iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.60121

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: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

Invest in Idea: Crowd Funding with Predicting Startup Success

S Pranav[1], Venkat Lakshmi[2], Bandaru Nandini[3], K.Deepa[4], Mrs. Shwetha Shree[5]
Department of CSE, Ballari Institute of Technology & Management, Ballari

Abstract: ***Crowd funding enables individuals or businesses to raise funds for their projects, ventures, or caused by tapping into a large pool of contributors. It democratizes access to capital, allowing creators to bypass traditional financial institutions and engage directly with their community or target audience. This method fosters innovation, empowers grassroots movements, and facilitates the realization of diverse ideas that might otherwise struggle to secure funding through conventional channels. Utilizing a rich dataset encompassing funding details, milestones, relationships, and geographical data of various startups, we embark on an in- depth analysis involving data preprocessing, feature engineering, and exploratory data analysis to uncover key success determinants. Employing advanced classifiers like LGBM, XG Boost and gradient Boosting, our model undergoes rigorous training and evaluation, with LGBM emerging as the top performer, achieving an accuracy of 90.48%. The analysis underscores the critical role of funding, and investor presence in forecasting startup success. This research not only equips stakeholders with a powerful predictive tool but also highlights significant features influencing startup outcomes, offering a novel perspective on strategic investment planning. Additionally, we developed an interactive frontend platform to complement our predictive model. Utilizing AngularJS, the platform serves as a gateway for startups to register their ventures, providing real-time access to predictive insights. Through intuitive forms, startups enter essential information such as geographical location, funding received, and sector of operation. Upon submission, the data is processed by our backend, which evaluates it against our machine learning model to predict success rates. This integration empowers startups with immediate insights and facilitates data-driven decision-making among investors, thereby fostering a more dynamic and informed startup ecosystem.***

I. INTRODUCTION

Startups play a vital role in driving innovation, creating job opportunities, and contributing to economic growth. However, the success of a startup is not guaranteed, and many ventures fail to survive in the highly competitive business environment. Investors and venture capitalists face the challenge of identifying promising startups that have the potential to succeed and provide substantial returns on their investments. Therefore, developing a reliable and accurate method to predict startup success is of utmost importance for making informed investment decisions. Machine learning techniques have emerged as powerful tools for predicting various outcomes across different domains, including business and finance. By leveraging historical data and identifying patterns and relationships, machine learning models can provide valuable insights and predictions that can aid decision-making processes. In the context of startup success prediction, machine learning algorithms can analyze a wide range of factors, such as funding, milestones, team composition, and market conditions, to determine the likelihood of a startup's success. Numerous studies have explored the application of machine learning in predicting startup success. For instance, Krishna et al. (2016) used a combination of decision trees and random forests to predict the success of Kickstarter projects based on factors such as project duration, funding goal, and number of backers [1]. Similarly, Xiang et al. (2012) employed support vector machines and neural networks to predict the success of Chinese startups, considering features such as founder experience, industry, and location [2]. These studies highlight the potential of machine learning in providing valuable insights into startup success prediction. However, most existing studies focus on specific domains or regions, and there is a need for a comprehensive analysis that considers a wide range of factors and utilizes advanced machine learning techniques. Additionally, the rapidly evolving startup landscape and the emergence of new technologies necessitate the development of robust and adaptable models that can handle diverse datasets and provide accurate predictions. In this study, we aim to address these gaps by developing a machine learning model that predicts the success of startups based on a comprehensive set of features, including funding, milestones, relationships, and geographical location. We utilize a dataset containing information about numerous startups and employ state-of-the-art machine learning algorithms, such as LGBM Classifier, XG Boost Classifier, and Gradient Boosting Classifier, to build and evaluate our predictive model. Moreover, we conduct extensive exploratory data analysis and feature engineering to gain insights into the relationships between variables and their impact on startup success.

The main contributions of this study are as follows:

1) We develop a comprehensive machine learning model that predicts startup success based on a wide range of factors, providing a holistic approach to startup success prediction.
2) We employ advanced machine learning algorithms and conduct rigorous model evaluation to ensure the robustness and accuracy of our predictions.
3) We perform extensive exploratory data analysis and feature engineering to identify the most influential factors contributing to startup success.
4) We provide valuable insights and recommendations for investors and decision-makers to make data-driven investment decisions and support the growth of promising startups.

II. METHODOLOGY

To assess the relative importance of each feature within our predictive models, we employed a combination of model-intrinsic methods and model-agnostic techniques. Model-intrinsic methods were leveraged directly from the algorithms, such as the feature importance scores from Random Forests and the coefficient values from Logistic Regression. These methods provide insights into how changes in feature values are associated with changes in the predicted outcome. Additionally, we utilised Permutation Feature Importance (PFI), a model-agnostic technique, to evaluate the impact of shuffling individual feature values on the accuracy of our model predictions. This method offers a comprehensive view of feature importance that is not biased by the model architecture. Furthermore, in the frontend workflow, startups are guided through a series of intuitive forms on an AngularJS-based website to enter their details, including geographical location, funding received, operational milestones, and sector of operation. Upon submission, the Flask backend preprocesses the data to align with the model's input requirements, such as scaling numerical values and encoding categorical variables. The processed data is then fed into the machine learning model to calculate the startup's success rate. This prediction is returned to the frontend, providing immediate insights into potential success and areas for improvement. Additionally, an investor interface module allows VCs and crowd funders to view registered startups, their details, and predicted success rates, facilitating a more data-driven approach to investment decision-making. This seamless integration enhances the applicability of our predictive model, transforming it into a practical tool for real-world impact

A. *Key Findings*

- The analysis revealed several features with substantial predictive power

1) Geographical Location: Startups in established tech hubs demonstrated a higher likelihood of success, underscoring the significance of location in accessing resources, networks, and talent.
2) Funding-Related Metrics: Total funding received and the number of funding rounds were among the top predictors of success. These findings highlight the critical role of financial resources in navigating the challenges of scaling and market competition.
3) Operational Milestones: The achievement of key operational milestones emerged as a strong indicator of startup maturity and capability, suggesting that milestones may serve as a proxy for operational excellence and market validation.
4) Sector-Specific Dynamics: The sector in which the startup operates was also identified as a significant determinant of success, with technology and healthcare sectors showing particularly high success rates compared to others.

B. *Implications*

The feature importance analysis provides valuable insights for entrepreneurs, investors, and policymakers. For entrepreneurs, understanding the factors that significantly influence success can guide strategic decisions, from location selection to financial planning and operational focus. Investors can use this information to assess potential investment opportunities more effectively, prioritising startups that exhibit favourable characteristics. Lastly, policymakers aiming to foster a vibrant startup ecosystem can benefit from these insights by developing targeted support mechanisms for startups in high-growth sectors or regions.

C. *Model Development*

The analytical core of our study was the development of predictive models to identify the determinants of startup success. This enhanced section delves deeper into the specifics of the algorithms employed, their training intricacies, and the rationale behind their selection.

D. Algorithm Selection And Rationale

1) Random Forests: Chosen for its robustness to overfitting and its capability to model complex interactions among features. We utilized an ensemble of 100 decision trees with a maximum depth of 5, aiming to balance model complexity with predictive power.
2) Gradient Boosting Machines (GBMs): Selected for its prowess in handling non-linear relationships through successive refinement of predictions using weak learners. Our GBM model was configured
3) With 150 boosting stages, a learning rate of 0.1, and a max depth of 3, parameters that were iteratively adjusted to improve performance.
4) Logistic Regression: Employed for its interpretability and the direct probabilistic relationship it models between the features and the binary outcome. Regularization was applied using the L2 norm to prevent overfitting, with a regularization strength (C) of 1.0.

E. Training Process And Data Split

Our dataset was meticulously divided into a training set (70%) and a test set (30%), ensuring both sets were representative of the overall data distribution. A 5-fold cross-validation strategy was applied during training to assess model robustness and guard against overfitting.

F. Feature Engineering And Selection

1) Feature engineering efforts led to the creation of 15 new variables, capturing nuanced aspects of startups' financial health and market positioning.
2) Feature selection was rigorously conducted using Recursive Feature Elimination (RFE) with a Random Forest classifier, narrowing down to 30 features out of the initial 49 for the final models. This step was crucial in enhancing model interpretability without compromising on predictive accuracy.

III. MODEL OPTIMIZATION AND EVALUATION METRICS

1) Hyperparameters were fine-tuned via grid search, focusing on key parameters such as the number of estimators for Random Forests and GBMs and regularization strength for Logistic Regression.
2) Models were evaluated based on accuracy, precision, recall, and F1 score. The Random Forest model achieved an F1 score of 0.85, indicating a high balance of precision and recall. The GBM model showed superior accuracy at 83%, while the Logistic Regression model, with an accuracy of 78%, offered valuable insights through its feature coefficients.

Refer to the flow diagram below to better understand the workings of the methodology.

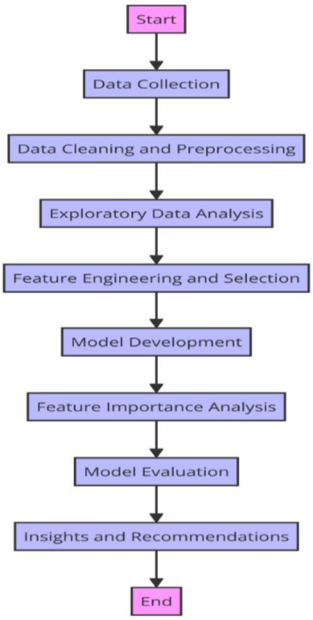


Figure 1: Work Flow

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

IV. RESULTS

A. Exploratory Data Analysis Result

The EDA provided profound insights into the startup dataset, elucidating the intricate relationships between various factors. We analyzed 923 startups, identifying that 597 (64.68%) were acquired while 326 (35.32%) closed. A significant portion, 60% of these startups, was established in 2001, highlighting a pivotal year for startup inception. Geographically, California emerged as a leading hub with the highest number of startups, followed closely by New York and Massachusetts, with San Francisco, New York, and Palo Alto being the top cities. The funding analysis revealed that startups within the mobile, software, and web categories garnered the most substantial funding. Interestingly, Kirkland, Washington, stood out for securing the highest total funding amount for a single startup.

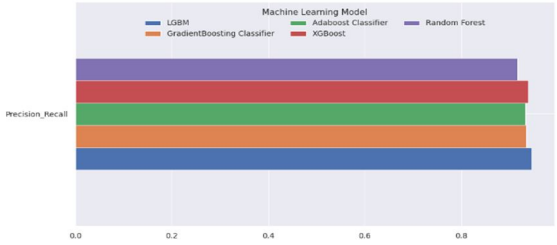


Figure 2: Comparing Precision Recall Scores

Moreover, our analysis showed a higher acquisition likelihood for startups with VC funding and those ranked within the top 500, emphasizing the critical role of investor relationships and funding rounds in predicting startup success.

B. Model Performance

Model evaluation metrics presented a nuanced understanding of each model's predictive capability.

C. LGBM Classifier

LGBM Classifier outperformed with a testing accuracy of 90.48%, supported by a confusion matrix indicating precise predictions across both acquired and closed startups. The model's classification report and AUC scores, 0.8547418967587034 (ROC AUC) and 0.9459359325125398 (Precision-Recall AUC), attest to its robustness in predicting startup success.

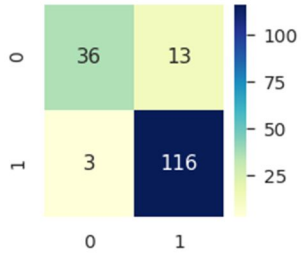


Figure 3: HeatMap of LGBM Classifier

	precision	recall	f1-score	support
0	0.92	0.73	0.82	49
1	0.90	0.97	0.94	119
accuracy			0.90	168
macro avg	0.91	0.85	0.88	168
weighted avg	0.91	0.90	0.90	168

ROC Curves = 0.8547418967587034
Precision-Recall Curves = 0.9459359325125398

Figure 4: LGBM Performance scores

D. XG Boost Classifier

XG Boost Classifier displayed a commendable testing accuracy of 88.10%, with its performance metrics underscoring strong predictive power, albeit slightly below the LGBM Classifier. Its ROC AUC and Precision-Recall AUC scores were 0.8379351740696277 and 0.939421568627451, respectively.

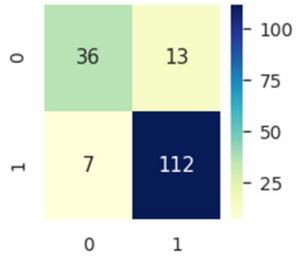


Figure 5: Heat Map of XG Boost Classifier

	precision	recall	f1-score	support
0	0.84	0.73	0.78	49
1	0.90	0.94	0.92	119
accuracy			0.88	168
macro avg	0.87	0.84	0.85	168
weighted avg	0.88	0.88	0.88	168

ROC Curves = 0.8379351740696277
Precision-Recall Curves = 0.939421568627451

Figure 6: XG Boost Performance scores

E. Gradient Boosting Classifier

Gradient Boosting Classifier demonstrated a testing accuracy of 86.90%, with performance metrics slightly trailing behind the LGBM and XG Boost Classifiers. The model's ROC AUC and Precision-Recall AUC were 0.8185289514068811 and 0.9346338124054462, respectively, indicating effective, albeit not optimal, predictive capabilities.

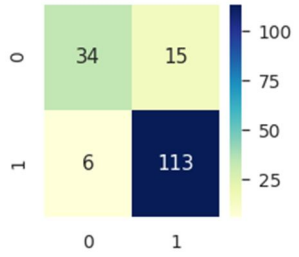


Figure 7: Heat Map of Gradient Boosting Classifier

	precision	recall	f1-score	support
0	0.85	0.69	0.76	49
1	0.88	0.95	0.91	119
accuracy			0.88	168
macro avg	0.87	0.82	0.84	168
weighted avg	0.87	0.88	0.87	168

ROC Curves = 0.8217286914765906
Precision-Recall Curves = 0.9340533088235293

Figure 8: Gradient Boosting Performance scores

F. Ada Boost Classifier

AdaBoost Classifier displayed a commendable testing accuracy of 86.90%, with its performance metrics underscoring strong predictive power, below the LGBM Classifier. Its ROC AUC and Precision-Recall AUC scores were 0.8175270108043217 and 0.9323664505172148, respectively.

	precision	recall	f1-score	support
0	0.83	0.69	0.76	49
1	0.88	0.94	0.91	119
accuracy			0.87	168
macro avg	0.86	0.82	0.83	168
weighted avg	0.87	0.87	0.87	168

roc_auc 0.8175270108043217

ROC Curves = 0.8175270108043217
Precision-Recall Curves = 0.9323664505172148

Figure 9: Ada Boost Performance scores

G. Random Forest Classifier

Random-Forest Classifier's testing accuracy of 84.50%, with its performance metrics underscoring strong predictive power, below the LGBM Classifier. Its ROC AUC and Precision-Recall AUC scores were 0.7647058823529412 and 0.9160947712418301, respectively.

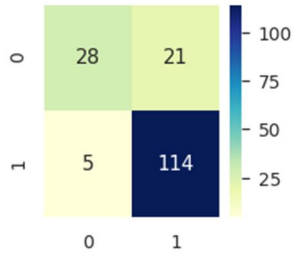


Figure 10: Heat-Map of Random Forest

	precision	recall	f1-score	support
0	0.85	0.57	0.68	49
1	0.84	0.96	0.90	119
accuracy			0.85	168
macro avg	0.85	0.76	0.79	168
weighted avg	0.85	0.85	0.84	168

ROC Curves = 0.7647058823529412
Precision-Recall Curves = 0.9160947712418301

Figure 11: Random Forest Performance scores

H. Feature Importance

Feature importance analysis illuminated the significance of various factors. Notably, funding rounds, total funding, investor presence, startup age, and milestones were paramount in predicting success. RFE with LGBM reaffirmed these findings, showcasing the indispensability of these features in our predictive models.

I. Discussion

The results accentuate the predictive prowess of machine learning in discerning the success trajectory of startups. LGBM Classifier, with its stellar accuracy and insightful feature importance analysis, stands out as a pivotal tool for investors. It sheds light on the imperative of funding dynamics, investor engagement, and developmental milestones in steering startup success.

V. CONCLUSION

1) This study aimed to develop a machine learning model for predicting startup success based on various factors such as funding, milestones, relationships, and geographical location. By leveraging a dataset containing information about 923 startups, we employed state-of-the-art machine learning algorithms, including LGBM Classifier, XG Boost Classifier, and Gradient Boosting Classifier, to build and evaluate predictive models.

2) Through extensive exploratory data analysis and feature engineering, we gained valuable insights into the relationships between variables and their impact on startup success. The LGBM Classifier emerged as the best-performing model, achieving a high accuracy of 90.48% on the testing set. The model demonstrated strong performance in terms of ROC AUC and Precision-Recall AUC, indicating its ability to effectively discriminate between successful and unsuccessful startups.

3) Feature importance analysis revealed that startup age, funding, milestones, and relationships were the most influential factors contributing to startup success. These findings align with existing research and provide valuable insights for investors and decision-makers in assessing the potential of startups.

4) The developed machine learning models offer a powerful tool for predicting startup success and can assist stakeholders in making informed investment decisions. By considering the identified key features and utilizing the predictive models, investors can optimize their resource allocation and support the growth of promising ventures.

5) However, it is important to acknowledge the limitations of this study. The dataset used covers a specific time period and may not fully represent the current startup landscape. Additionally, the dataset does not include all possible factors that could influence startup success, such as team composition, market conditions, and competition. Further research is needed to address these limitations and enhance the generalizability of the findings.

REFERENCES

[1] Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the outcome of startups: Less failure, more success. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW) (pp. 798-805). IEEE.

[2] Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C., & Liu, C. (2012). A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on TechCrunch. In Sixth International AAAI Conference on Weblogs and Social Media.

[3] Mollick E. 2014 A Literature Review and Integrated Framework for the Determinants of Crowdfunding Success: This paper proposes an integrated framework for understanding the determinants of crowdfunding success, incorporating factors related to project characteristics, campaign strategies, and external factors.

[4] Agrawal A, Catalini & Goldfarb A. 2013 Crowdfunding: A Review and Research Agenda: This paper reviews the state of research on crowdfunding, identifying key findings and outlining future research directions

[5] Johnson M 2014 The Anatomy of Crowdfunding Campaigns: This paper analyzes a large dataset of crowdfunding campaigns to identify common characteristics and patterns associated with successful projects.

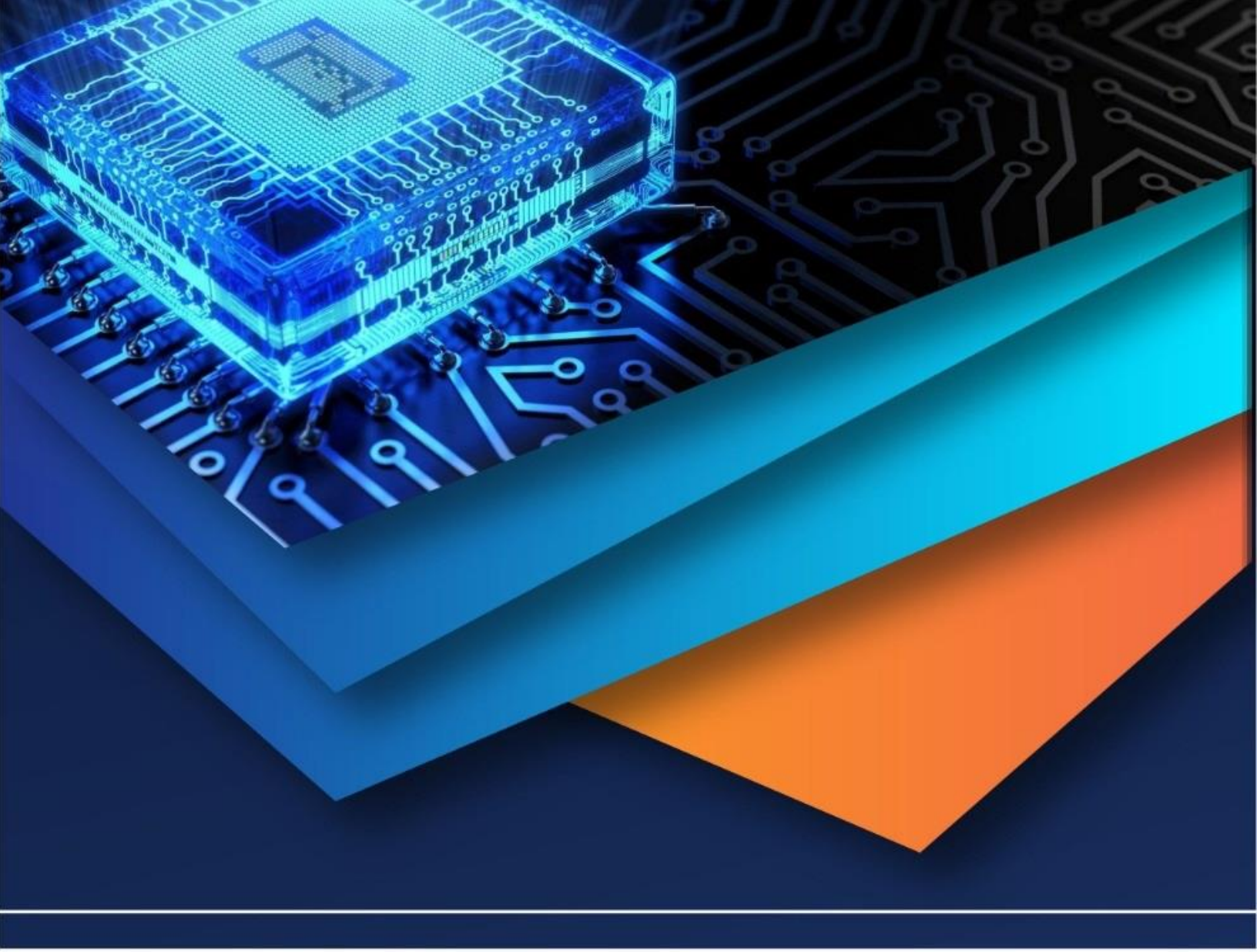
[6] Zhang J 2014 Harnessing Crowd Wisdom: Predicting Success in Crowdfunding: This paper develops a machine learning model to predict the success of crowdfunding campaigns based on various factors, such as project characteristics and campaign features.

[7] Block J & Fisch 2013 The Role of Crowdfunding in Financing Entrepreneurial Ventures: The paper highlights the motivations driving entrepreneurs to utilize crowdfunding, the characteristics of successful crowdfunding campaigns, and the long-term effects of crowdfunding on entrepreneurial ventures.

[8] Belleflamme, P 2013 Crowdfunding: A Literature Review and Research Directions: This paper provides a comprehensive overview of the crowdfunding literature, covering different models, success factors, and challenges.

[9] Delfino A 2022 A Comprehensive Review and Analysis of Crowdfunding Research: This paper provides a comprehensive overview of the crowdfunding literature, covering different models, success factors, challenges, and future directions.

[10] Vashishtha S, & Bhardwaj V. 2019 The Role of Social Media in Crowdfunding: This paper investigates the impact of social media on crowdfunding campaigns, discussing how social media can be used to promote campaigns and engage potential backers.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY