iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XII **Month of publication:** December 2021

DOI: https://doi.org/10.22214/ijraset.2021.39354

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

Harnessing the Predictive Power of Lower-division Statistics of Cricketers to Predict Their Rates of Success at the International Level

Vishal C V [1], Sathvik K B [2], Nischay N [3], Manoj Athreya H [4], Sagar N [5]
[1]*Data Scientist, Mazo Solutions*
[2, 3, 4]*UG Scholar, Dept. of CSE, JSS Academy of Technical Education, Bengaluru, Karnataka*
[5]*UG Scholar, Dept. of CSE, Dayananda Sagar College of Engineering, Bengaluru, Karnataka*

Abstract: ***Statistics has always been an integral part of the sporting world. Selectors pick players based on numerous factors such as averages, strike-rates, runs scored or goals scored. Teams have exclusive 'talent hunters', who spend weeks, if not months, trying to uncover talent from different parts of the world. With the rise of this new niche field called Sports Analytics, teams can now perform player evaluations on tons of data that is available. This paper aims to examine the factors that truly indicate the capacity of cricket players to perform at the top-most level – international cricket. Though this research has been carried out on cricket data, it is hoped that similar methods can be used to hunt for true talent in other sports!***
Keywords: Cricket Analytics, Random Forest, Principal Component Analysis, Dimensionality Reduction.

I. INTRODUCTION

Strike-rates, averages, centuries, wickets and economy-rates form the crux of cricket statistics. As a matter of fact, domestic cricket, just like international cricket, is played in three different forms – First-class cricket, List A cricket and T20 domestic. What form of the game is most indicative? For a batter, what parameter matters more – strike-rate or average? Does a wrist-spinner's statistics carry special weight in contrast to that of a finger-spinner? The purpose of this research was to find answers to many such questions. Some outcomes were as expected, whereas others were surprising. This paper discusses the findings, and how these findings can affect the future of cricket and sports analytics.

II. PROBLEM STATEMENT AND OBJECTIVES

As the title suggests, the main objective of this research was to determine the statistical importance of various numeric and categorical parameters and in turn, harness their predictive power to forecast the rate of success of a player. Exploratory Data Analysis has also been performed in an attempt to detect relationships between various parameters – something that Machine Learning may not always do. Some questions that have been answered are:

1) What format of domestic cricket is the most powerful predictor of success at the highest level?
2) Does strike-rate really matter at the domestic level?
3) Do left-hand batters and bowlers have a statistical edge over right-handers?
4) Which Machine Learning Algorithm yields best results in predicting international level success and why?

III.LITERATURE SURVEY

A. Satyam Mukherjee. (2012). Quantifying individual performance in Cricket − A network analysis of Batsmen and Bowlers.

This paper [1] provides a revised approach for determining the 'quality' of runs scored by a batter or wickets taken by a bowler in this paper. We look at how Social Network Analysis (SNA) can be used to evaluate the effectiveness of team members. Using the player-vs-player information available for Test and ODI cricket, they have created a directed and weighted network of batters-bowlers and also a network of batters and bowlers based on batters' dismissal records throughout cricket's history. Their method might be used to evaluate a player's performance in domestic contests, paving the path for a more balanced team selection for international matchups but ours is a more streamlined approach for analysing the data.

B. Vipul Punjabi, Rohit Chaudhari, Devendra Pal, Kunal Nhavi, Nikhil Shimpi, Harshal Joshi. (2019). A survey on team selection in game of cricket using machine learning.

This study [2] tries to predict player performance, such as how many runs each batter will score and how many wickets each bowler will take for both teams. Both issues are characterised as classification problems, with the number of runs and wickets falling into separate ranges. This paper focused more on the venue aspect and how the players are going to perform there. Some more data is required to be fed to come to a better conclusion and some more analysis of domestic cricket will be helpful like we do.

C. Saikia, Hemanta & Bhattacharjee, Dibyojyoti & Krishnan, Unni. (2016). A New Model for Player Selection in Cricket. International Journal of Performance Analysis in Sport.

The paper [3] offers a metric that can be used to convert a cricketer's performance into a single numerical value that can be used to calculate the player's cricketing efficiency. The distributional pattern of the performance metric is determined and then used to determine the best performers in various fields of expertise. The selectors' task is made easier as a result of the exercise because they now have a reduced set of options to choose from. This paper has used many formulas to calculate the performance which cannot be told in general for all players as there are various factors that need to be considered.

D. Subramanian Rama Iyer, Ramesh Sharadha. (2009). Prediction of athlete's performance using neural networks: An application in cricket team selection.

The paper [4] uses neural networks to forecast each cricketer's future results based on their previous performance and then divide them into three categories: performers, moderates, and failures. Based on the rating that the player has received, the model will recommend if the player should be included in the squad or not. Our paper takes more data into consideration and gives further analysis of the players by taking even their domestic stats into consideration.

IV. RESEARCH METHODOLOGY

Data of over 500 players was web-scraped from sites such as ESPNCricinfo and Cricbuzz using Beautiful Soup, a popular Python library for web scraping.

These players were first stratified into four categories – wicket-keeper batters, batters, bowlers and all-rounders. Less than 10% of the players were wicket-keeper batters.

Also, analysis suggested that no wicket-keeper who exhibited poor battership has been successful. So, wicket-keeper batters were categorized as batters, making the total number of categories three. Any analysis and model building were performed on the categories individually, since each category comprised of different parameters. Batting oriented parameters such as batting strike-rate and number of runs were dropped for bowlers.

