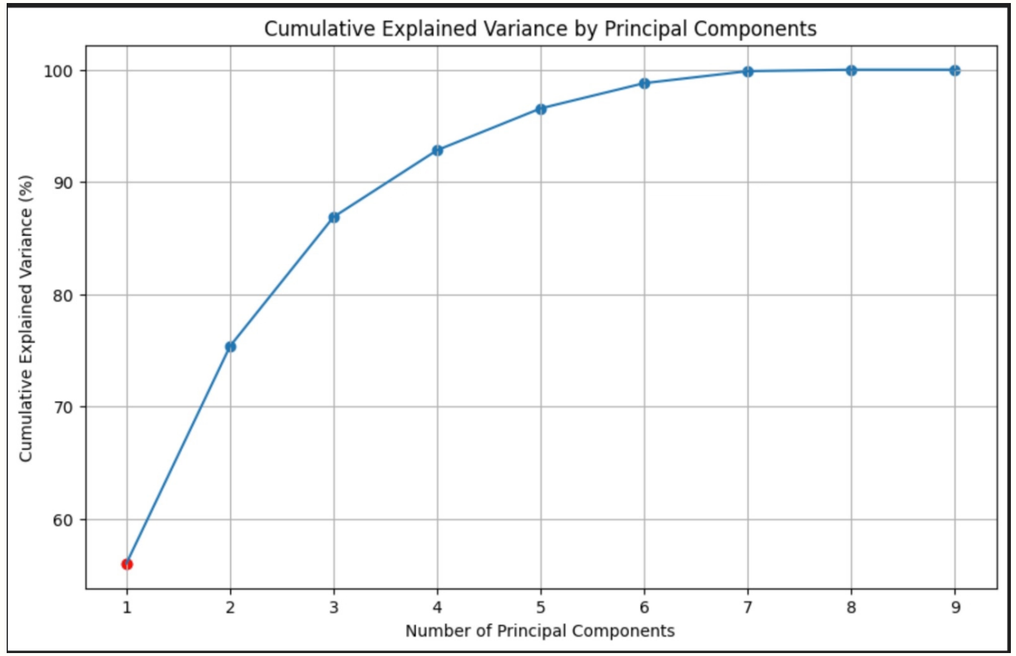

Relationship between Runs Scored Avg and Strike Rate
Batsmen
Line of best Fit
Max Avg: 39.78, Max SR: 135.52
Min Avg: 4.07, Min SR: 91.29
Strike Rate
Runs Scored Avg


Runs Scored by RG Sharma and V Kohli in Each Season
RG Sharma
V Kohli
Runs Scored
Season


Time Series Plot: R Ashwin Bowling Averages per Season
Bowling Average
Season


Time Series Plot: R Ashwin Balls per Wicket per Season
Balls per Wicket
Season

PCA



Principal Components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
CH Gayle	1.397111	4.690616	-0.235359	0.508584	0.449322	-0.344068	0.362841	0.104580	3.503520e-16
V Kohli	8.323528	-1.275992	-1.068945	0.197941	-0.330076	-0.063148	-0.057401	-0.097734	3.503520e-16
A Symonds	-2.162085	-1.550222	-0.700140	0.806282	1.836447	-0.581318	-0.242638	-0.090924	3.503520e-16
SK Raina	1.300175	-0.757902	1.798315	-0.226359	0.516492	0.838342	-0.437593	0.283384	3.503520e-16
SR Watson	-2.049509	1.551669	-1.563463	-1.816893	-0.102977	0.620176	-0.490022	-0.148242	3.503520e-16
RG Sharma	-0.240674	-0.366671	2.798067	-1.361994	-0.321777	-0.951703	0.156925	-0.102778	3.503520e-16
YK Pathan	-1.631477	-0.743767	0.668289	0.668595	-0.004736	1.097935	0.820065	-0.152665	3.503520e-16
ST Jayasuriya	-2.576108	0.346368	0.335978	1.821285	-1.393107	-0.175996	-0.565789	-0.033709	3.503520e-16
BA Stokes	-2.360962	-1.894097	-2.032743	-0.597440	-0.649588	-0.440221	0.453612	0.238088	3.503520e-16

Try Pitch

Truth Value

Bowling Points

Wicket Excluding Run Out	+25
Bonus (LBW / Bowled)	+8
3 Wicket Bonus	+4
4 Wicket Bonus	+8
5 Wicket Bonus	+16
Maiden Over	+12

Batting Points

Run	+1
Boundary Bonus	+1
Six Bonus	+2
30 Run Bonus	+4
Half-century Bonus	+8
Century Bonus	16
Dismissal for a duck Batter, Wicket-Keeper & All-Rounder	-2

Fielding Points

Catch	+8
3 Catch Bonus	+4
Stumping	+12
Run out (Direct hit)	+12
Run out (Not a direct hit)	+6

Other Points

Captain	2x
Vice-Captain	1.5x
In announced lineups	+4

Economy Rate Points

Min 2 Overs To Be Bowled

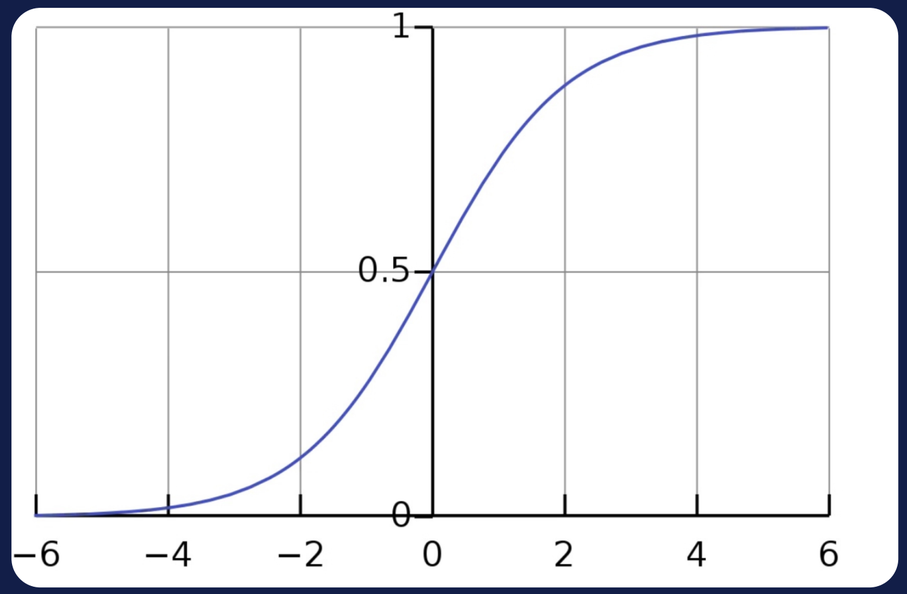
Below 5 runs per over	+6
Between 5-5.99 runs per over	+4
Between 6-7 runs per over	+2
Between 10-11 runs per over	-2
Between 11.01-12 runs per over	-4
Above 12 runs per over	-6

Truth Value

```
Final Team:                player          role  total_points  adjusted_points
0           RG Sharma       Captain         150.0          300.0
1         M Pathirana  Vice Captain         120.0          180.0
2          RD Gaikwad        Player          98.0           98.0
3              S Dube        Player          98.0           98.0
4           HH Pandya        Player          56.0           56.0
5         Tilak Varma        Player          46.0           46.0
6        Ishan Kishan        Player          42.0           42.0
7           R Ravindra        Player          35.0           35.0
8   Mustafizur Rahman        Player          31.0           31.0
9            MS Dhoni        Player          30.0           30.0
10          G Coetzee        Player          29.0           29.0
Total Team Points:  945.0
```

