Summary Report: Analysis of Indian team's batting performance

Sai Niketh (2020MT60895)
Sai Kiran (2020MT60889)

November 4, 2024

1 Introduction

In this analysis, we investigate India's batting performance in the **1st innings** and **2nd innings** to identify distinct trends based on the order of innings. The dataset spans from 2000 to 2010, allowing us to analyze each scenario separately before combining the results to assess overall performance. This approach provides a comprehensive view of India's batting capabilities throughout the decade.

2 Data Preprocessing

- Converted runs/wickets to two different columns by extracting runs and wickets separately.
- Converted win/loss to binary variable 1/0.
- Converted overs played into balls faced as overs are represented in base-6(senary) system.

3 Exploratory Data Analysis

- **Correlation Heatmaps**: The correlation heatmaps indicate a positive correlation between Runs and RPO, while Runs and RPO negatively correlate with Wickets. Furthermore, the Result negatively correlates with Wickets, as losing all 10 wickets typically leads to a match loss.
- From scree plots and histograms, we can observe that

  The team often brings the game down to the final 10 overs while chasing.

  While chasing, the team often loses more than 4 wickets.

  In the first innings, a team tends to score more and win matches when they lose fewer wickets. Conversely, losing more wickets generally results in scoring fewer runs and losing the match.

4 Fitting Multivariate Distribution

- The Q-Q plots show that Score, RPO, Wides, and Noballs nearly follow a normal distribution.
- From the Q-Q plots of Mahalanobis distance, we can identify the outliers (D^2>19.03 (at 2.5% threshold)), and we found 10 outliers in each dataset.

5 Covariance and Mean Comparison

- 5% significance level is used for all testing purposes.
- F-test is used to test the equivalence of covariance between two datasets; as we fail to reject, we proceed to test for equivalence of mean vectors using T-test.

Test	Statistic Value	Critical Value	Conclusion
Covariance Testing	$F = 0.71667369$	1.33042525	Fail to Reject Null Hypothesis
Mean Testing	$F = 10.13441376$	1.91627638	Reject Null Hypothesis

Table 1: Statistical Values for Covariance and Mean Testing

6 Profile Analysis

Test-1: Test for Parallel Profiles

Null Hypothesis:

$$H_0 : \mu_{1(k)} - \mu_{2(k)} = \mu_{1(k-1)} - \mu_{2(k-1)} \quad \text{for } k = 2, 3, \ldots, p$$

Alternative Hypothesis:

$$H_a : H_0 \text{ is not true}$$

The null hypothesis can be rewritten as:

$$C(\mu_1) = C(\mu_2)$$

where C is a $(p-1) \times p$ matrix defined as:

$$C(i,j) = \begin{cases} 1 & \text{if } i = j \\ -1 & \text{if } j = i - 1 \\ 0 & \text{otherwise} \end{cases}$$

Statistic	Value
T^2 Statistic	94.0205
F-Statistic	11.4433
F-Critical Value	1.9743
Conclusion	Reject the null hypothesis that profiles are parallel

Table 2: Results of the Profile Analysis Hypothesis Test

Based on the results, we can conclude that the assumption of parallel profiles doesn't hold, and significant differences exist.
As the null hypothesis that profiles are parallel is rejected, we won't test for coincident and equality of profiles.

7 MANOVA

7.1 MANOVA Across Datasets Using Dataset as Group

We conducted a MANOVA with `Group` (indicating dataset `df1` or `df2`) as the independent variable. The results are presented in Table 5 below.

Statistic	Value	Num DF	Den DF	F Value	Pr >F
Intercept					
Wilks' Lambda	0.0013	8	259	25090.6352	0.0000
Pillai's Trace	0.9987	8	259	25090.6352	0.0000
Hotelling-Lawley Trace	775.0003	8	259	25090.6352	0.0000
Roy's Greatest Root	775.0003	8	259	25090.6352	0.0000
Dataset					
Wilks' Lambda	0.7427	8	259	11.2170	0.0000
Pillai's Trace	0.2573	8	259	11.2170	0.0000
Hotelling-Lawley Trace	0.3465	8	259	11.2170	0.0000
Roy's Greatest Root	0.3465	8	259	11.2170	0.0000

Table 3: MANOVA Results Across Datasets Using Dataset as Group

7.2 MANOVA Across Datasets Using Result as Group

We also performed MANOVA with `Result` as the independent variable. This allows us to test if there is a difference in the multivariate means based on the outcome of the match. The results are presented in Table 6 below.

Statistic	Value	Num DF	Den DF	F Value	(Pr >F)
Intercept					
Wilks' Lambda	0.0014	8	259	23645.7964	0.0000
Pillai's Trace	0.9986	8	259	23645.7964	0.0000
Hotelling-Lawley Trace	730.3721	8	259	23645.7964	0.0000
Roy's Greatest Root	730.3721	8	259	23645.7964	0.0000
Result					
Wilks' Lambda	0.4973	8	259	32.7287	0.0000
Pillai's Trace	0.5027	8	259	32.7287	0.0000
Hotelling-Lawley Trace	1.0109	8	259	32.7287	0.0000
Roy's Greatest Root	1.0109	8	259	32.7287	0.0000

Table 4: MANOVA Results Across Datasets Using Result as Group

7.3 Interpretation

The p-values for both tests are less than 0.05, which suggests that there is a statistically significant difference in the multivariate means across both `Group` and `Result`. This implies that the variables differ significantly based on the dataset and match outcome, justifying further investigation into specific group differences.

8 Principal Component Analysis

- **PC0**: **Score** (0.6056) and **RPO** (0.5278) have the highest positive loadings, indicating that higher scores and run rates contribute significantly to this component. This suggests that PC0 captures the overall scoring efficiency or success regarding scoring and run rates.
- **PC1**: It captures, **Wickets** (0.6015) and **RPO** (-0.3295), indicating an inverse relation of Wickets and RPO as it truly is, extras like lb, wide, no-ball are also captured by this component.
- The first five components explain around 90% of the total variance (PC1 = 31.67%, PC2 = 19.25%, PC3 = 13.46%, PC4 = 11.05%, PC5 = 10.25%). In contrast, the sixth component accounts for about 9.01% of the variance. The eighth component contributes negligibly (only 0.07%), indicating that 5-6 components are enough to capture the total variance.

9 Logistic Regression

9.1 Results for 1st innings data

Accuracy = 0.7692
ROC-AUC Score = 0.9273

Class	Precision	Recall	F1-Score	Support
0	0.67	0.91	0.77	11
1	0.91	0.67	0.77	15

Figure 1: Classification Report

	Predicted 0	**Predicted 1**
Actual 0	10	1
Actual 1	5	10

Figure 2: Confusion Matrix

In conclusion, the logistic regression model shows satisfactory performance with an accuracy of 0.77, balanced precision and recall, and a high ROC-AUC score of 0.93, indicating its effectiveness in predicting the binary outcome.

9.2 Results for 2nd innings data

Accuracy = 0.8929
ROC-AUC Score = 0.9896

Class	Precision	Recall	F1-Score	Support
0	0.80	1.00	0.89	12
1	1.00	0.81	0.90	16

Figure 3: Classification Report

	Predicted 0	**Predicted 1**
Actual 0	12	0
Actual 1	3	13

Figure 4: Confusion Matrix

In conclusion, the logistic regression model achieves an accuracy of 0.89, with precision values of 0.80 for class 0 and 1.00 for class 1. The confusion matrix shows all class 0 instances correctly identified, and the ROC-AUC score of 0.99 highlights the model's strong classification ability.

10 Factor Analysis

Variable	**Factor 1**	**Factor 2**	**Factor 3**
Score	-0.9933	-0.0917	0.0055
Balls	-0.5918	-0.7955	-0.0611
RPO	-0.8075	0.5811	0.0396
Result	-0.1420	0.3447	-0.8470
b	-0.1930	-0.0328	0.2278
lb	-0.4280	-0.1025	0.0162
w	-0.2606	-0.1874	0.0881
nb	-0.0816	-0.1589	0.0733
Wickets	0.1491	-0.4899	0.8022

Table 5: Factor Loadings

- **Factor 1** has high negative loadings for *Score* and *RPO*, suggesting that this factor may represent overall batting performance.
- **Factor 2** shows high loadings for *Balls*, which could relate to the duration of play.
- **Factor 3** has a strong loading for *Wickets* and a negative loading for *Result*, possibly indicating match outcomes influenced by dismissals.

11 Linear Discriminant Analysis

- The LDA plot shows clear separation between Class 0 and Class 1, with Class 0 clustering around lower values and Class 1 towards higher values.
- High absolute values of *Score*, *RPO*, and *Wickets* strongly influence the *Result*, with a negative coefficient for wickets indicating that losing more wickets typically leads to match losses, while higher scores and RPO are associated with wins.
- The negative coefficient of *Balls* suggests that teams tend to win matches by chasing quickly, resulting in fewer balls faced.

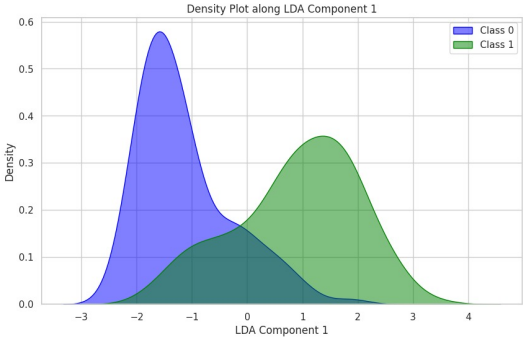

Figure 5: Kernel Density

Parameter	Value	Description
Score	3.40	Indicates score impact
RPO	-1.41	Rate of scoring
Wickets	-2.22	Number of wickets lost

Figure 6: Caption for the table

12 Conclusion

From the tests we have conducted, we can conclude that the batting approach of Team India is different compared to 1st and 2nd innings. The essential factors for the match results are mainly *Score*, *RPO*, *Wickets*, and *Balls*, and we have found them to be essential factors while performing Factor Analysis and Linear Discriminant Analysis. *Extras (b, lb, w, nb)* have less contribution towards the match result.