Economy rates and bowling averages were dropped for batters. To ensure the significance of any outcome or result, players who played in less than 40 international matches were dropped from each of the categories. The final dataset consisted data of 148 batters, 128 bowlers and 85 all-rounders.

A. Data Pre-processing and Analysis

After the classification of data, null values had to be handled. T20 and franchise is relatively new to the sport. In fact, from the inception of T20 cricket in 2004, up till 2010, less than 130 T20 International Matches had been played. By 2020, however, more than 1000 T20 matches had been played. In fact, many top cricketers have not even played T20 cricket. This is a case of MNAR (Missing not at random). To replace null values in the T20 average and strike rate columns, the Iterative Imputer was used. Sklearn's Iterative Imputer models zero-null features as a function of features with null values, in turn computing values that replace null values.

Multiple imputation ensures that all features have been taken into consideration in the computation of a value of the given variable. In computing T20 strike-rates, considering the year of debut becomes as important as any other substantial factor. Majority of players who debuted before 2007 have strike-rates under 80 whereas majority of the players who debuted after 2007 have strike-rates over 80. One of main reasons for this change in trends is due to the introduction of T20 cricket in 2006, which changed the face of the game. Not only did batters start scoring at a high SR in T20 cricket, this change was brought about in ODIs as well as in the domestic equivalent of ODIs, List A cricket.

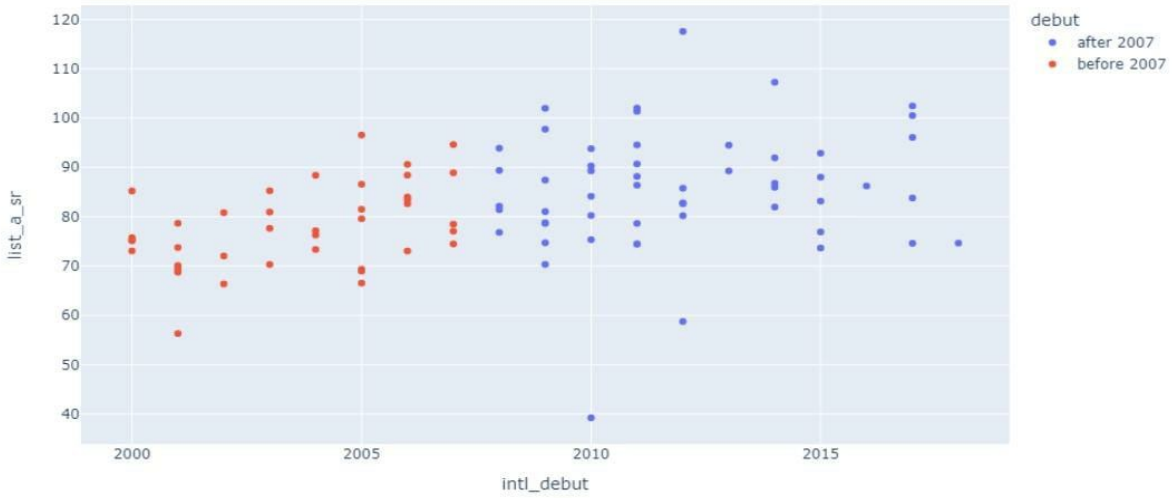


Figure 1 Notice the change of trends in List A Strike rate of batters before and after 2007

T20 averages and strike rates shared some sort of a linear relationship with other parameters such as first-class and list-a averages and strike-rates. Hence, the Bayesian Ridge form of multiple imputation was used.

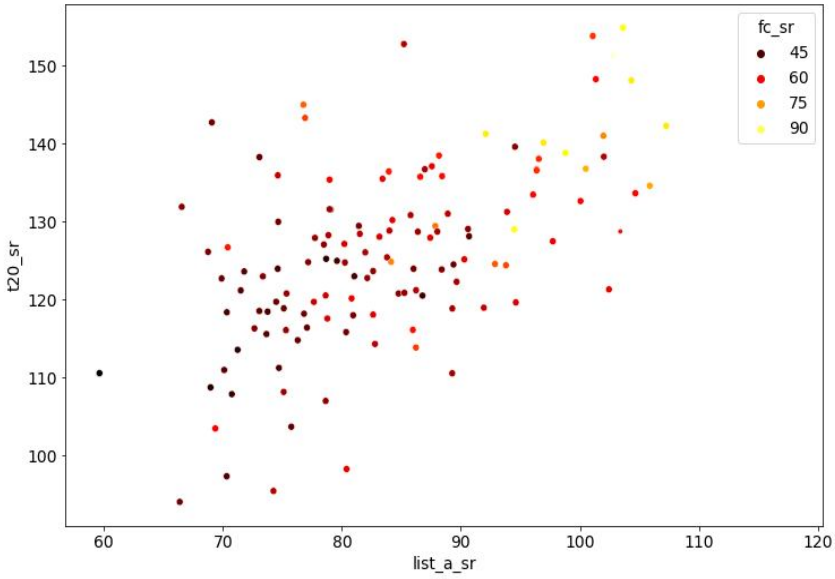


Figure 2 Scatter-plot exhibiting relationships between T20, FC and List-A strike-rates

Bayesian Ridge is a version of linear regression where in point estimates are replaced by probability distributors (i.e. y is not an exclusive output, but is rather drawn from a Gaussian Distribution).

$$Posterior = \frac{Likelihood * Prior}{Normalization}$$

Figure 3 Simple expression for Bayesian ridge/linear regression

After multiple imputation of missing data, Exploratory Data Analysis was carried out, whose results have been discussed in the next section.

In order to prepare the data for model building, it was scaled using the standard scalar, given the symmetric distribution of data.

$$x_{new} = \frac{x - \mu}{\sigma}$$

Figure 4 standard scaler is computed by subtracting the mean of all observations from x and dividing the resultant by the standard-deviation of all observations

Models were fit on each of the three data-frames, in an attempt to accurately predict the success rate of an international cricketer. The number of player of the match awards as a percentage of international matches played was taken as the metric for success-rate. F(x) was the man of the match percentage, referred throughout this paper as 'motm_perc'.
The next section is divided into four parts – Impact of multiple imputation, Exploratory Data Analysis, Models' performance before Principal Component Analysis, and Models' performance after Principal Component Analysis.

V. RESULTS AND INTERPRETATION

A. *Impact of multiple imputation*

The foremost question after the pre-processing of data was 'How well did the iterative imputer perform?' The only way to judge is by examining the computed values. Given below is the data for Indian Cricketer VVS Laxman. VVS Laxman played no T20 matches before his international debut. Hence, his T20 strike-rate and average are null.

	player	country	batting_arm	years_of_domestic_before_debut	intl_debut	total_intl_matches	fc_avg	list_a_avg	t20_avg	fc_sr	list_a_sr	t20_sr	n
45	VVS Laxman	india	right	3	1996	223	45.97	30.76	NaN	49.37	71.23	NaN	

Figure 5 Data Before Multiple Imputation

The data after multiple imputation is shown below.

	years_of_domestic_before_debut	total_intl_matches	fc_avg	list_a_avg	t20_avg	fc_sr	list_a_sr	t20_sr
71	3.0	223.0	45.97	30.76	27.38655	49.37	71.23	119.620309

Figure 6 Data After Multiple Imputation

Had VVS played a decent bit of T20 cricket before his international debut, he would have averaged around 27 and struck at 119. Now, how believable are these numbers? Ask cricket fans from the late 90s and early 2000s, they are going to tell you it is quite believable! This shows the strength of statistical methods such as multiple imputation. When enough original data is available, it always makes sense to prefer multiple imputation over simple imputation.

B. *Exploratory Data Analysis*

Figure 7 exhibits the upward trend of list a strike-rates as time progressed. But what effect did these have on the 'motm_perc'?