match_id	dream11_team
335982	['BB McCullum', 'AB Agarkar', 'SC Ganguly', 'AB Dinda', 'I Sharma', 'RT Ponting', 'JH Kallis', 'AA Noffke', 'P Kumar', 'Z Khan', 'LR Shukla']
335983	['MEK Hussey', 'JR Hopes', 'KC Sangakkara', 'IK Pathan', 'SK Raina', 'S Badrinath', 'Joginder Sharma', 'Yuvraj Singh', 'ML Hayden', 'P Amarnath', 'PA Patel']
335984	['G Gambhir', 'MF Maharoof', 'S Dhawan', 'R Bhatia', 'SR Watson', 'GD McGrath', 'RA Jadeja', 'DL Vettori', 'D Salunkhe', 'SK Warne', 'V Sehwag']
335985	['Z Khan', 'MV Boucher', 'ST Jayasuriya', 'RV Uthappa', 'B Akhil', 'Harbhajan Singh', 'AM Nayar', 'V Kohli', 'SM Pollock', 'R Dravid', 'LRPL Taylor']
335986	['M Kartik', 'DJ Hussey', 'WPUJC Vaas', 'PP Ojha', 'Mohammad Hafeez', 'AB Agarkar', 'A Symonds', 'I Sharma', 'AB Dinda', 'AC Gilchrist', 'RP Singh']
335987	['SR Watson', 'SK Warne', 'Yuvraj Singh', 'SK Trivedi', 'RA Jadeja', 'IK Pathan', 'PP Chawla', 'Kamran Akmal', 'JR Hopes', 'MM Patel', 'KC Sangakkara']
335988	['V Sehwag', 'RG Sharma', 'Mohammad Asif', 'R Bhatia', 'MF Maharoof', 'S Dhawan', 'RP Singh', 'VY Mahesh', 'G Gambhir', 'AC Gilchrist', 'WPUJC Vaas']

Logistic Regression



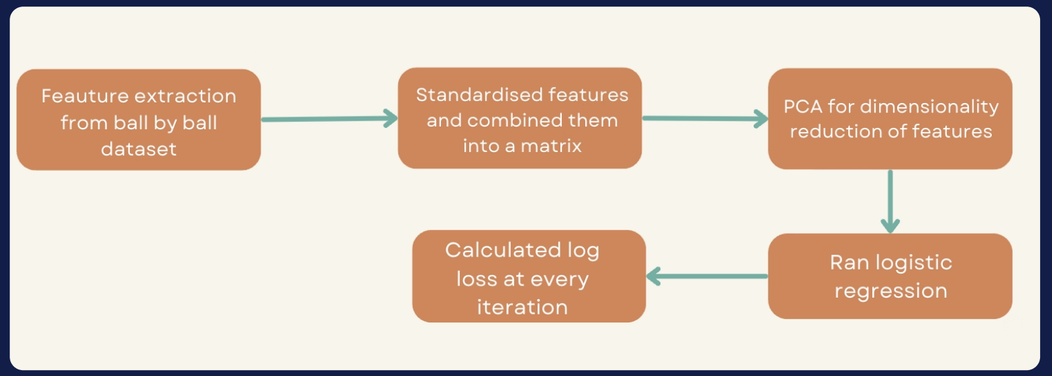
We ran logistic regression using all the features and using 5 Principal components which captured roughly 95% of the variance.

