Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2026 (44) 2** : 1–11.
ISSN-0037-8712
doi:10.5269/bspm.80527

Decoding Start-up Success: Predicting Future Through Founder Profiles - A Statistical Analysis

V. Ganesh Kumar, K. Jhansi Lakshmi Bai, B. Pragna, Ch. Sharanya, H. Karthika

ABSTRACT: Startups stimulate economic growth, innovation, and job creation, but they need on outside funding to be viable. Based on the literature, founder traits such as social capital, human capital, and entrepreneurial experience influence investment performance nevertheless, empirical research is inconsistent and context dependent. Using Data Labs, LinkedIn, and Crunchbase data on 300 Indian start-ups, this analysis quantifies the association between founder traits and success rates. We use multi-linear and logistic regression models to evaluate how past entrepreneurial experience, education, industry exposure, and connectivity affect fundraising success. We also show how expertise and education boost fundraising possibilities, especially in urban areas and fintech, where investors favor innovation scalability above revenue creation. The findings provide practical advice for entrepreneurs, accelerators, and lawmakers who want to make start-up investments easier, and they also illustrate that variables at the founder level are predictive of success rates

Keywords: Logistic regression, multiple linear regression, Random Forest, funding success prediction, data-driven entrepreneurship.

Contents

1. Introduction

Human capital is a weak predictor of admission into fledgling entrepreneurship, according to a study by [1], but it is a strong predictor of successfully completing the start-up process. Colombo and Grilli [2] collaborated to analyze the success of 439 new technology-based businesses in Italy was impacted by the founders' human capital and their ability to obtain venture capital funding. The study was limited to Italy, but it did indicate that founders' human capital drives venture growth through access to venture financing. In a study involving 42 independent samples, the authors in [3] identified a significant relationship between EET and assets related to human capital in entrepreneurship, as well as the outcomes of entrepreneurship. Jyoti and Singh [4] investigated the connection between socioeconomic characteristics of start-ups and their size in Gujarat, India. It also assessed the determinants affecting the annual sale of start-ups.

A Critical Success Factor stage model for FinTechs was presented by Barz, Lindeque, and Hedman [5] which incorporates insights from 18 interviews in the Danish FinTech industry as well as existing research on success determinants, while Subrahmanya [6], [7] conducted an exploration and examination of the structure, evolution, and growth of ecosystems pertaining to technology startups within the context of Bangalore and Hyderabad. According to Metrick and Yasuda [8], new businesses play a crucial role in today's economies by creating jobs, technological innovations, and spillovers. Many basic and advanced statistical techniques such as regression and applied regression models are discussed by the authors of [10], [11], [13]. The Founder's Dilemmas [14] is the first book to examine the early decisions by entrepreneurs that can make or break a startup and its team. It reveals the common pitfalls founders face and how to avoid them.

2020 *Mathematics Subject Classification:* 62J02, 62J05, 62-07, 00A66.
Submitted December 01, 2025. Published March 14, 2026

Present study fills these gaps with a quantitative data-driven approach. Using a dataset of 300 Indian start-ups drawn from Datalabs [16], Crunchbase [17] and LinkedIn [18], it employs Random Forest (RF), logistic regression (LR), and multiple linear regression (MLR) algorithms to predict funding success. In contrast to earlier studies, proposed work specifically analyzes connections between founder qualities and ecosystem elements using digital traces as measurable measures of human and social capital.

2. Methodology

This article proposes the statistical analysis of Indian startups across sectors and analyzes the characteristics to identify distinct patterns or trends in the influencing factors. Machine learning models are proposed to analyze the data which involves the factors funding, sector, headquarters, educational background, funding rounds, and launch year. The methodology encompasses data processing, data visualization using descriptive statistical analysis. ML models like LR, RF, and MLR models are applied. Entire code is implemented using R-programming (version 4.3.2) as a software tool.

2.1 Variable Consideration

Target variable - Net profit margin (NPM) Explainable variables - Educational Background, Sector, Head Quarters, Launch Year, Stage, Total Funding (in CR), Total Revenue (in CR), Total expenses (in CR), Total assets, Total rounds.

2.2 Data Visualization Approach

This study presented a visualization method called multi-layered for a comprehensive analysis of startup patterns to compare performance indicators. Frequency distributions, pie chart, and histograms are utilized to visualize the data which identify capture evolutionary trends, sectoral concentrations, and uncover intricate linkages. The process begins with univariate distributions and advances to multivariate relationships; each visualization style being selected to correspond to the data properties and analytical objectives.

2.3 Statistical Modelling Methodology

To determine the degree of dependency among the variables, a simple linear correlation analysis is employed. MLR and LR models are framed to look at how several factors affected startup profitability at the same time. Variance inflation factor (VIF) is analyzed to identify possible multicollinearity.

The initial analytical phase utilized logistic regression to model the binary outcome of funding acquisition. This model specification incorporated founder level predictors including educational background, prior entrepreneur experience, and professional network indicators, along with startup level variables such as sector classification and geographical location. We employed maximum likelihood estimation to determine the probability of funding success given the predictor variable. Model diagnostics included examination of Hosmer-lemeshow goodness-of-fit statistics and receiver operating characteristics (ROC). Curve analysis to assess predictive accuracy.

Present study provided a non-parametric alternative that accommodates complex interaction effects and non-linear relationships. The algorithm's inherent feature of prior measures helped identify the most influential predictors of startup success, while it is a bagging approach mitigative over fitting. We took other parameters through error minimization.

3. Figures and Results

Data visualization helps summarize complex datasets into meaningful visual insights. Extensive data visualization is performed using R programming and the ggplot2 library to uncover hidden patterns, identify trends and summarize the structural characteristics of start-up related variables before applying predictive models. Using R and ggplot2, we constructed a series of plots-bar chart, line chart, box plot, scatter plot and pie chart- to analyze various factors of a dataset.

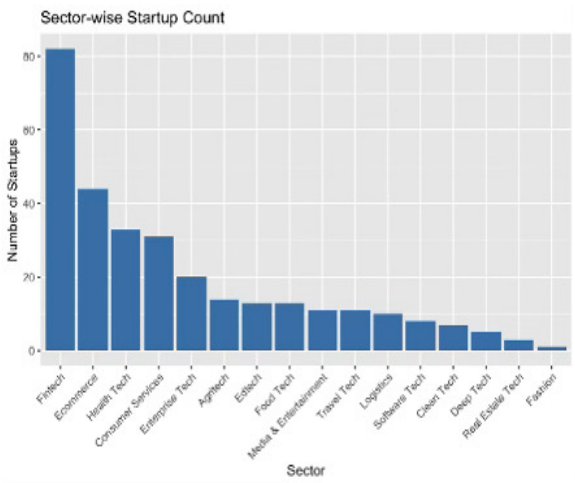

Figure 1: Startup Count

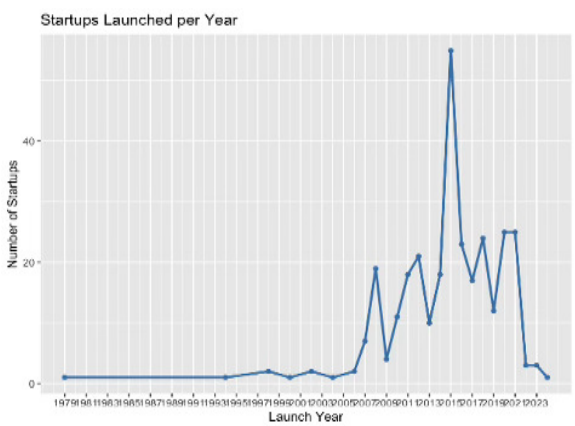

Figure 2: Startups Launched per Year

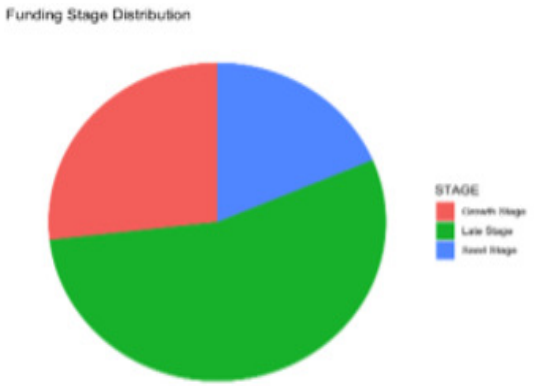

Figure 3: Funding Stage distribution

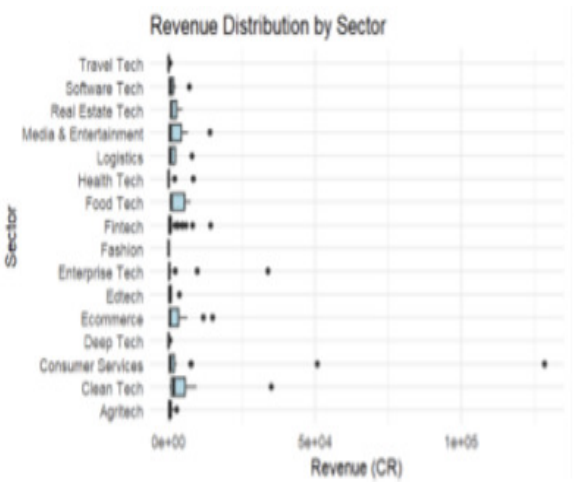

Figure 4: Revenue distribution by sector

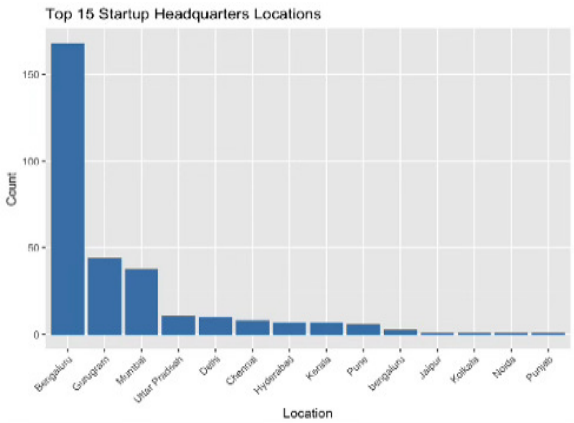

Figure 5: Top 15 startup Headquarters Locations

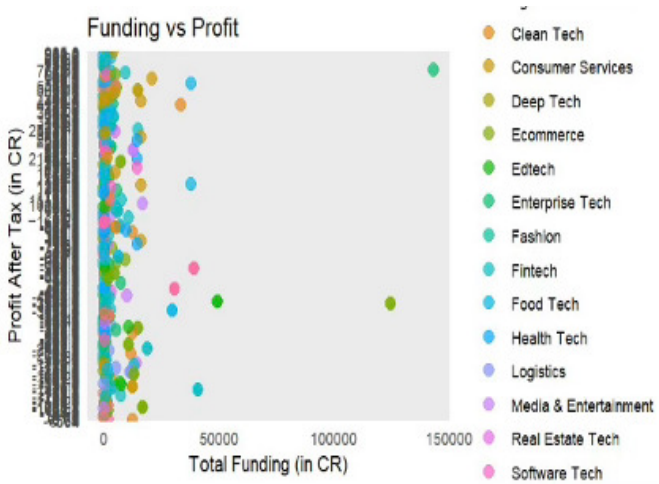

Figure 6: Funding vs Profit

Fig.1 represents sector-wise startup distribution, X-axis representing different sectors and Y-axis representing startup count. This visualization revealed that Fintech, E-commerce and Health tech are the most active sectors, indicating high entrepreneurial activity, investor interest and funding is in these consumer-centric sectors. This insight provided a foundational understanding of where entrepreneurial activity is most prominent. Fig. 2 shows a line chart that was used to analyze patterns in the formation of startups over time. The Y-axis shows the total number of startups, while the X-axis shows the year of inception. During 2005 to 2016, there was a notable growth phase, and a modest dip were observed.

The distribution of funding stages was presented in Fig. 3. The green colour indicates late-stage startups (55%), while red and blue signify the growth-stage startups (26%) and seed-stage startups (19%) respectively. The distribution of revenue across industries is illustrated in Fig. 4. There is a logarithmic-like scientific scale for revenue on the X-axis and a Y-axis for various industries. The plot effectively summarized the median, and outliners of revenue values for a sector. Sectors like Enterprise Tech and Fintech showed higher median revenues and wider spreads indicating both potential and variability, while sectors such as Clean Tech and Deep Tech exhibited significant outliners, pointing to a few exceptionally high performing startups.

Fig.5 graph analysis geographical distribution of startups in India, with X-axis representing locations

and Y-axis representing the count of startups in respective locations. This bar analysis showed that Bengaluru, followed by Gurugram and Mumbai, accounted for the highest number of startups. This pattern reflects the availability of tech infrastructure, tech talent and investor access in Tier-1 cities. Fig.6 is generated to evaluate total funding and profit. The plot generally shows the correlation between two continuous variables. In the above graph profit after tax and funding are considered as they are two continuous variables with Funding on X-axis and Profit on Y-axis. Data points are color coded by sector to indicate clustering patterns. There was no definitive linear correlation between funding and profitability. The importance of efficient capital utilization is highlighted by the observation that, although finance can support expansion, it does not ensure profitability.

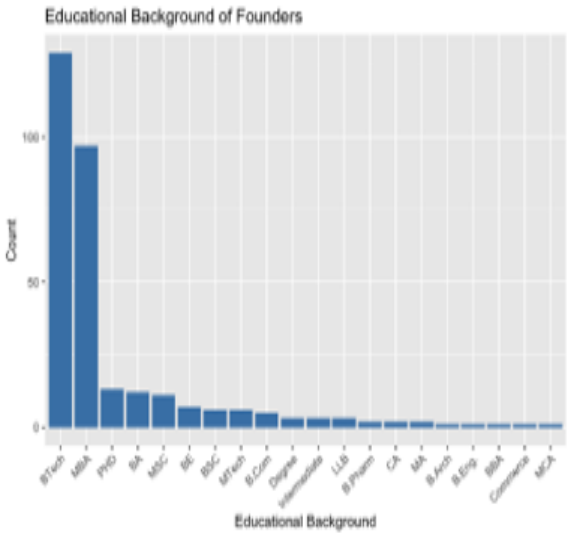

Figure 7: Educational Background of Founders

Fig. 7 is representing the educational background of founders. It reveals a predominance of engineering degrees, emphasizing the importance of technical expertise in building scalable ventures. The preceding S-curve plotting presents a detailed yet straightforward understanding of a logistic regression model's ability in correctly predicting the success of a startup. On the chart, every cross stands for a startup on the X-axis and its linear combination for predictors, with the Y-axis showing the model's probability estimate of the success of a particular startup (success is considered when NPM=1). From the above graph, we can conclude that if the input values are very low, the cross marks are closer to zero, i.e., we see the red crosses nearer to zero; whereas, if the input is high with crosses marked in blue, they tend to be nearer to one. In the curve, overlapping between red and blue.

4. Model Discussion

4.1 Regression Model Analysis

4.1.1 Simple Linear Regression

NPM was significantly and positively correlated with Total Revenue (r = 0.84) and Stage Numbering (r = 0.72). Funding and revenue logarithm values (r = 0.22 and 0.62), Launch Year (r = 0.32), and Total Funding (r = 0.27), also showed moderately strong positive correlations, indicating that companies with higher revenues, better funding, and a newer launch have higher profit margins. Conversely, factors such as Sector Numbering, Headquarters Numbering, Education Numbering, Total Expenses, and Total Assets exhibited low correlation with NPM, indicating minimal direct impact on profitability. The total number of funding rounds exhibited a slightly adverse trend.

4.1.2 Multiple Linear Regression

The multiple linear regression model developed to forecast Net Profit Margin (NPM) incorporates various independent variables, including industry classification, education background, year launched, headquarters location, stage, number of education programs, total expenses, log-transformed revenue funding, total assets, number of funding rounds, demonstrating satisfactory overall performance. With a Multiple R-squared value of 0.9922, the model predicts nearly 99.22% of NPM variance, which is extremely unprecedented. This means that the selected set of independent variables are predicting NPM at an extremely high level of precision. But when model complexity is determined by number of predictors, Adjusted R-squared decreases to 0.8327, showing that almost 83.27% of the variance in NPM is being explained after having accounted for model complexity. The decrease shows that there is a chance some of the variables are not contributing much and are rather causing the model to overfit and providing it with an overestimated performance. An Adjusted R-squared value of more than 0.8 also shows a good model. The F-statistic value of 6.221 and possessing a very insignificant p-value of 0.0001887 proves that the overall regression model itself is statistically significant. That is, there is strong evidence to prove that there is at least one independent variable which has a significant association with the dependent variable, NPM. It is also of note that four observations were removed due to missing data, which reduced the sample size and degrees of freedom slightly (14). The model appears statistically adequate and sound, though perhaps it could be improved by dropping weaker predictors to more strictly uphold parsimony.

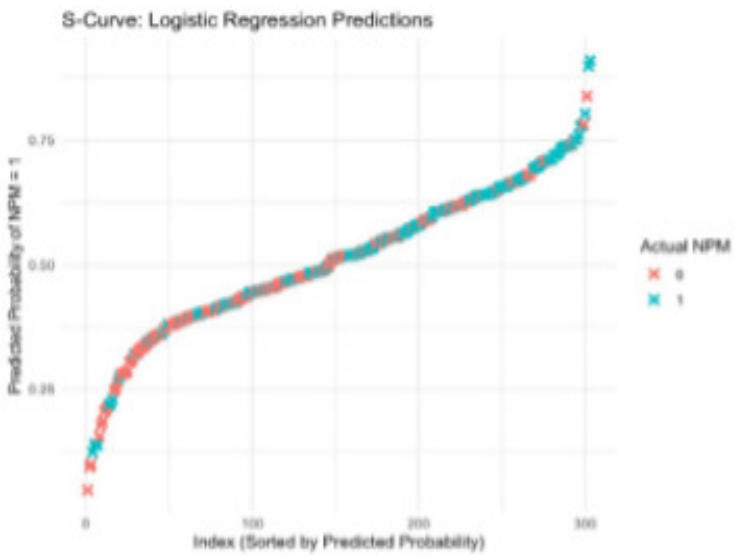

Figure 8: Prediction by Logistic Regression

Figure 9: Correlation

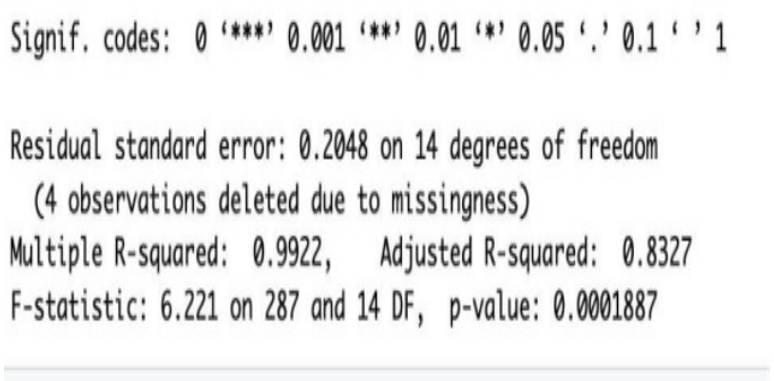

Figure 10: Summary of MLR

```
Call:
 randomForest(formula = NPM ~ ., data = data, ntree = 500, mtry = 3,      importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 2.33%
Confusion matrix:
    0   1 class.error
0 146   1 0.006802721
1   6 148 0.038961039
```

Figure 11: Summary of RF

4.2. Random Forest Regression Model

The goal was to predict the Net Profit Margin (NPM) with a Random Forest classification model using numerous measures of finance. A Random Forest (RF) classification model was created to predict net profit margin with several measures of finance. The model was trained on a cleaned dataset without any missing values, and the target variable was reshaped in a categorical format. The RF model was grown with three variables sampled randomly at each node and with 500 trees. The model performed excellently with an out-of-bag (OOB) error rate of only 2.33 percentage. The confusion matrix illustrates a high classification accuracy with class 0 correctly classified 146 of 147 times and class 1 classified correctly 148 of 154 times. The overall error rates for each class were 0.68 and 3.90 for classes 0 and 1 respectively, the model generalized well to the unseen dataset and could discriminate well. Also, the variable importance plot shows the strongest predictors that have contributed to making the decision of the model which could illustrate the strength and interpretability of the Random Forest approach.

5. Conclusion

Sharp uptick in startup formations between 2005 and 2016 are aligned with rising internet penetration, smartphone adoption, and pro-business government policies. It also proving that entrepreneurial growth was fostered by favorable policy environments and technological advancements. Over half of the startups funded are in late stages, indicating the maturity of the Indian startup ecosystem and investor conviction in scalable enterprises. It can be observed that Bengaluru, Gurugram, and Mumbai are the leading startup hubs. Findings from the education background indicate that founders with engineering degrees are the predominant figures in the ecosystem, highlighting the significance of technical expertise in creating scalable enterprises. The RF Model had an Out-Of-Bag error rate of 2.33 percentage, indicating strong classification accuracy and generalization. Both SLR and MLR models indicated that funding, revenue, and launch year serve as strong positive predictors of profitability. The MLR model established a strong capability in explaining variance (Adjusted $R^2 = 0.8327$), as financial and time variables collectively serve as robust predictors of profit margins.

The present study provides empirical evidence that the success of startups can be studied through recognizable and data-supported paths, with a particular emphasis on the qualities of the founders and the characteristics of the setting in which they operate. Large-scale data visualization and statistical modeling reveal startup financial and structural tendencies. Financial indicators are the best predictors of company success, showing that ensemble learning algorithms can understand complex financial processes. According to research by sector, the most promising new businesses in the tech-driven service and financial sectors are those in the health technology, e-commerce, and financial technology sectors. It is well-established that difficulties encountered by founders are a good predictor of the ultimate funding outcomes and offer valuable insights that investors, entrepreneurs, and lawmakers can use to their advantage. This methodology quantifies ecosystem exposure, social capital, and human capital to improve startup investment and advance data-driven entrepreneurship.

Our future project scope is as follows:

- Adding start-ups from other nations may help determine if the findings apply to other economic and cultural situations. A comprehensive sector-specific SNA beyond FinTech may reveal variations in investor behavior across different sectors.
- In longitudinal research, the impact of second or simultaneous processes on subsequent stages of a firm is analyzed, along with the evolution of founder qualities and their impact on long-term performance, compared to financial concerns.
- Analyzing dynamic models Using founder data, market trends, economic indicators, and investor attitude to assess fundability in a changing environment may require more research.

Acknowledgments

We thank all the contributors.

References

1. P. Davidsson, B. Honig, *The role of social and human capital among nascent entrepreneurs*, Journal of Business Venturing. 18, 301-331, (2003).
2. Colombo, M.G., and Grilli, L., *On growth drivers of high-tech start-ups: Exploring the role of founders' human capital and venture capital*, Journal of Business Venturing. 25, 610 -626 (2010).
3. Martin,B. C., McNally, J. J., and Kay,M. J., *Examining the formation of human capital in entrepreneurship: A meta-analysis of entrepreneurship education outcomes*, Journal of Business Venturing. 28, 211-224 (2013).
4. Jyoti, B., Singh, A. K., *Characteristics and Determinants of New Startups in Gujarat*, India, 2020.
5. Barz, L., Lindeque, S., Hedman, J., Critical success factors in the FinTech World: A stage model, Electronic Commerce Research and Applications. 60, 101280 (2023).
6. Subrahmanya, M. H. B., *Comparing the Entrepreneurial Ecosystems for Technology Startups in Bangalore and Hyderabad, India*, Technology Innovation Management Review. 2017.
7. Subrahmanya, M. H. B., *Entrepreneurial Ecosystems for Tech Start-ups in India: Evolution, Structure and Role*, Boca Raton, CRC Press. 2021.
8. Metrick, A., Yasuda, A., *Venture capital and the finance of innovation*, 2021.
9. Aulet, B., *Disciplined Entrepreneurship: 24 Steps to a Successful Startup, Expanded and Updated*, Hoboken, NJ: John Wiley and Sons, 2024.
10. James, G., Witten, D., Hastie, T., Tibshirani, R., *An Introduction to Statistical Learning*, New York, NY: Springer, 2013.
11. Weisberg, S., *Applied Linear Regression*, 3rd ed. Hoboken, NJ: John Wiley and Sons, 2005.
12. St. John, R. C., *Applied Linear Regression Models*, Journal of Quality Technology. 15, 201-202 (2018).
13. Field, A., Field, Z., Miles, J., *Discovering Statistics Using R*, London: SAGE Publications, 2012.
14. Wasserman, N., *The Founder's Dilemmas: Anticipating and Avoiding the Pitfalls That Can Sink a Startup*, Princeton, NJ: Princeton University Press, 2012.
15. Autio, E., Sapienza, H. J., Almeida, J. G., *Effects of Age at Entry, Knowledge Intensity, and Imitability on International Growth*, Academy of Management Journal. 41, 921-945 (2017).
16. *Datalabs website* - https://inc42.com/datalabs/
17. *Crunchbase website* - https://www.crunchbase.com/organization/crunchbase
18. *LinkedIn website* - https://gb.linkedin.com/company/linkedin

V. Ganesh Kumar, Assistant Professor,
Department of Mathematics,
Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology,
India.
E-mail address: ganeshkumar68@gmail.com

and

K. Jhansi Lakhmi Bai, Assistant Professor,
Department of CSE,
Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, India.
E-mail address: jhansi_cse@vnrvjiet.in

and

B. Pragna, Student,
Department of CSE-CSBS,
Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, India.

and

Ch. Sharanya, Student,
Department of CSE-CSBS,
Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, India.

and

H. Karthika, Student,
Department of CSE-CSBS,
Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, India.