Data Analytics Approach to Fraud Detection in Indian IT Companies: Beneish Model and Predictive Analysis

Dr. Shraddhaben Girvatsinh Raj
Research Assistant,
P. G. Department of Education,
Sardar Patel University,
Vallabh Vidyanagar

Dr. Vijaykumar C Modesara
Assistant Professor, Deen Dayal Upadhyay Kaushal Kendra,
C P Patel & F H Shah Com. College Anand
Sardar Patel University, Vallabh Vidyanagar

Abstract
Financial statement fraud puts investors, regulators, and other stakeholders at serious risk and damages the reliability of financial reporting. The Indian Information Technology (IT) sector, one of the most rapidly expanding and highly competitive industries in the world, is under increased pressure to maintain steady financial performance, which could encourage earnings manipulation. This paper focuses on fraud detection and prediction in selected Indian IT companies. Fraud detection is assessed using the Beneish M-Score model, and predictive analysis is conducted using logistic regression. Secondary data have been collected from the Capitaline database for 8 large-cap IT companies over a 14-year period from 2011-12 to 2024-25. As an initial screening method, the Beneish M-Score was used to categorise companies according to their risk of manipulation. The likelihood of earnings manipulation was determined, and the relative significance of the Beneish variables was assessed using binary logistic regression with standardized predictors, while inter-firm differences were examined using Welch's ANOVA and Games-Howell post hoc tests. The earnings manipulation is largely episodic, as indicated by the Beneish M-Score. The study found a manipulation-level risk among selected IT Companies. High classification accuracy, i.e., 85.7% found by the result of logistic regression. It has an R^2 value of 69.9%, indicating the model's predictive power. The main determinants of the M-Score were the Days Sales Receivable Index, Asset Quality Index, Gross Margin Index, Leverage Index, and Sales Growth Index, as per the study's results. These findings reinforce theoretical predictions derived from agency and fraud-based frameworks, emphasising the role of financial pressure and reporting incentives.

Keywords: Earnings Manipulation, Beneish M-Score, Logistic Regression, ANOVA, Indian IT Sector

1. Introduction

Financial statements are the most important document for any corporate entity. It gives the reliability and transparency information to the stakeholders of any company. It provides reliable information to investors, regulators, and other stakeholders. The complexity of financial reporting and competitive markets pushes stakeholders to assess financial conditions, risk and return, earnings manipulation, and fraud in statements. In recent years, many financial frauds and risks have been faced by corporations. As a result, it is important to identify such risk manipulation, which is the main task in research in the field of accounting and financial analysis. The Beneish M-Score model is widely regarded as a valuable tool for detecting potential earnings manipulation through various financial ratios. Recently, the incorporation of statistical and predictive analytics has further improved the ability to uncover irregularities in financial reporting. In this context, the current research investigates the probability of earnings manipulation among specific Indian IT firms by utilising the Beneish M-Score model and conducting statistical analysis.

2. Theoretical Framework

Beneish (1999) studies the detection of earnings manipulation in 2,406 US firms by taking 8 variables. The M-Score was calculated in the firms from 1982 to 1992. Probit regression was also applied to examine the level of manipulation among US firms. The author successfully developed a model that detects earnings manipulation, and it is widely used by corporations for fraud detection. Furthermore, Beneish, Lee, and Nichols (2013) examined earnings manipulation and expected returns at the global level, using data on 1,200 firms from 1993 to 2010. Here, the authors applied logistic regression and calculated the M-score for these firms. The result showed an abnormal accrual pattern of manipulation. The study suggested that investors should use the M–Score to avoid fraud. In a similar way, **Roxas (2011)** used logistic regression to examine the model's effectiveness and to calculate fraud. He had considered 218 firms from emerging markets and calculated the M-score by considering 5 variables. It indicated that these indicators are useful in developing economies. **Spathis (2002) and Sharma & Bansal (2018)** both used ratio analysis in fraud detection, and regression analysis indicated that this technique was useful for predicting fraud in their studies. Furthermore, another study by Kaminski, Wetzel, and Guan (2004) used the same methodology and found that fraud firms exhibit abnormal financial ratio patterns. So, the auditors must monitor the ratios. **Dechow et al. (2011)** found strong predictive power for fraud detection by examining the financial ratios of 2190 firms in the USA. The study also identified key predictors of accounting misstatements. In the previous year, **Dechow, Ge, and Schrand (2010)** reviewed the literature on global firms and evaluated their earnings quality using quantitative methods. **Minny Narang (2023)** examined financial indicators of fraud among the top 200 BSE-listed companies. According to the M-Score results, certain companies were found to have a manipulation level. Significant statistical differences were observed in selected profitability, liquidity, leverage, and efficiency ratios between the two groups of firms.The study has used the M-Score model to detect fraud in IT companies in India and has also predicted the level of fraud manipulation using advanced statistical methods, including logistic regression, in recent years, following changes in the pattern of financial reporting.

3. Research Methodology

3.1 Objectives

1. To detect the financial fraud in Indian IT Companies through Beneish M-Score Model
2. To examine the prediction of risk manipulation of the M-Score in Indian IT Companies

3.2 Hypotheses

H_{01}: Beneish M-Score cannot detect fraud in Indian IT companies.
H_{02}: Financial ratios do not significantly predict fraud probability.

3.3 Research Design: Quantitative research design applied to check the financial statement fraud in Indian IT companies by using forensic accounting tools and predictive analytics. The study applies the Beneish M-Score model, Descriptive statistical analysis, and a Logistic regression predictive model to identify earnings manipulation and predict the probability of fraud.

3.4 Sample Selection: The sample consists of large-cap Indian IT companies listed on the National Stock Exchange of India and the Bombay Stock Exchange. Sample companies include - Tata Consultancy Services**,** Infosys**,** HCL Technologies**,** Wipro**,** Tech Mahindra**,** LTIMindtree**,** Mphasis**,** Persistent Systems**,** Coforge, and L&T Technology Services.

3.5 Time Period: Five years from 2020-21 to 2024-25

3.6 Data Collection: The researchers used secondary data collected from the Capitaline database. The financial ratios have been calculated to detect the level of fraud among IT companies. The selected ratios were as follows:

Table 3.1 List of ratios of Beneish M-Score

Name of the Ratio	Abbreviation	Formula
Days Sales Receivable Index	DSRI	$(\text{Receivables}_t / \text{Sales}_t) \div (\text{Receivables}_{t-1} / \text{Sales}_{t-1})$
Gross Margin Index	GMI	$\text{Gross Margin}_{t-1} \div \text{Gross Margin}_t$
Asset Quality Index	AQI	$[1 - (\text{Current Assets}_t + \text{PPE}_t)/\text{Total Assets}_t] \div [1 - (\text{Current Assets}_{t-1} + \text{PPE}_{t-1})/\text{Total Assets}_{t-1}]$
Sales Growth Index	SGI	$\text{Sales}_t \div \text{Sales}_{t-1}$
Depreciation Index	DEPI	$(\text{Dep}_{t-1}/(\text{Dep}_{t-1}+\text{PPE}_{t-1})) \div (\text{Dep}_t/(\text{Dep}_t+\text{PPE}_t))$
SG&A* Index	SGAI	$(\text{SG\&A}_t/\text{Sales}_t) \div (\text{SG\&A}_{t-1}/\text{Sales}_{t-1})$
Leverage Index	LVGI	$(\text{Liabilities}_t/\text{Assets}_t) \div (\text{Liabilities}_{t-1}/\text{Assets}_{t-1})$
Total Accruals to Total Assets	TATA	$(\text{Net Income} - \text{CFO}) \div \text{Total Assets}$
***SG&A = Selling, General & Administrative expenses**		

3.7 Data Analysis: The researchers have applied the Beneish M-Score algorithm to detect levels of fraud in Indian IT Companies. Second, fraud prediction is performed using Logistic regression analysis. Descriptive and inferential statistics were applied using SPSS and Microsoft Excel. The algorithm of M-Score and variables for the predictive models is as follows:

3.8 Beneish M-Score Analysis

The Beneish M-Score model is used to detect earnings manipulation.

$$\textbf{M-Score} = -4.84 + 0.920(\text{DSRI}) + 0.528(\text{GMI}) + 0.404(\text{AQI}) + 0.892(\text{SGI}) + 0.115(\text{DEPI}) - 0.172(\text{SGAI}) + 4.679(\text{TATA}) - 0.327(\text{LVGI})$$

Interpretation:

- **M-Score > −2.22 → Likely Manipulator**
- **M-Score < −2.22 → Non-Manipulator**

Table 3.2 List of Independent and Dependent Variables

	Name of Variable	Description
Dependent Variable	Fraud indicator (Binary variable): 1 = Likely Manipulator 0 = non-Manipulator	If M-Score > −2.22 = Manipulator (1), and M-Score < −2.22 = Non-Manipulator (0)
Independent Variables	DSRI	Days Sales Receivable Index
	GMI	Gross Margin Index
	AQI	Asset Quality Index
	SGI	Sales Growth Index
	DEPI	Depreciation Index
	SGAI	Selling, General & Administration expenses Index
	LVGI	Leverage Index
	TATA	Total Accruals to Total Assets

4. Result of the Study

The study calculated the Beneish M-Score for each IT company in each year to assess the level of manipulation in Indian IT Companies. Furthermore, the prediction analysis was examined using logistic regression. The Beneish M – Score, One Way ANOVA test with fulfilment of all assumptions and logistic regression analysis result presented as under:

A. Bneish M-Score

Table 4. 1 Beneish M-Score Result of Selected IT Companies

Year End	Coforge	HCL Technologies	Infosys	Mphasis	Persistent Systems	TCS	Tech Mahindra	Wipro
2012	-2.873	-1.866	-0.984	-2.328	-1.992	-1.551	-2.122	-2.039
2013	-1.742	-2.313	-0.928	-2.451	-2.307	-1.969	-2.047	-2.484
2014	-1.682	-2.293	-1.267	-0.515	-2.449	-1.607	-0.434	-2.150
2015	-2.335	-1.727	-0.782	-2.045	-2.383	-2.224	-2.039	-2.466
2016	-2.452	-1.979	-0.983	-2.273	-2.060	-1.294	-2.461	-2.027
2017	-3.180	-1.992	-0.822	-2.065	-2.214	-2.115	-2.557	-2.447
2018	-2.108	-2.000	-1.283	-2.204	-2.512	-2.054	-2.297	-2.387
2019	-2.220	-3.742	-1.262	-2.214	-2.434	-1.846	-2.138	-2.484
2020	-2.190	-2.262	-1.346	-2.505	-14.772	-42.772	-2.631	-2.380
2021	-2.871	-2.419	-1.898	-2.212	-2.852	-2.270	-2.696	-2.691
2022	-2.857	-2.327	-1.684	-1.333	-2.805	-1.841	-1.612	-2.200
2023	-2.343	-2.172	-1.891	-1.888	-1.852	-1.837	-2.413	-2.300
2024	-2.282	-2.586	-1.689	-4.189	-2.223	-1.934	-2.925	-2.641
2025	-1.892	-2.507	-2.145	-2.160	-1.920	-1.990	-2.394	-2.456

Source: Author Calculation

Note: Red Colour indicates Likely Manipulated, and Green Colour indicates non-manipulated M-score. According to the Beneish M-Score results for the chosen IT companies (2012–2025), most companies consistently fall below the (–2.22) threshold. It indicated a lower level of earnings manipulation. Many companies have found periodic manipulation risk, as indicated by the M-Score results. Coforge, Infosys, and Mphasis display more regular threshold crossings, whereas TCS, Wipro, and HCL Technologies show relatively consistent reporting patterns. The major negative results in the year 2020 are probably due to unusual accounting shifts rather than manipulation. In general, the evidence points to periodic rather than regular earnings management in the Indian IT industry.

B. One-Way Anova Result

4.2. Descriptive statistics of M-score

	N	Mean	Std. Deviation
Coforge	14	2.732	0.684
HCL Technologies	14	1.570	0.648
Infosys	14	2.038	0.277
Mphasis	14	1.460	0.863
Persistent Systems	14	1.448	3.322
TCS	14	-0.213	10.941
Tech Mahindra	14	2.110	0.791
Wipro	14	0.638	0.335
Total	112	1.473	4.042

Source: Output of SPSS

4.3 Tests of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
Based on Mean	3.544	7	104	0.002
Based on Median	1.001	7	104	0.435
Based on Median and with adjusted df	1.001	7	15.361	0.467
Based on trimmed mean	1.360	7	104	0.230

Source: Output of SPSS

4.4 Standard ANOVA Result of M-Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	82.041	7	11.720	0.704	0.669
Within Groups	1731.305	104	16.647		
Total	1813.346	111			

Source: Output of SPSS

4.5 Welch ANOVA Result of M-Score

	Statistic[a]	df1	df2	Sig.
Welch	25.581	7	43.310	0.001
a. Asymptotically F distributed.				

Source: Output of SPSS

4.6 Games–Howell Post Hoc Results of M-Score

Group 1	Group 2	Mean Difference(I–J)	p-value	Interpretation
Coforge	HCL Technologies	1.162	.002	Coforge > HCL
Coforge	Infosys	0.694	.042	Coforge > Infosys
Coforge	Mphasis	1.272	.005	Coforge > Mphasis
Coforge	Wipro	2.094	<.001	Coforge > Wipro
HCL Technologies	Wipro	0.932	.003	HCL > Wipro
Infosys	Wipro	1.400	<.001	Infosys > Wipro
Tech Mahindra	Wipro	1.473	<.001	Tech Mahindra > Wipro

Source: Output of SPSS

The tables above show the results of the one-way ANOVA test of the M-score among the selected IT Companies. The researchers have applied the test and met the assumptions. According to the Levene test, the variances are not equal, with a 3.544 statistic and a 0.002 p-value. Hence, the researchers used Welch ANOVA, and the test result was highly significant (p=0.001). It indicated a significant difference in the mean M-Score between selected IT companies. Furthermore, the researchers used the Games-Howell post hoc test to determine significant differences between groups. A total of seven pairwise significant differences were found in the test results. The mean M Score values of Coforge were significantly higher than those of HCL Technologies, Infosys, Mphasis, and Wipro. Furthermore, Wipro's M_Score values were significantly lower than those of Tech Mahindra, Infosys, and HCL Technologies. At the 5% level, all of the other pairwise comparisons were statistically significant.

Predictive Analysis

The logistic regression model has been applied to predict the probability of fraud in Indian IT Companies. The researchers used the Beneish M-Score as the dependent variable (1 = Likely Manipulator, 0 = Non-Manipulator), and the Days Sales Receivable Index, Gross Margin Index, Asset Quality Index, Sales Growth Index, Depreciation Index, SGA Index, Leverage Index, and Total Accruals to Total Assets as independent variables. The researchers had transformed independent variables into standardized form. Z-score transformation was used to standardize independent variables to ensure scale consistency and reduce estimation instability (Tabachnick & Fidell, 2019). Standardization increases the robustness of the model and enables the analysis of coefficients in terms of standard deviation changes (Field, 2018). The study has satisfied the assumptions of logistic regression. Standardized residuals were used to measure outliers; values greater than ±3 were regarded as indicative of extreme outliers (Field, 2018). Variance Inflation Factor (VIF) values below 5 suggest acceptable levels of multicollinearity, while values above 10 indicate major multicollinearity concerns (Hair et al., 2019; Menard, 2002). The model satisfies these predetermined statistical thresholds, according to the test results.

4.7 Model Fit Statistics of Logistic Regression

Statistic	Value
-2 Log Likelihood	71.790
Omnibus Chi-square	82.903
df	8
Sig.	< .001
Cox & Snell R^2	0.523
Nagelkerke R^2	0.699
Hosmer–Lemeshow Chi-square	4.214
Hosmer–Lemeshow Sig.	0.837
Overall Classification Accuracy	85.7%
Sample Size (N)	112

Source: Output of SPSS

4.8 Classification Accuracy

Observed Category	Correctly Classified	Percentage Correct
Non-Manipulated (0)	43 / 52	82.7%
Manipulated (1)	53 / 60	88.3%
Overall Accuracy	—	**85.7%**

Source: Output of SPSS

4.9 Logistic Regression Coefficients (Final Model)

Variable	B	Sig.	Exp(B)	Interpretation
Days Sales Receivable Index	3.534	<.001	34.270	Significant positive effect
Gross Margin Index	-3.265	<.001	0.038	Significant negative effect
Asset Quality Index	0.871	.041	2.390	Moderate positive effect
Sales Growth Index	50.397	.024	Very Large	Significant positive effect
Depreciation Index	4.393	.150	80.913	Not significant
SGA Index	-51.142	.024	≈0	Significant negative effect

Leverage Index	-39.941	<.001	≈0	Significant negative effect
Total Accruals to Total Assets	0.272	.363	1.313	Not significant
Constant	-3.108	.029	0.045	—

Source: Output of SPSS

4.1 Formula of Logit (Linear Predictor) Form

p denote the probability that a firm is classified as **Likely Manipulated (1)**.

$$\ln\left(\frac{p}{1-p}\right) = -3.108 + 3.534(\mathrm{DSRI}) - 3.265(\mathrm{GMI}) + 0.871(\mathrm{AQI}) + 50.397(\mathrm{SGI}) + 4.393(\mathrm{DEPI}) - 51.142(\mathrm{SGAI}) - 39.941(\mathrm{LVGI}) + 0.272(\mathrm{TATA})$$

4.2 Formula of Logistic Probability Form

The probability of earnings manipulation is:

$$p = \frac{1}{1+e^{-Z}}$$

$$Z = -3.108 + 3.534(\mathrm{DSRI}) - 3.265(\mathrm{GMI}) + 0.871(\mathrm{AQI}) + 50.397(\mathrm{SGI}) + 4.393(\mathrm{DEPI}) - 51.142(\mathrm{SGAI}) - 39.941(\mathrm{LVGI}) + 0.272(\mathrm{TATA})$$

Note: DSRI = Days Sales Receivable Index, GMI = Gross Margin Index, AQI = Asset Quality Index, SGI = Sales Growth Index, DEPI = Depreciation Index, SGAI = Sales, General & Administrative Index, LVGI = Leverage Index, and TATA = Total Accruals to Total Assets

The tables above show the results of the binary logistic regression. The impact of the M-Score model on the probability of earnings manipulation in selected Indian IT Companies was examined through logistic regression. A statistically significant overall model ($\chi^2(8) = 82.903$, $p < .001$) indicated that the set of independent variables reliably distinguished between manipulated and non-manipulated IT companies. The model showed significant explanatory power, with a Nagelkerke R2 of 0.699, representing that the included factors explain approximately 69.9% of the variation in earnings-manipulation status. The Hosmer–Lemeshow test was non-significant ($p = 0.837$), indicating good model fit. With 88.3% of manipulated cases and 82.7% of non-manipulated cases correctly classified, the model's predictive power is further supported by its 85.7% classification accuracy. Statistically significant variables that influence earnings manipulation were a number of predictors: Days Sales Receivable Index ($p <.001$), Gross Margin Index ($p <.001$), Asset Quality Index ($p =.041$), Sales Growth Index ($p =.024$), SGA Index ($p =.024$), and Leverage Index ($p <.001$). In particular, stronger gross margin stability decreased the probability of manipulation, but increases in receivables and sales growth significantly increased the likelihood of manipulation. Whereas the Depreciation Index and the Total Accruals to Total Assets were statistically significant predictors. Overall, the results show that the Indian IT sector's earnings manipulation risk is significantly predicted by the important Beneish Model. The Beneish framework is robustly validated in this context using logistic regression, as evidenced by excellent predictive accuracy and strong model fit.

5. Findings, Implications, and Conclusion of the Study

The results of the Beneish M-Score analysis indicated that the likelihood of earnings manipulation varied among the selected IT companies over the study period. Some firms reported M-Score values that approached or exceeded the benchmark threshold of −2.22 in certain years, indicating potential signals of manipulation, whereas others consistently fell below the threshold, suggesting a lower risk of such activities. The standard ANOVA indicated no statistically significant differences in mean M-Score values among the firms ($p = 0.669$). However, the Welch ANOVA test revealed significant variability among companies ($p = 0.001$), indicating differences in manipulation risk within the IT industry. The Games–Howell post hoc test subsequently identified notable pairwise differences between multiple companies. The findings from the logistic regression indicated that DSRI, AQI, and SGI significantly increased the likelihood of earnings manipulation, whereas GMI, SGAI, and LVGI showed significant inverse associations. Additionally, DEPI and TATA were found to have no significant impact on the prediction of manipulation in the companies analysed.

The results of the research offer significant insights for investors, auditors, regulators, and policymakers. The findings initially underscore the effectiveness of the Beneish M-Score model for detecting potential financial statement manipulation in the IT industry. Investors can apply this model as an initial filter before making investment decisions. Secondly, the significant impact of indicators such as DSRI, AQI, and SGI indicates that stakeholders should pay close attention to receivables management, shifts in asset quality, and trends in rapid sales growth when assessing a company's financial statements. These metrics could act as early warning signs of possible earnings manipulation. To enhance fraud detection, auditors and regulatory bodies ought to integrate data analytics and forensic accounting methods with conventional auditing practices. Implementing predictive analytics models, such as logistic regression and financial ratio analysis, could enhance the effectiveness of financial statement evaluation and corporate governance efforts. Additionally, organisations should bolster their internal control mechanisms and enhance transparency in financial disclosures to minimise the chances of manipulation and foster investor trust.

In Conclusion, it indicates that some companies displayed signs of potential manipulation during particular years, whereas others exhibited comparatively lower risk levels. Additionally, the analysis indicated that certain financial indicators played a significant role in uncovering possible manipulations. The results imply that the Beneish M-Score model, supported by statistical analysis, can be an effective analytical tool for detecting potential financial statement manipulation in the IT industry.

References

1. Beneish, M. D. (1999). The detection of earnings manipulation. Financial Analysts Journal, 55(5), 24–36. https://doi.org/10.2469/faj.v55.n5.2296
2. Beneish, M. D., Lee, C. M., & Nichols, D. C. (2013). Earnings manipulation and expected returns. Financial Analysts Journal, 69(2), 57–82.
3. Capitaline. (2025). Capitaline corporate database. Capital Market Publishers (India) Pvt. Ltd. https://www.capitaline.com
4. Dechow, P. M., Ge, W., & Schrand, C. M. (2010). Understanding earnings quality: A review of the proxies, their determinants, and their consequences. Journal of Accounting and Economics, 50(2–3), 344–401.
5. Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. Contemporary Accounting Research, 28(1), 17–82.
6. Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). Sage Publications.
7. Gupta, M., & Gill, N. (2012). Financial ratio analysis for fraud detection. International Journal of Financial Research.
8. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.
9. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.
10. Kaminski, K., Wetzel, T. S., & Guan, L. (2004). Can financial ratios detect fraudulent financial reporting? Managerial Auditing Journal, 19(1), 15–28.
11. Menard, S. (2002). Applied logistic regression analysis (2nd ed.). Sage Publications.
12. Narang, M. (2023). Fraud detection in companies listed on the Bombay Stock Exchange using the Beneish M-Score model and logistic regression analysis. Journal of Financial Crime, 30(4), 1023–1041. https://doi.org/10.1108/JFC-03-2022-0054
13. Roxas, M. L. (2011). Financial statement fraud detection using ratio and digital analysis. Journal of Leadership, Accountability and Ethics, 8(4), 56–66.
14. Sharma, A., & Bansal, R. (2018). Financial ratio analysis for fraud detection in Indian firms. Journal of Accounting and Finance.
15. Spathis, C. (2002). Detecting false financial statements using published data. Managerial Auditing Journal, 17(4), 179–191.
16. Tabachnick, B. G., & Fidell, L. S. (2019). Using multivariate statistics (7th ed.). Pearson.