With PCA

```
Iteration 100, Loss: 0.6927
```

Without PCA

```
Iteration 100, Loss: 0.6857
```



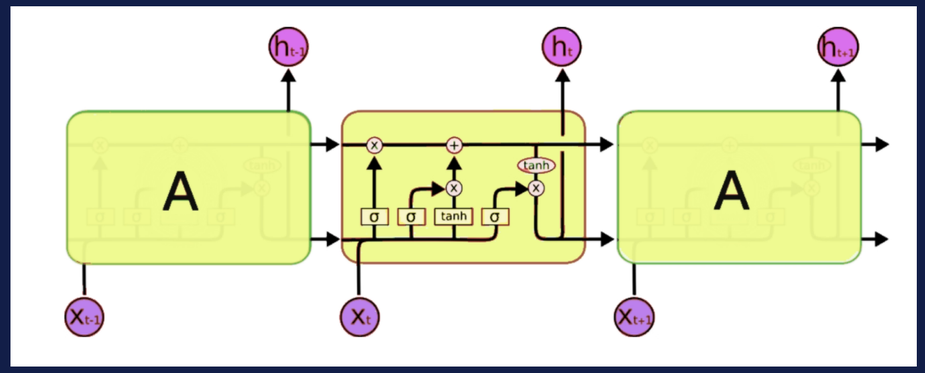
Log loss is a value between 0 and 1, 0 being the best and 1 being the worst. As we can see the value of log loss is quite high, therefore we would require a more complex model to solve the problem efficiently.

Note: Various combinations of hyperparamters were used but this was the best output we received. PCA didn't improve model performance in this case

Try Pitch

LSTM- Long Short Term Memory

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) that is well-suited for handling sequential data and capturing both short-term and long-term dependencies. It can effectively learn and remember patterns from recent time steps (short-term memory) while also selectively retaining and utilizing relevant information from farther back in the sequence (long-term memory).

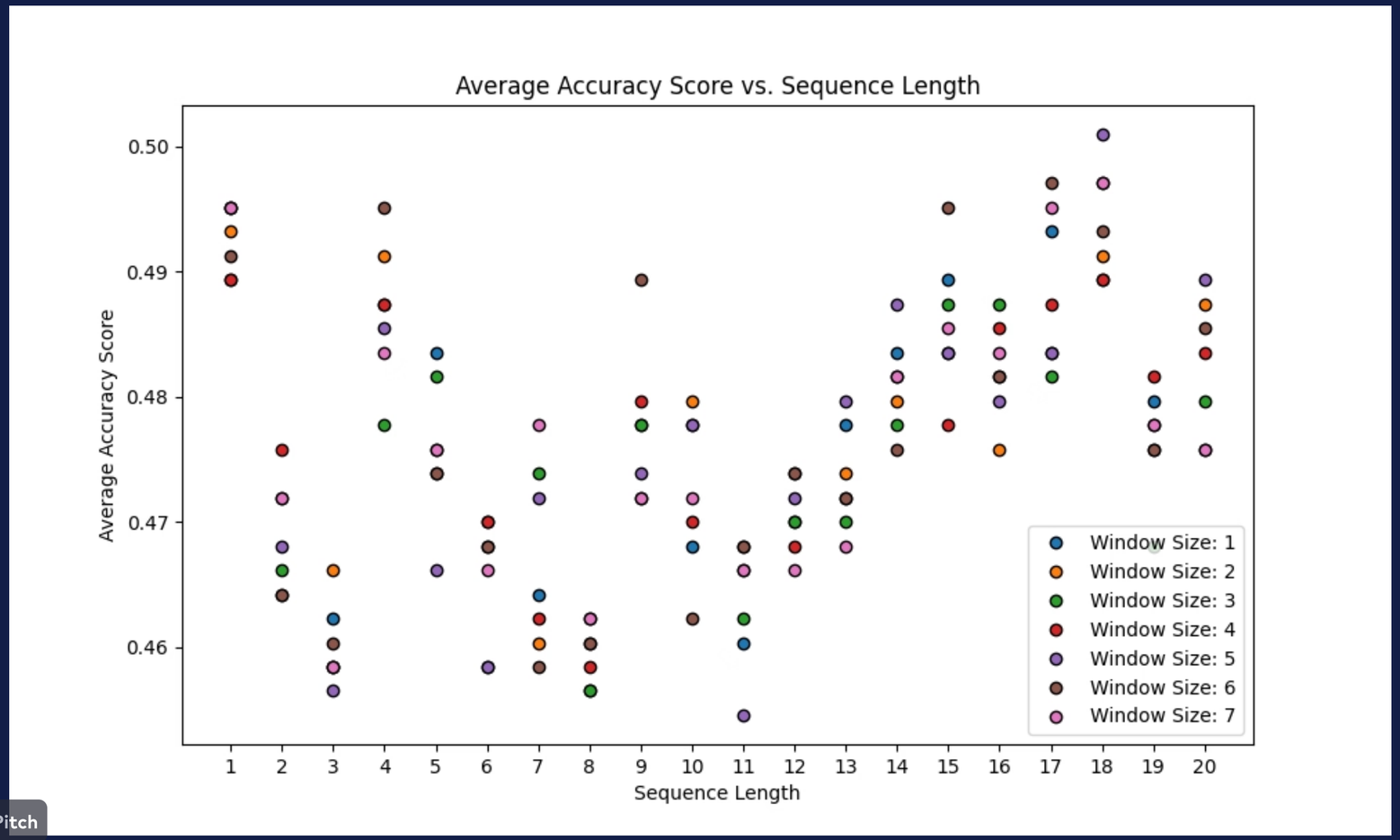


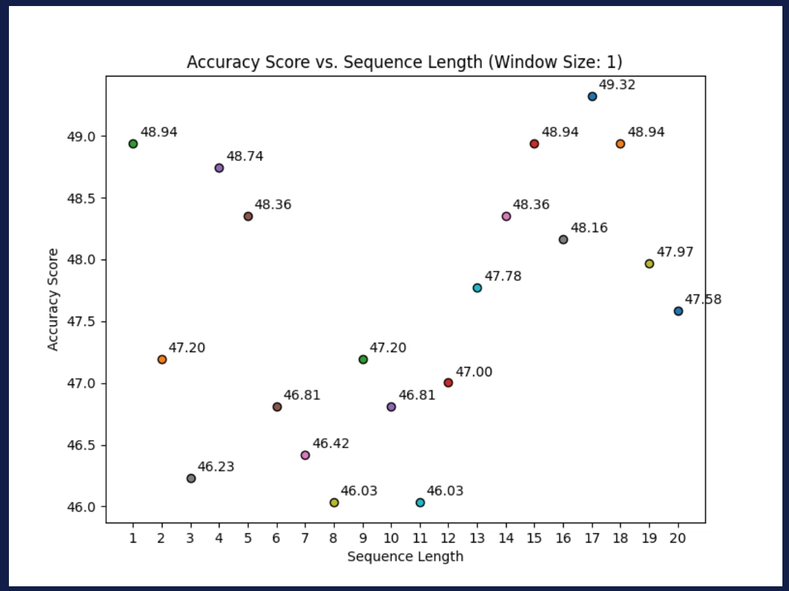
Using LSTM enabled us to give higher weightage to the current form of the player, which will help us generate a better team.

Try Pitch

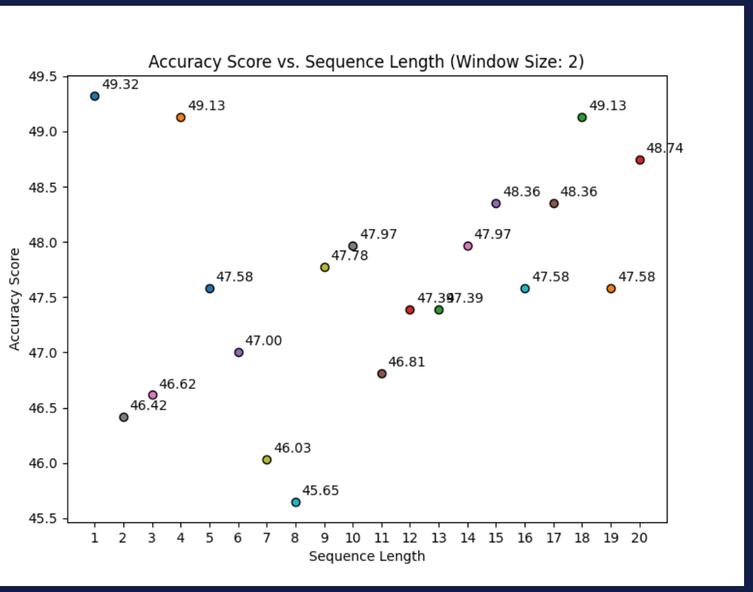
ML Methodologies

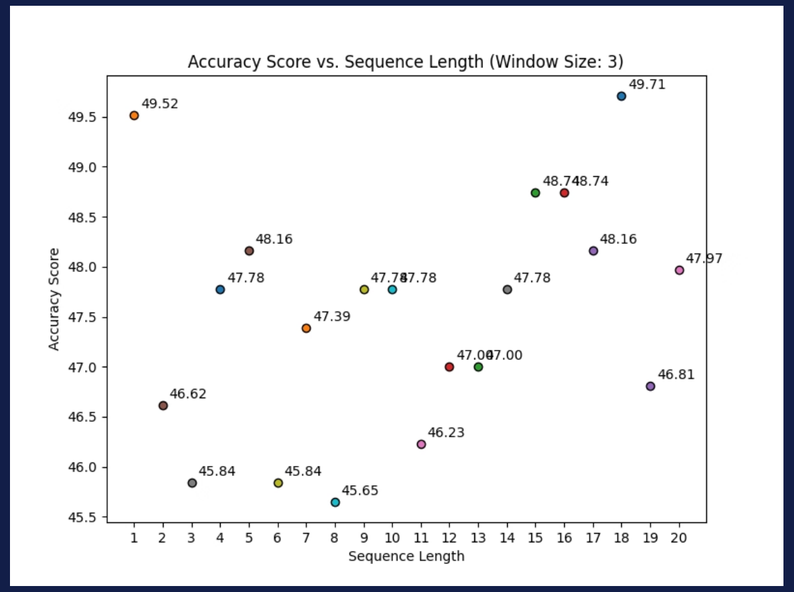
LSTM- Long Short Term Memory

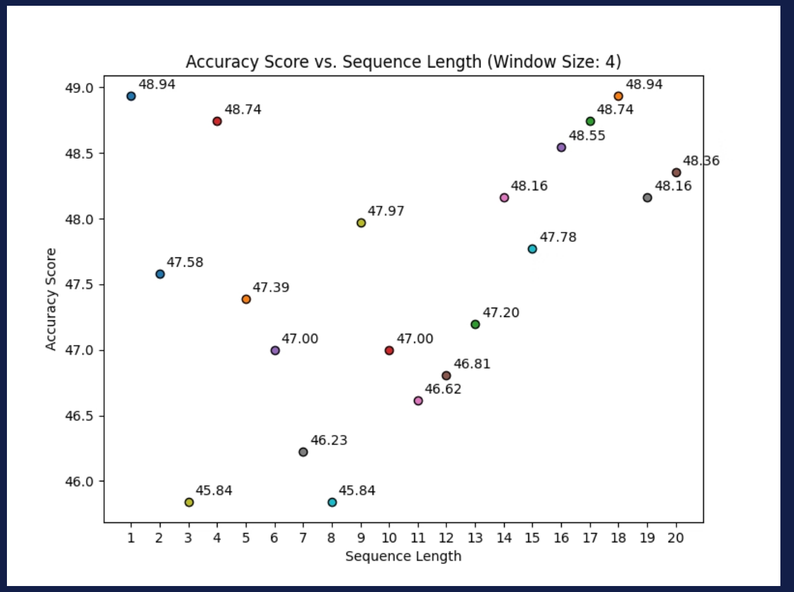


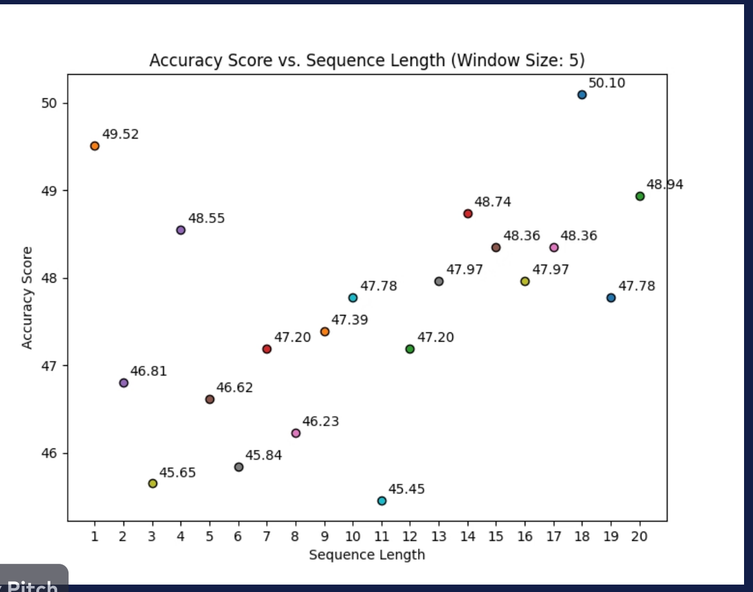


Try Pitch

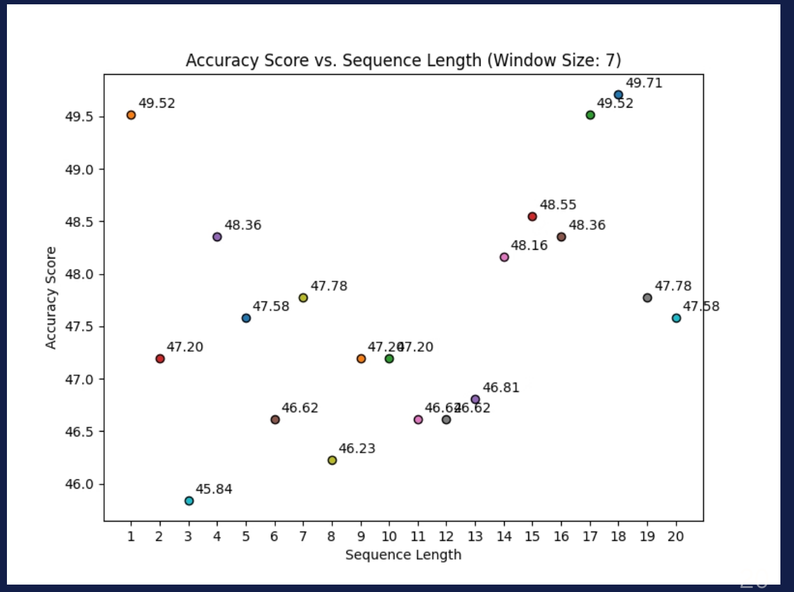

Accuracy Score vs. Sequence Length (Window Size: 2)
Accuracy Score
Sequence Length


Accuracy Score vs. Sequence Length (Window Size: 3)
Accuracy Score
Sequence Length


Accuracy Score vs. Sequence Length (Window Size: 4)
Accuracy Score
Sequence Length


Accuracy Score vs. Sequence Length (Window Size: 5)
Accuracy Score
Sequence Length


Accuracy Score vs. Sequence Length (Window Size: 6)
Accuracy Score
Sequence Length


Accuracy Score vs. Sequence Length (Window Size: 7)
Accuracy Score
Sequence Length

Try Pitch

ML Methodologies

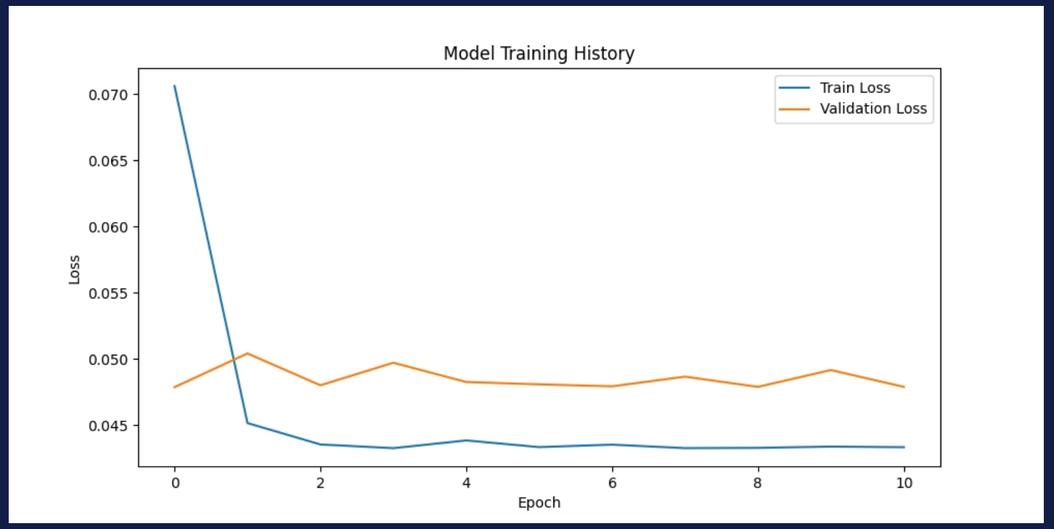
LSTM- Long Short Term Memory

$$\text{Accuracy} = \frac{\text{Number of Players Common Between Our Team and Dream Team}}{11}$$

- **Overall Algorithm Accuracy : 50.09%**
- **Best Accuracy: 81.82% [9/11 players]**

GT vs CSK
59th Match, IPL 2024
10th May 2024

Player Name	Fantasy Score
RD Gaikwad	0.35166809
B Sai Sudharsan	0.30831939
S Dube	0.302668035
Kartik Tyagi	0.296838075
Noor Ahmad	0.293360144
Shubman Gill	0.291186512
M Shahrukh Khan	0.274190277
TU Deshpande	0.273265511
MM Sharma	0.261645019
SN Thakur	0.255419314
DA Miller	0.245473266



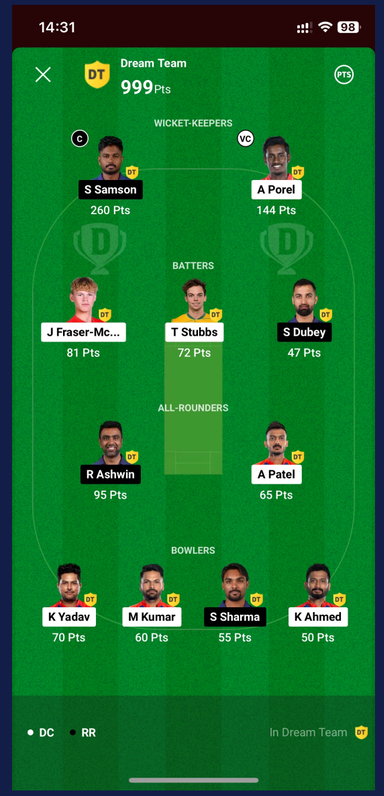
ML Methodologies

DC vs RR
56th Match, IPL 2024
7th May 2024

Accuracy for this match ≈ 54.5%

657 points

Our ML Team



999 points

Best Dream11 Team

Challenges

- Improve model performance. We are trying to implement LSTM which in the scope of our literature review no one seemed to have used. This makes it difficult to know how to tune the hyperparameters of our model. It requires a lot of hit and trial.
- Constant modifcation of the dataset & subsequent files to include recent matches.
- Cricket is generally not very predictable. Even if our model logically gives us the best possible team, it still may not be the actual best team of that match.
- We might run into some problems with Dream11 due to the nature of our project and their business model.

Try Pitch

Group 13
Hemant Gupta
Madhvendra Singh
Samarth Anand

Thank you.