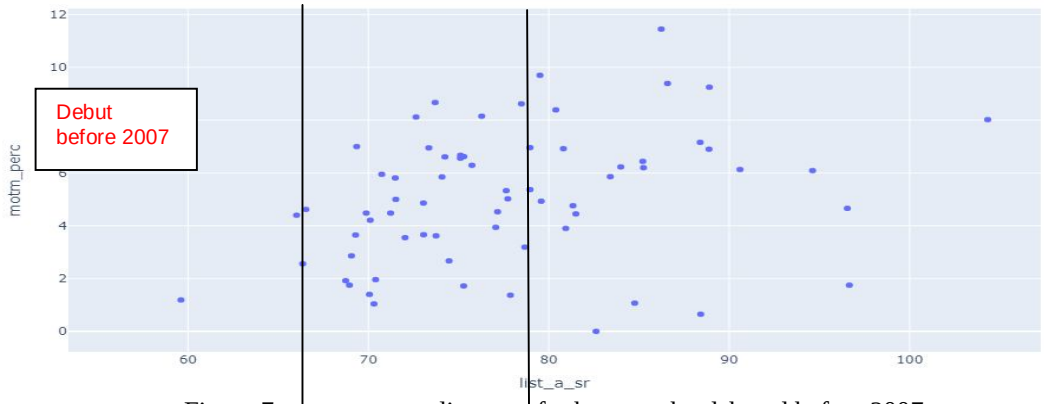


Figure 7 motm_perc vs list_a_sr for batters who debuted before 2007

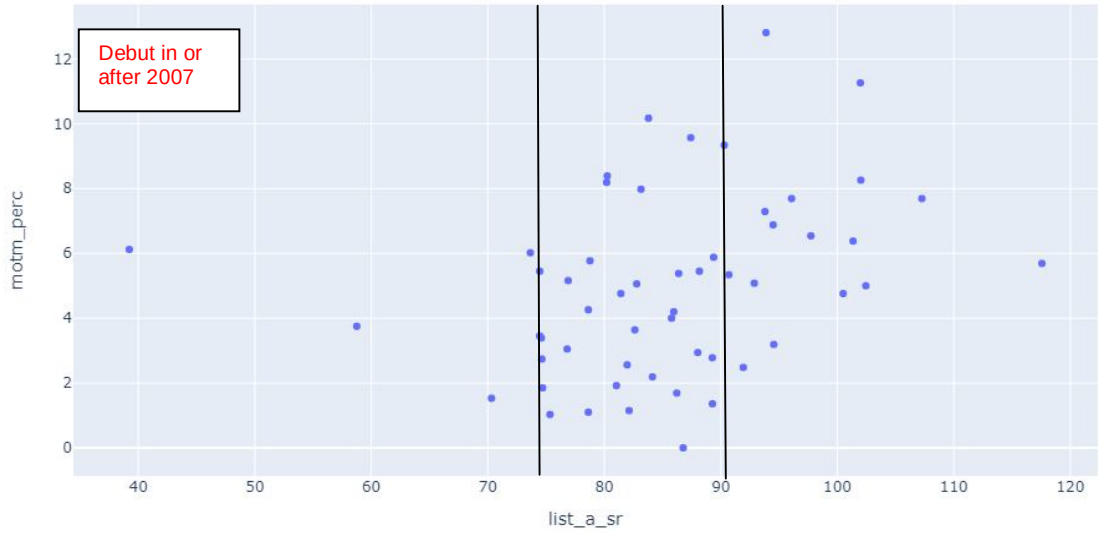


Figure 8 motm_perc vs list_a_sr for batters who debuted in or after 2007

Figure 7 shows that for batters who debuted before 2007, the average 4 to 8 motm_perc mark was in the 70-85 strike-rate range. Figure 8 shows that this motm_perc marked moved about 15 strike-rates units to the right! That is incredible! This indicates that the bench-mark 50-over --cricket scores went up by about 45 runs in the post T20 era.
Enough about batters! Here is something for the bowling fans.

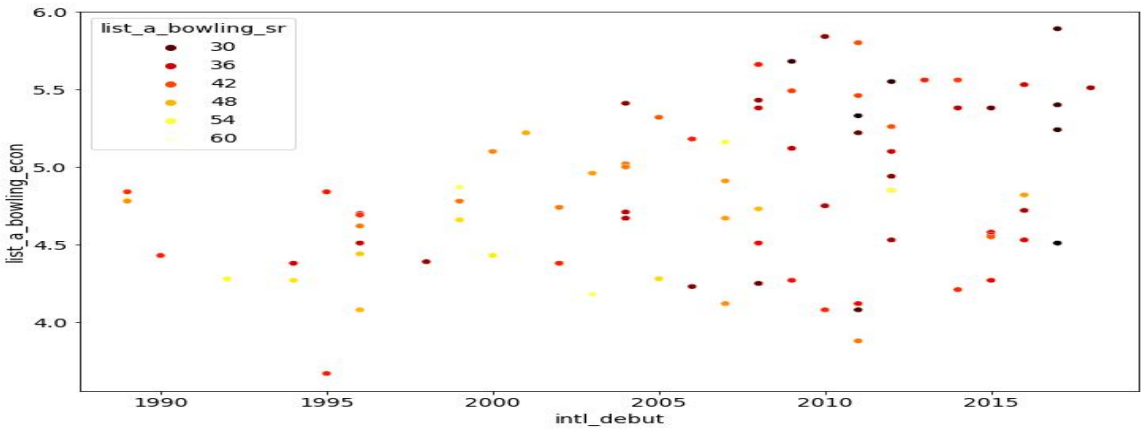


Figure 9 list_a_economy vs year of debut (points colour encoded with respect to bowling s_r)

```
In [577]: sum(df[df['intl_debut']<=2006]['list_a_bowling_econ'])/len(df[df['intl_debut']<=2006]['list_a_bowling_econ'])
Out[577]: 4.653055555555556

In [578]: sum(df[df['intl_debut']>2006]['list_a_bowling_econ'])/len(df[df['intl_debut']>2006]['list_a_bowling_econ'])
Out[578]: 4.9636734693877544
```

Figure 10 average economy rates for players who debuted in or before 2006 and players who debuted after 2006

Figures 9 and 10, together, clearly indicate that the mean and median economy rates started climbing with time. But what is interesting is that post 2006, teams started to prefer bowlers with lower strike rates (bowlers who could pick up wickets more frequently). This means that teams didn't mind bowling expensive bowlers, as long as they could bowl out the opposition quickly.

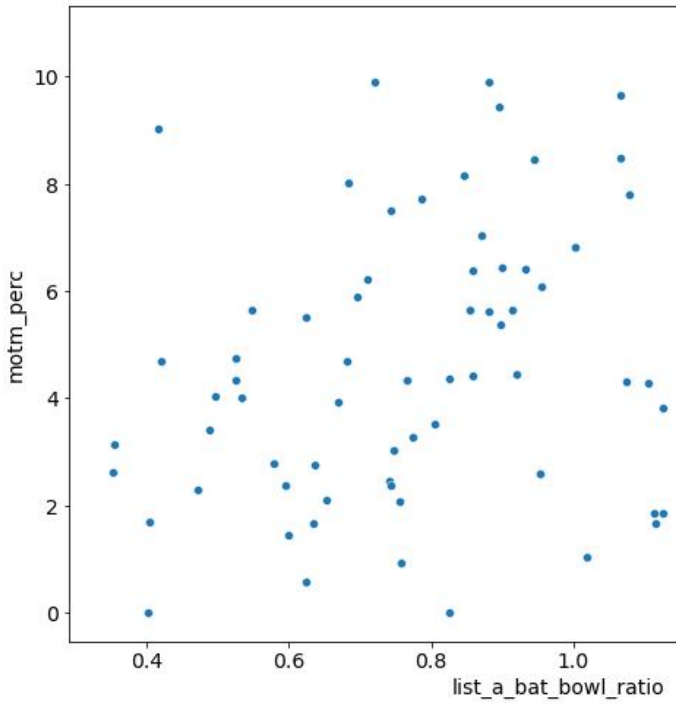


Figure 11 List_a_bat_bowl_ratio vs motm_perc

Figure 11 describes the relationship between the golden ratio of cricket and motm_perc. The golden ratio is the ratio between the batting average and bowling average of an all-rounder. It is a general belief that a ratio closer to 1 is indicative of a better all-rounder. The graph above seems to validate the belief!

C. Before PCA

As the figure below suggests, motm_perc shared no strong linear relationship with other parameters.

```
In [425]: imputed_df.corr()['motm_perc']

Out[425]: years_of_domestic_before_debut    0.005190
          total_intl_matches                0.470409
          fc_avg                            0.462487
          list_a_avg                        0.433977
          t20_avg                           0.365262
          fc_sr                             0.211951
          list_a_sr                         0.201142
          t20_sr                            0.155761
          motm_perc                         1.000000
          country_australia                 0.085146
          country_bangladesh               -0.108348
          country_england                   0.147675
          country_india                     0.118473
          country_new zealand              -0.011610
          country_pakistan                  0.000922
          country_south africa              0.012267
          country_sri lanka                -0.112443
          country_west indies              -0.176039
          batting_arm_left                  0.057784
          batting_arm_right                -0.057784
          Name: motm_perc, dtype: float64
```

Figure 12 Relationship of motm_perc with other parameters

As expected, statsmodel's GLM did not yield favorable results even after filtering of independent variables based on p-values and variation inflation factors (VIF). Though the R-square score was nearly 70% for the training data, it fell to a feeble 23% when the linear regression model was fit on the testing data. Similar problems persisted for all three data-frames – Batters, Bowlers and All-rounders.

Feature Engineering, Recursive Feature Elimination and penalizing the linear model (Lasso and Ridge regression) did not mitigate the overfitting. So, a Random Forest model was tried on the data.

```
]: rf = RandomForestRegressor(random_state=42)
params = {
    'max_depth': [2,5,7],
    'min_samples_leaf': [5,10,20,50,100],
    'n_estimators': [3,5,7,9,11, 13, 15, 17]
}
grid_search_1 = GridSearchCV(estimator=rf,
                             param_grid=params,
                             cv = 3,
                             n_jobs=-1 , verbose=1, scoring="neg_mean_absolute_error")
grid_result_1 = grid_search_1.fit(X_train, y_train)

Fitting 3 folds for each of 120 candidates, totalling 360 fits
```

Figure 13 Random Forest model code

The metrics used to evaluate the performances of the Random Forest models were Negative Mean Absolute Error (NMAE) and Mean Absolute Percentage Error (MAPE).

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Figure 14 MAE is the summation of the absolute difference between the predicted and true values, unit number of observations

Mean Absolute Percentage Error is similar to MAE, but in percentage form.
GridSearchCV identified that the best performing model had a NMAE of -1.829 and MAPE of 65.43%, on the test data. The NMAE suggests that the predicted motm_perc on an average, was off by 1.8, which might not seem too bad. In simpler terms, it suggests that the model may predict that a batters will win you 10 in 100 matches, when he actually might win you anywhere between 8 and 12 matches. However, the MAPE suggests that the error rate is over 60%, which may not be that bad for this sort of a scenario, but can be better.
Before applying Principal Component Analysis (PCA) on the data-frame, the important features and the extent of their importance was visualized using Random Forest's feature_importances.
The importance of features is calculated based on node impurities. A node probability is computed by weighing the impurity of a given node against the probability of a tree reaching that node. Higher the node probability, higher the importance.

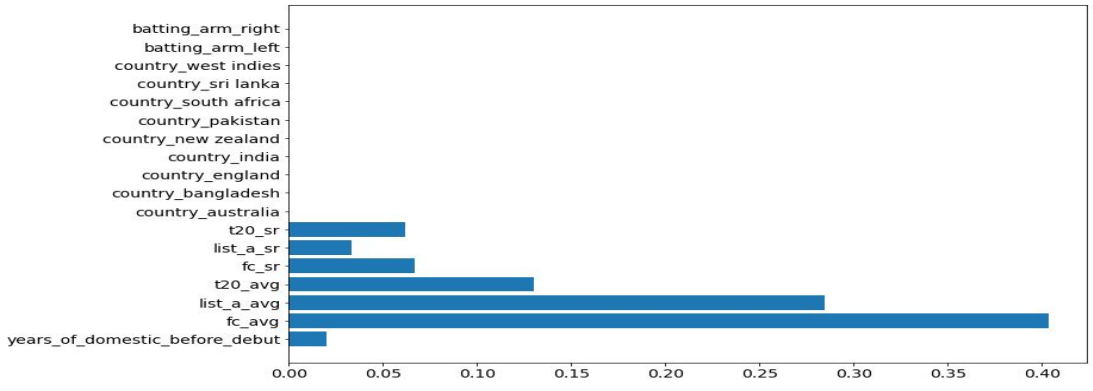


Figure 15 Feature importance (Feature vs node probability)

It is clear that some features are more important than others, in predicting the international success of the batters. Coaches are probably right - a good batter invariably does well in first-class cricket!

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

D. After PCA

PCA was applied on the data-frame. PCA transforms the dataset onto a lower dimensionality subspace. The aim is to absorb as much as information as possible from as fewer parameters as possible. Linear transformations cause a change in the values of the data.

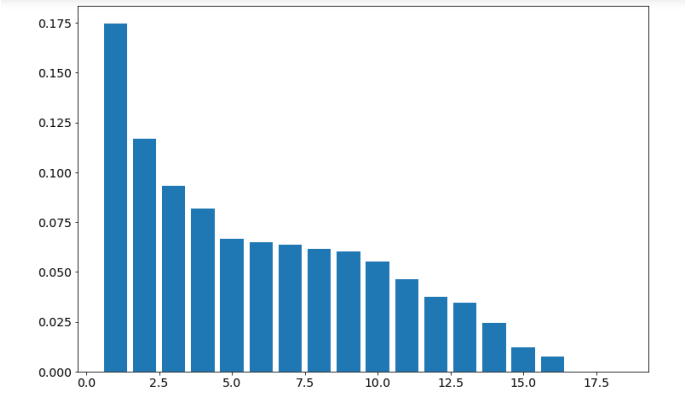


Figure 16 Variance Ratio vs Feature

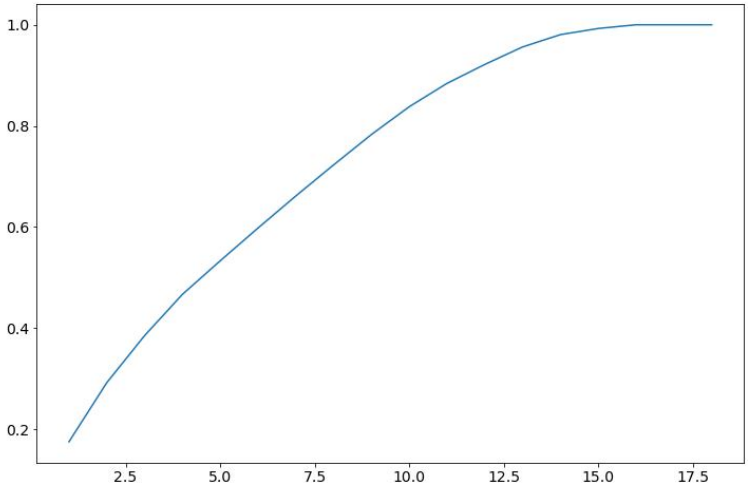


Figure 17 cumulative variance ratio vs number of features

PCA suggested that 12 variables captured more than 90% of the data.

```
pc2 = PCA(n_components=12 , random_state =42)
newdata=pc2.fit_transform(X)
col= []
for i in range(1,13):
    col.append('PC'+str(i))
df_2 = pd.DataFrame(newdata , columns = col)
df_2
```

Figure 18 PCA

On fitting a Random Forest model on this data, the results significantly improved. MAPE of test data was now only 22%. Observe the predicted values before and after PCA.

	motm_perc	predicted	change	pred_after_pca	change after pca
125	1.75	1.932696	0.182696	2.454899	0.704899
51	6.96	7.744301	0.784301	7.106342	0.146342
139	7.25	3.646846	3.603154	9.251233	2.001233
19	1.27	3.977911	2.707911	1.933995	0.663995
104	8.19	6.506323	1.683677	9.133441	0.943441
12	8.42	6.622569	1.797431	10.333920	1.913920
76	4.45	4.936973	0.486973	3.904990	0.545010
31	5.16	4.900926	0.259074	6.144462	0.984462
81	7.00	4.465660	2.534340	7.013344	0.013344
9	2.38	5.839898	3.459898	1.344211	1.035789
26	7.29	6.766169	0.523831	7.009129	0.280871
96	2.94	6.161555	3.221555	2.794222	0.145778
144	7.69	5.276823	2.413177	7.781205	0.091205
67	4.86	3.464934	1.395066	5.091234	0.231234
135	3.62	6.046430	2.426430	4.620988	1.000988
66	5.00	6.675433	1.675433	7.020311	2.020311
18	6.72	7.508529	0.788529	8.934758	2.214758
69	3.75	3.910522	0.160522	3.453504	0.296496
124	5.86	7.728582	1.868582	7.129390	1.269390
30	1.69	6.011795	4.321795	4.349232	2.659232

Figure 19 After fitting a Random Forest model on data

A similar approach was directed towards the bowler and all-rounder data-frames.

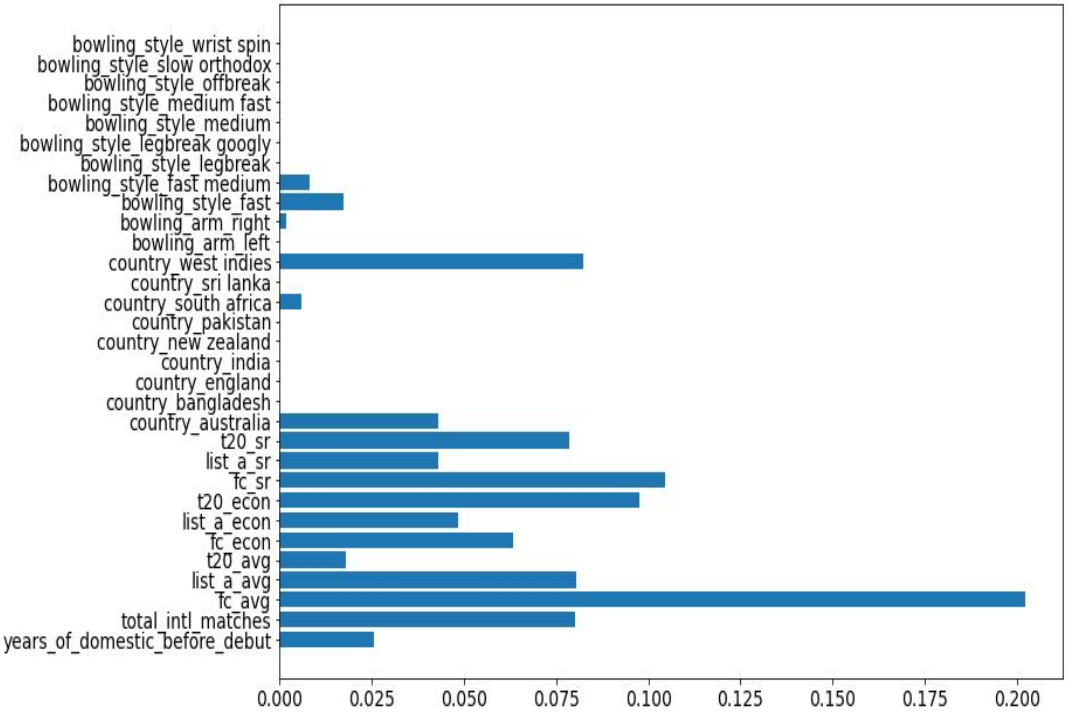


Figure 20 Feature importance for bowlers

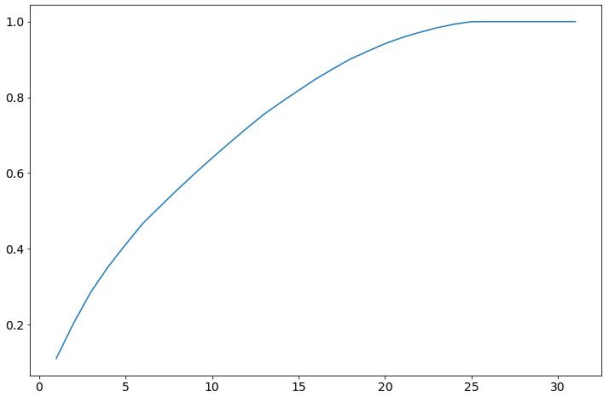


Figure 21 PCA on bowler's dataset

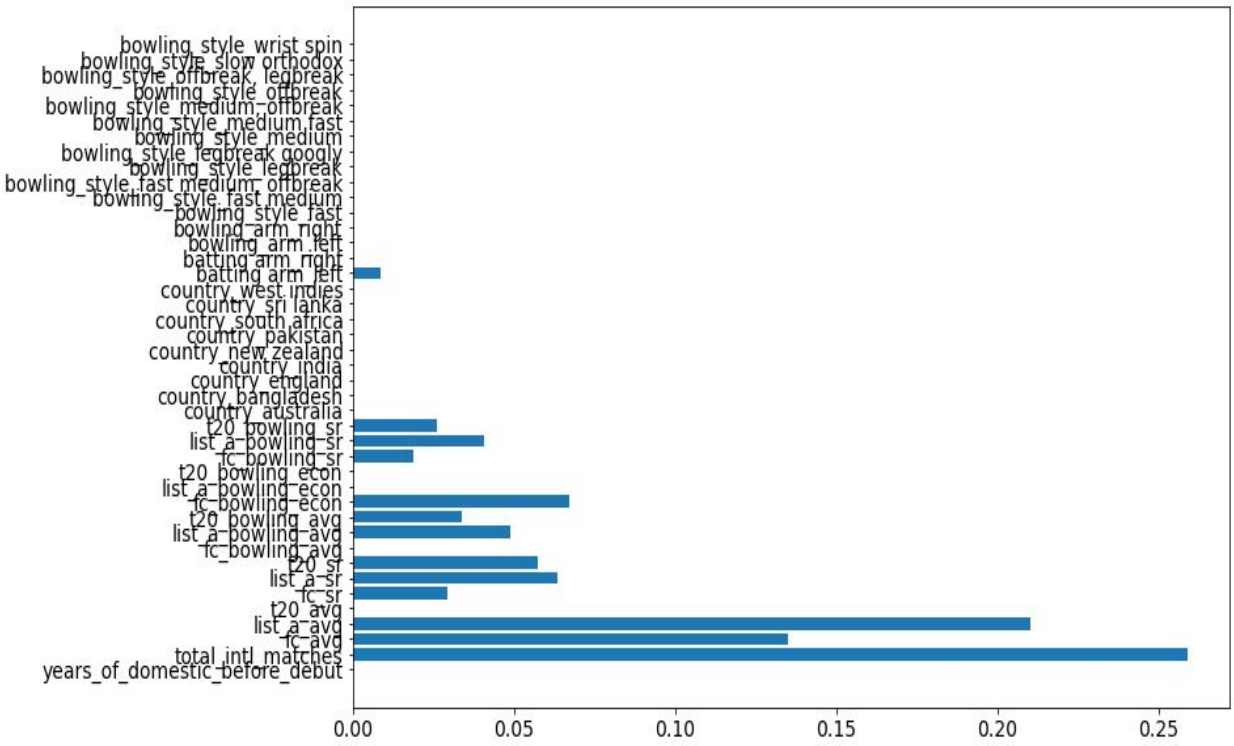


Figure 22 Feature importance for all-rounders

Figure 22 suggests that all-rounders are more credited for their experience followed by batting skills, bowling economy and strike-rate. In fact, a player with a lower economy-rate is in more demand!

TABLE I – Positive effect of PCA

Category (in increasing order of dimensionality)	Random Forest MAPE before PCA	Random Forest MAPE after PCA
Batters	0.64	0.22
Bowlers	0.71	0.38
All-rounders	0.91	0.41

The table above exemplifies the positive effect of PCA, especially when working with high dimensionality datasets. Reduction in dimensionality and capturing of relevant information is the key to obtaining stronger models.

VI. CONCLUSIONS

Currently, Sports Analytics is sparse in India but it is definitely the future. The introduction of T20 format has definitely changed the face of cricket. With limited time and resources, were able to predict the success rate of a player in international cricket just by using the players' domestic stats. Imagine what specialists could do with extensive data such as ball-by-ball and match-by-match stats, the performance of other players in the same match, venue of the matches, weather conditions and various other factors. Such is the power of Sports Analytics. Not only cricket, Analytics could do wonders in other sports as well.

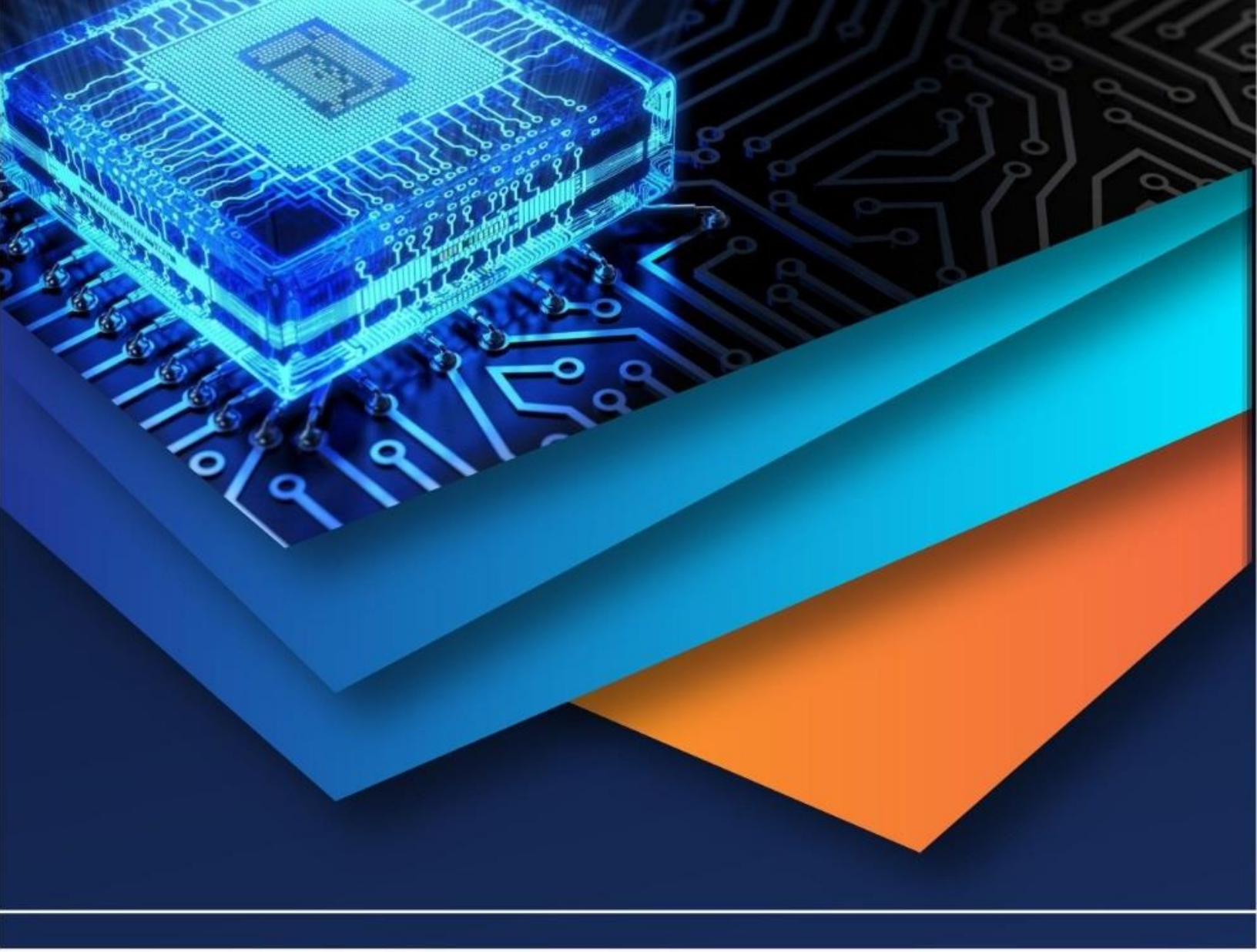
REFERENCES

[1] Satyam Mukherjee, Quantifying individual performance in Cricket — A network analysis of batsmen and bowlers, Physica A: Statistical Mechanics and its Applications, Volume 393, 2014, Pages 624-637, ISSN 0378-4371, https://doi.org/10.1016/j.physa.2013.09.027.

[2] Vipul Punjabi, Rohit Chaudhari, Devendra Pal, Kunal Nhavi, Nikhil Shimpi, Harshal Joshi. (2019). A survey on team selection in game of cricket using machine learning, International Research Journal of Engineering and Technology.

[3] Hemanta Saikia, Dibyojyoti Bhattacharjee & Unni Krishnan Radhakrishnan (2016) A New Model for Player Selection in Cricket, International Journal of Performance Analysis in Sport, 16:1, 373-388, DOI: 10.1080/24748668.2016.11868893

[4] Subramanian Rama Iyer, Ramesh Sharda, Prediction of athlete's performance using neural networks: An application in cricket team selection, Expert Systems with Applications, Volume 36, Issue 3, Part 1, 2009, Pages 5510-5522, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2008.06.088.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY