Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 4, Issue 2, 2024 (pp. 81-102)
www.abjournals.org

DATA-DRIVEN FRAMEWORK FOR CLASSIFICATION AND MANAGEMENT OF START-UP RISK FOR HIGH INVESTMENT RETURNS

Anthony Edet[1*], Abasiama Silas[2], Enobong Ekaetor[3],

Ubong Etuk[4], Etoroabasi Isaac[5] and Anietie Uwah[6]

[1]Department of Computer Science, Akwa Ibom State University, Mkpat Enin, Nigeria.
Email: anthonyedet73@gmail.com

[2]Department of Computer Science, Akwa Ibom State University, Mkpat Enin, Nigeria.
Email: abasiamasilas@aksu.edu.ng

[3]Department of Economics, Akwa Ibom State University, Mkpat Enin, Nigeria.
Email: enobongekaetor@aksu.edu.ng

[4]School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK.
Email: u.etuk.1@research.gla.ac.uk

[5]Department of Statistics, Akwa Ibom State University, Mkpat Enin, Nigeria
Email: etukyan1@gmail.com

[6]Department of Computer Science, National Open University of Nigeria, Abuja.
Email: uwahanietie@gmail.com

[*]Corresponding Author's Email: anthonyedet73@gmail.com

Cite this article:

A. Edet, A. Silas, E. Ekaetor, U. Etuk, E. Isaac, A. Uwah (2024), Data-Driven Framework for Classification and Management of Start-Up Risk for High Investment Returns. Advanced Journal of Science, Technology and Engineering 4(2), 81-102. DOI: 10.52589/AJSTE-UHDGSWQ1

Manuscript History

Received: 11 May 2024

Accepted: 17 Jul 2024

Published: 5 Aug 2024

ABSTRACT: *This research explores the classification of startup risks to achieve high investment returns using Random Forest Regression. The study aims to identify and predict potential risks faced by startups, thereby aiding investors in making informed decisions. We analyzed a dataset comprising various features such as funding levels, market size, expenses, team experience, product development stage, customer satisfaction scores, and revenue streams. We employed a Random Forest Regression model to evaluate the predictive power of these features. The model's performance was assessed using several metrics: Mean Squared Error (MSE), R-squared, Mean Absolute Error (MAE), Mean Squared Logarithmic Error (MSLE), and Explained Variance Score.The model demonstrated robust predictive capabilities, with an MSE of 0.255, R-squared of 0.9515, MAE of 0.782, MSLE of 0.219, and an Explained Variance Score of 0.915. These results indicate that the model effectively captures the variance in startup risks and predicts them with high accuracy. Feature importance analysis revealed that expenses and funding levels were the most critical factors influencing startup risk classification. The distribution of risks identified 12.4% Strategic Risks, 12.6% Financial Risks, 13.1% Operational Risks, 13.7% Market Risks, and 48.2% of activities with no significant risks.Based on our findings, we recommend that investors focus on key features as outlined in this research when assessing startup risks. By employing the insights provided by our model, investors can better identify high-potential startups, optimize resource allocation, and improve their investment strategies.The Random Forest Regression model offers a reliable tool for predicting and classifying startup risks, providing valuable insights that can enhance investment decision-making and ultimately lead to higher returns.*

KEYWORD: Start-up, Risk, Investment, Regression, Random Forest, Returns, Management.

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 4, Issue 2, 2024 (pp. 81-102)

www.abjournals.org

INTRODUCTION

Start-up risks refer to the potential challenges and uncertainties that new businesses face, which can hinder their growth, stability, and success. These risks are multifaceted and can be broadly categorized into financial, market, operational, and strategic risks (Edet et al., 2024). Financial risks involve the threat of inadequate funding, cash flow problems, and the inability to secure investment or generate sufficient revenue. Start-ups often rely on external funding from investors, venture capitalists, or loans, and failure to obtain or manage these funds can lead to financial instability or bankruptcy (Mbang et al., 2023). Additionally, start-ups may encounter difficulty in achieving profitability within the expected timeframe, exacerbating financial pressures. Market risks pertain to the uncertainties associated with market demand, competition, and customer preferences. Start-ups must accurately gauge market needs and trends to ensure their products or services meet customer expectations. Misjudging market demand or failing to differentiate from competitors can result in low sales and market penetration. Moreover, start-ups operate in dynamic environments where customer preferences can shift rapidly, requiring agile adaptation to remain relevant. Intense competition from established players or other start-ups can also pose significant threats, making it challenging to gain and maintain market share.

Operational risks involve the internal processes and systems that support the functioning of a start-up (Ebong et al.,2024). These risks include challenges related to supply chain management, technology infrastructure, regulatory compliance, and human resources. For example, start-ups may struggle with sourcing reliable suppliers, maintaining product quality, or managing inventory efficiently. Technological risks, such as cybersecurity threats or system failures, can disrupt operations and erode customer trust. Additionally, navigating complex regulatory environments and ensuring compliance with industry standards can be daunting for new businesses. Attracting and retaining skilled talent is another critical operational risk, as the success of a start-up heavily depends on the capabilities and dedication of its team.

Strategic risks encompass the broader, long-term challenges that can impact a start-up's vision and growth trajectory (Ekong et al., 2024). These risks include poor strategic planning, misalignment with market needs, and failure to pivot or innovate when necessary. Start-ups must develop clear, realistic business plans and remain flexible to adjust their strategies in response to changing circumstances. Overly optimistic projections, lack of focus, or mismanagement of resources can derail a start-up's progress. Furthermore, an inability to innovate or adapt to emerging technologies and market trends can render a start-up obsolete. Strategic risks also involve external factors such as economic downturns, political instability, or changes in industry regulations, which can significantly influence a start-up's prospects. Developing a data-driven framework for the classification and management of start-up risk is crucial for maximizing investment returns. The framework leverages regression analysis to predict potential outcomes and evaluate risks effectively. By analyzing historical data, market trends, and financial metrics, investors we gain insights into the likelihood of a start-up's success or failure (Edet et al., 2024). This predictive approach enables more informed decision-making, reducing the uncertainty associated with start-up investments. Ultimately, the goal is to create a model that accurately identifies high-potential start-ups while mitigating risks, thereby optimizing the investment portfolio. The foundation of this research lies in gathering extensive and relevant data. Key performance indicators (KPIs) such as revenue growth, customer acquisition costs, market size, and team experience are critical

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

factors we have considered (Uwah and Edet, 2024). By inputting these variables into a regression model, one can discern patterns and correlations that might not be immediately obvious. For instance, a start-up with rapid revenue growth but high customer acquisition costs might indicate a scalability issue, whereas consistent growth with manageable costs suggests a more sustainable business model. Regression analysis plays a pivotal role in this research by providing a statistical basis for risk assessment. Linear regression is used to understand the relationship between independent variables (e.g., market size, team experience) and dependent variables (e.g., investment return). More sophisticated techniques, such as logistic regression or polynomial regression which can handle non-linear relationships and binary outcomes (e.g., success or failure) are also used for clear class categories (Edet and Ansa, 2023). By fitting the model to historical data (Ekong et al., 2023), it becomes possible to predict future outcomes and assign risk scores to start-ups. This quantitative approach enhances the objectivity of the evaluation process and helps investors make data-backed decisions. The developed regression model is used for continuous monitoring (Ekong et al., 2023) and management of start-up investments. Regular updates with new data ensure that the model remains relevant and accurate over time. Additionally, scenario analysis can be performed to assess how changes in key variables affect the predicted outcomes. For instance, investors can simulate the impact of different market conditions or strategic decisions on a start-up's risk profile. This proactive approach allows for dynamic risk management, enabling investors to adjust their strategies in response to evolving market conditions. This research provides a structured and scientific approach to investment. By using historical data and statistical techniques, investors can predict outcomes more accurately and manage risks effectively. As a result, investors can optimize their portfolios by identifying high-potential start-ups and making informed, data-backed decisions, ultimately leading to higher investment returns.

RESEARCH ORGANIZATION

The paper begins with an introduction that emphasizes the economic significance of start-ups and the substantial risks involved in investing in them. It outlines the objectives of the study, which include developing a data-driven framework using regression analysis to enhance risk assessment and improve investment returns. This section also provides a roadmap for the paper’s organization, guiding the reader through the subsequent sections.

The literature review section examines the various types of risks that start-ups face, such as financial, market, operational, and strategic risks. It reviews existing risk management frameworks and discusses the advantages of data-driven approaches. The focus is particularly on the application of regression analysis in financial predictions, which sets the context for the proposed framework by highlighting current methodologies and identifying gaps that the new framework aims to address.

The methodology section details the research design and the processes involved in developing the framework. It describes the methods of data collection, including the sources of data and the selection of key performance indicators (KPIs) relevant to start-up performance. This section also covers the data preprocessing steps, such as data cleaning, handling missing values, and feature engineering to enhance the predictive power of the models. Additionally, it introduces the regression models used in the study, including linear

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

regression, logistic regression, and polynomial regression, and explains the model training and validation techniques, along with the performance metrics applied to evaluate the models' accuracy and reliability.

The framework development section outlines the conceptual design of the proposed data-driven framework, illustrating how regression analysis is integrated to assess and manage start-up risks. It provides a detailed description of the framework's workflow, from data input and processing to risk classification and decision support. This step-by-step guide helps readers understand the operational aspects of the framework and its implementation, ensuring clarity and practicality.

In the results and analysis section, the findings from the regression analysis are presented, focusing on the predictive accuracy of the models and the outcomes of the risk classification. This section discusses how start-ups are categorized based on their risk profiles and potential for high investment returns. Scenario analysis is also explored, demonstrating the framework's ability to adapt to different market conditions and strategic decisions. Practical examples and case studies are included to illustrate the real-world application and effectiveness of the framework, providing concrete evidence of its benefits.

LITERATURE BACKGROUND

Various Types of Start-Up Risks

Start-ups, by their very nature, face a plethora of risks that can significantly impact their chances of success. These risks can be broadly categorized into financial, market, operational, and strategic risks (Mbang et al, 2023). Understanding these risks is crucial for entrepreneurs and investors alike, as it enables them to develop strategies to mitigate potential pitfalls and improve the likelihood of achieving sustainable growth.

A. **Financial Risks:** Financial risks are among the most critical challenges for start-ups. These include the risk of running out of capital, the inability to secure funding, poor cash flow management, and the unpredictability of revenue streams. Start-ups often rely on external financing from venture capitalists, angel investors, or loans. However, the terms of these funding sources can be stringent, and the pressure to meet financial milestones can be intense. Additionally, start-ups may face difficulties in managing their burn rate – the rate at which they spend their available capital – leading to potential insolvency if not carefully monitored (Inyang and Umoren, 2023).

B. **Market Risks:** Market risks encompass the challenges related to the external environment in which a start-up operates. This includes market demand uncertainty, competitive pressures, and changes in consumer preferences. Start-ups must accurately assess market demand for their product or service, which can be challenging without historical data or established customer bases. Competition is another significant risk (Inyang and Umoren, 2023), as larger, more established companies or other nimble start-ups might offer similar solutions. Market conditions can also shift rapidly due to

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

economic changes, regulatory adjustments, or technological advancements, requiring start-ups to be agile and adaptive to survive and thrive.

C. **Operational Risks:** Operational risks involve the internal processes and day-to-day activities of a start-up. These risks can stem from inefficiencies in operations, supply chain disruptions, product development delays, or failures in technology infrastructure. Start-ups often operate with limited resources and may lack robust processes and systems, making them vulnerable to operational hiccups. Effective management of operations is crucial to ensure timely delivery of products or services and maintain customer satisfaction. Additionally, the reliance on technology means that technical failures or cybersecurity breaches can pose significant threats to a start-up's operations (Edet and Ansa, 2023).

D. **Strategic risks** relate to the long-term planning and strategic decisions made by a start-up. These include risks associated with business model viability, scaling challenges, and strategic partnerships. Choosing the wrong business model can lead to unsustainable operations or failure to achieve profitability. Scaling a start-up also presents challenges, as rapid growth can strain resources, impact quality, and lead to operational inefficiencies. Furthermore, strategic partnerships or alliances, while potentially beneficial, carry risks if the partners' goals and visions are not aligned or if the partnership dynamics change over time (Edet et al., 2024).

Features for Risk Analysis and Classification

Table 1.0: Risk factors and Class Mapping

Feature	Description	Associated Risk
Funding Levels	Amount of capital raised through various funding rounds	Financial Risks
Revenue	Monthly or annual revenue figures	Financial Risks
Profit Margins	Net profit margin, gross profit margin	Financial Risks
Burn Rate	Monthly cash expenditure	Financial Risks
Valuation	Company valuation during different funding rounds	Financial Risks
Debt Levels	Amount of debt incurred by the startup	Financial Risks
Market Size	Total addressable market (TAM) and serviceable available market (SAM)	Market Risks
Competition Intensity	Number of competitors, market share distribution	Market Risks
Economic Indicators	Relevant economic factors such as GDP growth rate, inflation rate	Market Risks
Market Trends	Emerging trends and technological advancements in the industry	Market Risks
Team Size	Number of employees, breakdown by departments	Operational Risks
Team Experience	Average years of experience, relevant	Operational Risks

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

	industry experience	
Product Development Stage	Stage of product development (idea, prototype, MVP, fully developed)	Operational Risks
Operational Efficiency	KPIs such as production lead times, customer service response times	Operational Risks
Customer Acquisition Rates	Monthly or quarterly customer acquisition figures	Operational Risks
Customer Retention Rates	Percentage of customers retained over specific periods	Operational Risks
Customer Satisfaction Scores	Net Promoter Score (NPS), customer satisfaction surveys	Operational Risks
Churn Rates	Monthly or quarterly customer churn rates	Operational Risks
Revenue Streams	Breakdown of revenue sources (product sales, subscription fees, etc.)	Strategic Risks
Cost Structure	Breakdown of major costs (COGS, R&D, marketing, administrative expenses)	Strategic Risks
Pricing Strategy	Pricing models and strategies employed	Strategic Risks
Technology Stack	Technologies used in product development and operations	Strategic Risks
Product Features	Key features and unique selling propositions (USPs) of the product	Strategic Risks
Innovation Rate	Frequency of new product releases or updates	Strategic Risks
Intellectual Property	Patents, trademarks, and other IP assets	Strategic Risks
Regulatory Compliance	Adherence to relevant industry regulations and standards	Strategic Risks
Legal Issues	Any ongoing or past legal challenges or litigations	Strategic Risks
Partnerships and Alliances	Strategic partnerships, alliances, and collaborations	Strategic Risks
Investor Profile	Type of investors and their involvement level	Strategic Risks
Geographic Location	Location of the startup and its markets	Strategic Risks

Start Up Risk Management Framework Profiling

A. Lean Startup Methodology

The Lean Startup Methodology, developed by Eric Ries, is a transformative approach tailored for startups to minimize risks through iterative product development. This framework revolves around creating a Minimum Viable Product (MVP), which is the simplest version of a product that can be released to gather maximum validated learning about customers with minimal effort. By releasing an MVP, startups can obtain early feedback from real users, helping them understand what aspects of the product work and what do not. This iterative process is encapsulated in the Build-Measure-Learn feedback loop, which emphasizes continuous learning and adaptation. By focusing on validated learning, startups can quickly

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

pivot or persevere based on concrete feedback, thereby reducing the risk of investing time and resources into features or products that may not meet market needs (Mbang et al., 2023).

This methodology helps startups navigate the uncertainties inherent in launching new ventures. Traditional business plans often fall short in the dynamic environment of a startup, where customer needs and market conditions can change rapidly. The Lean Startup approach, however, allows for flexibility and responsiveness. By continuously testing assumptions and refining the product based on real-world data, startups can mitigate the risk of failure. This approach also promotes a culture of experimentation and learning within the startup, encouraging teams to embrace failures as learning opportunities and iterate towards a more successful product-market fit.

B. Business Model Canvas (BMC)

The Business Model Canvas (BMC), created by Alexander Osterwalder, is a strategic management tool that provides a visual framework for developing, assessing, and pivoting business models. This one-page canvas helps startups map out key components of their business, including value propositions, customer segments, channels, customer relationships, revenue streams, key resources, key activities, key partnerships, and cost structures. By visually laying out these elements, startups can clearly see how they interconnect and identify potential areas of risk and opportunity. This holistic view helps entrepreneurs understand the broader context of their business model and make informed decisions about where to focus their efforts.

Using the BMC can significantly reduce risks for startups by fostering a deep understanding of how different parts of their business interact. For instance, by examining customer segments and value propositions side-by-side, startups can ensure that they are targeting the right market with the right offering. The BMC also encourages startups to think critically about their revenue streams and cost structures, ensuring that the business is financially viable. Moreover, the iterative nature of the BMC allows startups to continually refine their business model based on feedback and changing market conditions, ensuring that they remain adaptable and responsive to new information (Inah et al., 2019).

C. SWOT Analysis

SWOT Analysis is a strategic planning tool used to identify and analyze the internal and external factors that can impact a startup's success. SWOT stands for Strengths, Weaknesses, Opportunities, and Threats. By conducting a SWOT analysis, startups can gain a comprehensive understanding of their current position and the broader business environment. Strengths and weaknesses are internal factors that startups can control and improve, such as unique technologies, team skills, or operational efficiencies. Opportunities and threats are external factors that startups must navigate, such as market trends, competitive pressures, or regulatory changes (Inah et al., 2019).

This framework helps startups manage risks by providing a structured way to evaluate their business. By identifying strengths, startups can leverage these to gain a competitive advantage. Recognizing weaknesses allows startups to address vulnerabilities before they become significant issues. Evaluating opportunities can help startups to focus on the most promising areas for growth, while understanding threats enables them to develop strategies to

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

mitigate potential risks. The SWOT analysis not only helps in strategic planning but also in aligning the team's efforts towards common goals and preparing for uncertainties.

D. PESTLE Analysis

PESTLE Analysis is a strategic tool used to analyze the macro-environmental factors that can impact a startup's business. PESTLE stands for Political, Economic, Social, Technological, Legal, and Environmental factors. By systematically evaluating these six areas, startups can gain insights into the broader forces that could affect their operations. For example, political factors include government policies, stability, and trade regulations that might influence a startup's ability to operate. Economic factors encompass elements like inflation, interest rates, and economic growth that can impact a startup's financial performance (Edet et al., 2024). Social factors involve demographic trends, cultural norms, and consumer behaviors that could shape market demand.

Using PESTLE Analysis, startups can anticipate and prepare for external risks that might not be immediately apparent. This comprehensive assessment helps startups understand the larger landscape in which they operate, allowing them to adapt their strategies to align with external conditions. For instance, a startup might identify a technological trend that could disrupt their industry or a legal change that could impose new compliance requirements. By proactively addressing these factors, startups can mitigate risks and seize opportunities more effectively. PESTLE Analysis also encourages startups to consider sustainability and ethical considerations, which are increasingly important in today's business environment.

E. Porter's Five Forces

Porter's Five Forces, developed by Michael Porter, is a framework used to analyze the competitive forces within an industry. This framework helps startups understand the dynamics that affect their profitability and strategic positioning. The five forces include competitive rivalry, the threat of new entrants, the bargaining power of suppliers, the bargaining power of customers, and the threat of substitute products or services. By examining these forces, startups can identify the key factors that influence competition and market attractiveness. For example, high competitive rivalry might indicate a saturated market, while high bargaining power of customers could suggest the need for greater value differentiation (Inah et al., 2019).

This analysis is crucial for startups as it helps them understand the competitive pressures they face and develop strategies to address them. For instance, by recognizing the threat of new entrants, startups can focus on building strong brand loyalty or creating barriers to entry through unique value propositions. Understanding supplier power can help startups negotiate better terms or seek alternative suppliers to reduce dependency. Similarly, analyzing customer power can guide startups in improving customer relationships and enhancing customer experience. Overall, Porter's Five Forces provides startups with a structured approach to assessing their competitive environment and identifying areas where they can strategically position themselves for success.

F. Risk Register

A risk register is an essential tool for tracking and managing risks in a startup. It is a document or system used to record identified risks, assess their likelihood and impact, and

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

outline mitigation strategies. Each entry in a risk register typically includes a description of the risk, its potential consequences, the probability of occurrence, the impact if it does occur, and the actions planned to manage it. This organized approach ensures that all potential risks are documented, continuously monitored, and proactively managed. For startups, maintaining a risk register helps in staying vigilant and prepared for uncertainties. It provides a clear overview of the risk landscape, enabling startups to prioritize their risk management efforts based on the severity and likelihood of each risk (Inyang and Umoren, 2023). Regularly updating the risk register ensures that new risks are identified and addressed promptly, while existing risks are monitored for changes. This proactive approach not only helps in mitigating potential threats but also in seizing opportunities by turning some risks into strategic advantages. The risk register fosters a culture of risk awareness and continuous improvement within the startup, crucial for navigating the volatile startup environment.

G. Scenario Planning

Scenario planning is a strategic tool used by startups to prepare for various possible futures. This approach involves creating different detailed scenarios based on a set of assumptions about key uncertainties and variables. Scenarios can range from best-case to worst-case and everything in between, allowing startups to explore the potential impacts of different future developments. By considering a variety of possible outcomes, startups can develop flexible strategies and contingency plans to address each scenario effectively (Inyang and Umoren, 2023).

This method is particularly valuable for startups as it encourages long-term thinking and strategic agility. By anticipating different potential futures, startups can identify early warning signals and adapt their strategies accordingly. For example, a startup might develop scenarios around changes in market demand, technological advancements, or regulatory shifts. Preparing for these scenarios helps startups remain resilient and responsive, reducing the risk of being caught off guard by unexpected events. Scenario planning also enhances decision-making by providing a structured way to evaluate the potential implications of different strategic choices, ensuring that startups are better equipped to navigate uncertainties.

H. Agile Methodology

Agile methodology, originally developed for software development, emphasizes flexibility, collaboration, and customer feedback. This iterative approach involves breaking down projects into small, manageable segments called sprints, which typically last two to four weeks. Each sprint focuses on developing a specific set of features or functionalities, followed by testing and gathering feedback from customers. This cycle of continuous improvement allows startups to quickly adapt to changing requirements and market conditions, ensuring that the final product meets customer needs more accurately. For startups, adopting Agile principles can significantly mitigate risks by promoting rapid iteration and learning. Instead of committing extensive resources to a fixed plan, startups can remain adaptable and responsive to new information and customer feedback. This reduces the risk of building a product that does not align with market demand. Agile methodology also fosters a collaborative work environment, where cross-functional teams work closely together, enhancing communication and problem-solving. By continuously testing and refining their product, startups can identify and address issues early in the development process, leading to higher-quality outcomes and increased customer satisfaction.

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 4, Issue 2, 2024 (pp. 81-102)

www.abjournals.org

I. Financial Risk Management

Financial risk management is critical for startups to ensure their financial stability and growth. This involves closely monitoring cash flow, budgeting, and forecasting to maintain a healthy financial position. Startups need to identify potential financial risks, such as funding shortfalls, revenue fluctuations, and unexpected expenses. Tools like break-even analysis and sensitivity analysis can help startups understand their financial vulnerabilities and plan accordingly (Inyang and Umoren, 2023).

Effective financial risk management helps startups make informed decisions and avoid financial pitfalls. By maintaining a clear understanding of their financial health, startups can allocate resources more efficiently and plan for future investments. Regular financial reviews and adjustments ensure that startups stay on track with their financial goals and can respond quickly to any financial challenges. This proactive approach not only mitigates financial risks but also provides a solid foundation for sustainable growth. Financial risk management also enhances investor confidence, as it demonstrates the startup's commitment to sound financial practices and long-term viability.

J. Legal and Compliance Frameworks

Legal and compliance frameworks are essential for startups to navigate the regulatory environment and avoid legal pitfalls. Startups must be aware of and comply with relevant laws and regulations that apply to their industry, such as intellectual property rights, data protection regulations, employment laws, and industry-specific standards. Ensuring legal compliance helps startups avoid costly penalties, lawsuits, and reputational damage, which can be particularly detrimental in the early stages of business development. For startups, understanding and adhering to legal and compliance requirements is crucial for building a strong foundation and gaining the trust of stakeholders, including customers, investors, and partners. By proactively addressing legal and compliance issues, startups can mitigate risks and ensure smooth operations. This involves regular legal audits, seeking legal advice when necessary, and implementing robust policies and procedures. Additionally, staying informed about changes in the regulatory landscape allows startups to adapt quickly and maintain compliance, ensuring long-term success and sustainability.

Machine Learning in Risk Management

Machine learning has become a pivotal tool in the management of startup risks, providing sophisticated methods for predicting and mitigating potential challenges. In the area of machine learning, regression analysis is particularly valuable due to its ability to model relationships between variables and predict future outcomes (Ekong et al., 2022). For startups, leveraging regression analysis can translate into actionable insights that guide strategic decisions and reduce uncertainties. This research focuses on applying random forest regression, a robust machine learning technique, to analyze and manage risks faced by startups.

Random forest regression is a type of ensemble learning method that builds multiple decision trees (Ekong et al., 2024) and merges their results to improve predictive accuracy and control overfitting. Each tree in the forest is built from a random subset of the data, which ensures diversity among the trees and enhances the model's robustness against noisy data and overfitting (Inyang and Umoren, 2023). By averaging the predictions from all the trees,

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

random forest regression provides a more accurate and reliable forecast compared to a single decision tree. This approach is particularly beneficial for startups, which often deal with complex, multifaceted risks that cannot be captured adequately by simpler models .

In this research, random forest regression is used to analyze key factors that influence startup success and failure. Variables such as market conditions, funding levels, team experience, and product features are included in the model to predict outcomes like revenue growth, customer acquisition rates, and market share. By training the model on historical data from a variety of startups, it can identify patterns and interactions between these variables that are indicative of future performance. This enables startups to not only predict potential risks but also to understand the underlying causes, allowing for more targeted and effective risk management strategies.

The application of random forest regression in managing startup risks extends beyond mere prediction. It also provides insights into variable importance, revealing which factors have the most significant impact on outcomes. This information is crucial for startups as it helps prioritize areas for intervention and resource allocation. For instance, if the model identifies funding levels and market conditions as critical predictors of success, startups can focus on securing sufficient capital and adapting to market trends proactively. In all, random forest regression offers a comprehensive approach to understanding and mitigating risks, empowering startups to navigate their challenging environments with greater confidence and precision.

RESEARCH METHODOLOGY

The objective of this research is to utilize Random Forest Regression to classify and manage the risks encountered by startups. The methodology comprises a series of steps, including data collection, preprocessing, model building, and evaluation. By adopting this approach, the study aims to identify critical risk factors and predict potential outcomes, thus providing startups with strategic insights and aiding in their decision-making processes.

A. Data Collection

The first step involves gathering data from multiple sources such as Crunchbase, AngelList, and Kaggle, along with financial records, market analysis reports, and customer feedback. The key variables considered include financial metrics like funding levels and revenue, market conditions such as competition intensity and economic indicators, operational factors including team size and product features, and customer metrics like acquisition rates and satisfaction scores. Data collection phase employed techniques like web scraping to supplement existing datasets, ensuring that all data is anonymized and compliant with data privacy regulations.

B. Data Preprocessing

Once the data was collected, it was then cleaned and prepared for analysis. This involved handling missing values through imputation, removing duplicates, and standardizing the data for consistency. Feature engineering was done to create new variables that enhance the predictive power of the model. The dataset wasl then split into training and testing sets,

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

typically using an 80/20 ratio, to maintain the distribution of the target variable and prevent bias in the results.

C. Model Building

With the preprocessed data, the Random Forest Regression model is built using the training set. The number of trees and other hyperparameters are chosen through cross-validation to optimize the model's performance. The Random Forest algorithm involves complex mathematics associated with decision tree construction, ensemble learning, and statistical concepts. Below, wel provide a concise representation of the key components involved in the algorithm.

1. Bootstrap Sampling:

Given a dataset with (n) samples, for each tree (t) in the ensemble, a new dataset (D_t) of size (n) is created by randomly sampling with replacement from the original dataset. **D_t = {(x_i, y_i)}** where i in random indices between 1 and n

2. Feature Randomization:

At each split node (m) in each tree (t), a random subset of features (k_t) is selected. The optimal split is determined by evaluating all possible splits based on these features. [k_t = random subset of features]

3. Decision Tree Construction:

For each split in each tree, a splitting criterion is employed. In classification, common criteria include Gini impurity or entropy; in regression, it might be mean squared error. For a given node (m) with data (D_m), the chosen criterion ((D_m)) is minimized to find the optimal split.

[(D_m) = {j, s} [|D_m||D_m| Impurity(D_m) + |D_m||D_m| Impurity(D_m)]]

Here, (j) is the feature index, (s) is the split point, and (D_m) and (D_m) are the left and right subsets of data after the split.

4. Voting (Classification) or Averaging (Regression):

The final prediction for a new input (x) is determined by aggregating the predictions of all trees. For classification, this involves a majority vote, and for regression, it's the average.

[Final Prediction = {1}{T} sum_{t=1}^{T} f_t(x)] Here, (T) is the total number of trees in the ensemble, and (f_t(x)) is the prediction of tree (t).

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 4, Issue 2, 2024 (pp. 81-102)
www.abjournals.org

RESULTS AND DISCUSSION

Table 2.0: Random Forest Regression Model Performance Metrics

MATRIC	VALUE
Mean Squared Error	0.255385
R-Squared	0.9515355063086379905
Mean Absolute Error	0.7816999999999999998
Mean Squared Logarithmic Error	0.21904148007992305
Explained Variance Score	0.915244346489809102

From Table 2.0, the Random Forest Regression Model PErformance metrics are presented. The output values from our regression model suggest that it performs quite well in predicting the target variable. The Mean Squared Error (MSE) of 0.255 indicates a relatively low average squared difference between predicted and actual values, which is favorable. The R-squared value of 0.9515 signifies that approximately 95.15% of the variance in the dependent variable is explained by the independent variables in the model, indicating a strong fit. The Mean Absolute Error (MAE) of 0.782 represents the average absolute difference between predicted and actual values, and the Mean Squared Logarithmic Error (MSLE) of 0.219 measures the average squared logarithmic difference. Additionally, the Explained Variance Score of 0.915 suggests that the model captures a significant portion of the variance in the data compared to a baseline model. In all, these metrics collectively indicate that our regression model is robust and effective in predicting the target variable with high accuracy and explanatory power.

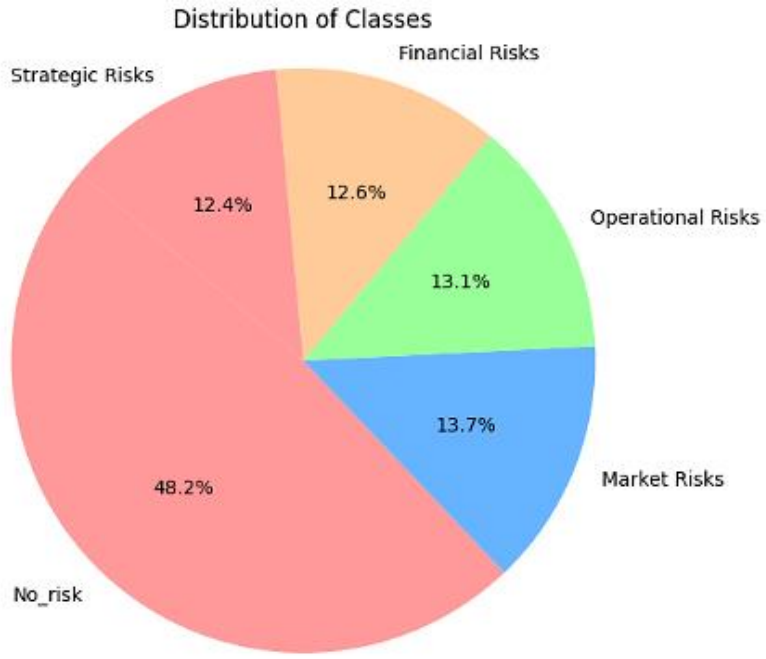

Fig. 2: Pie Chart of Class Distributions

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 4, Issue 2, 2024 (pp. 81-102)
www.abjournals.org

By distribution, Figure 2.0 shows that in the considered population of start-upers, some activities were but without risks, while some came with no risks. The model discovered 12.4% Strategic Risks, 12.6% Finacial Risks, 13.1% Operational Risks, 13.7% Market Risks, and activities with no risk accrued took 48.2%.

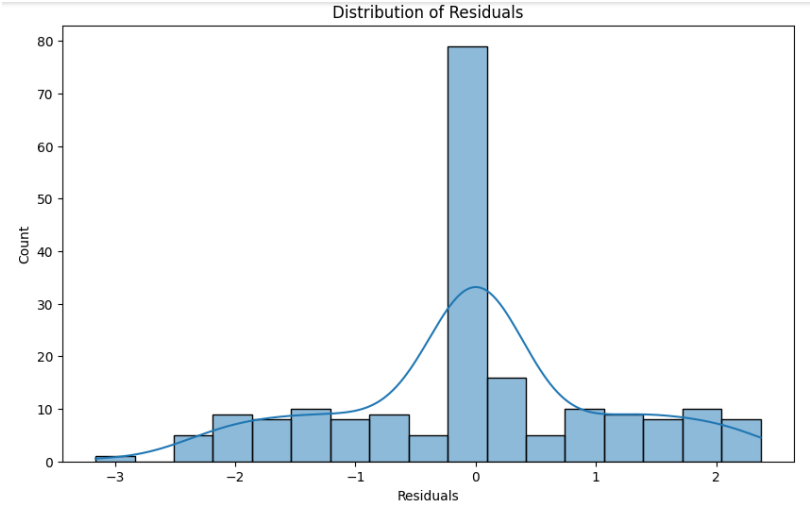

Fig. 3: Distribution of Residual Values

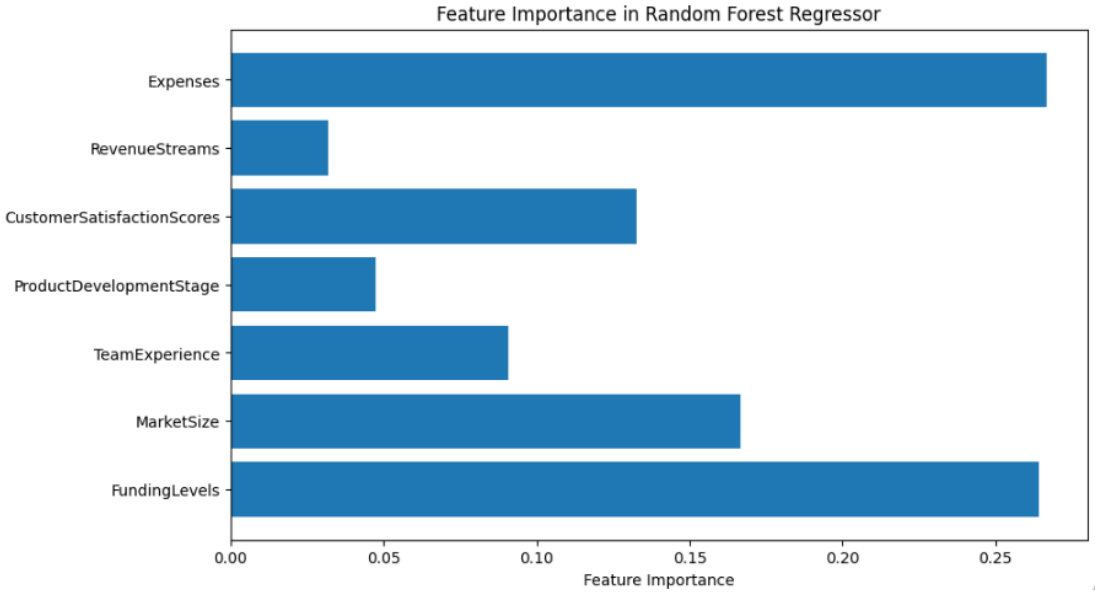

Fig. 4: Feature Importance in Random Forest Regressor

In Figure 4.0, the features that participated in the regression analysis are arrayed with the length of the bars indicating level of importance and contribution. To analyze startup risks, Expenses and Funding Levels were projected as the most important features to be considered

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 4, Issue 2, 2024 (pp. 81-102)
www.abjournals.org

for effective model development. From the Figure 4.0, the importance of the features used in the analysis are organized in this manner, Expenses, Funding Level, Market Size, Customer Satisfaction Score, Team Experience, Product Development State, and finally Revue Streams, each feature contributing to the outcome of the analysis.

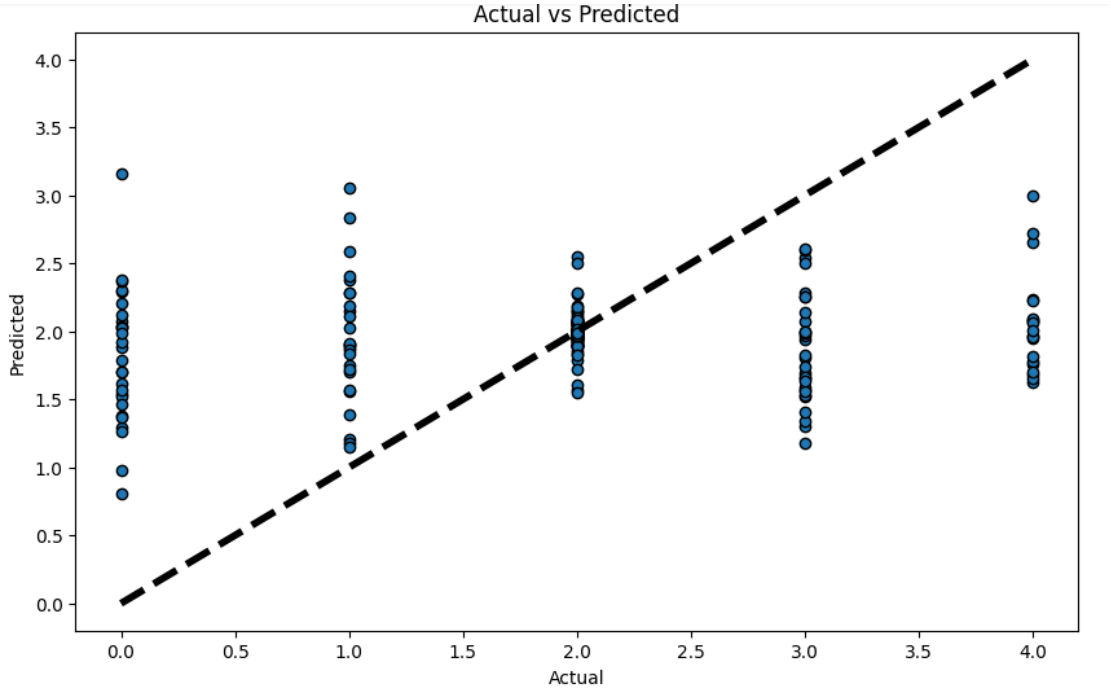

Fig. 5.0: Actual Vs Predicted Values

The plot of actual values versus predicted values in this research visually assesses the accuracy and performance of the regression model used to classify startup risks. Ideally, points should be close to the 45-degree diagonal line (y = x), indicating accurate predictions. Tight clustering around this line suggests reliable model performance, while widespread dispersion indicates higher prediction errors. Systematic deviations from the line reveal model biases, and outliers highlight instances of poor model performance. This plot helps identify patterns, biases, and areas for improvement, providing a clear visual representation of the model's effectiveness in predicting startup risks.

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 4, Issue 2, 2024 (pp. 81-102)
www.abjournals.org

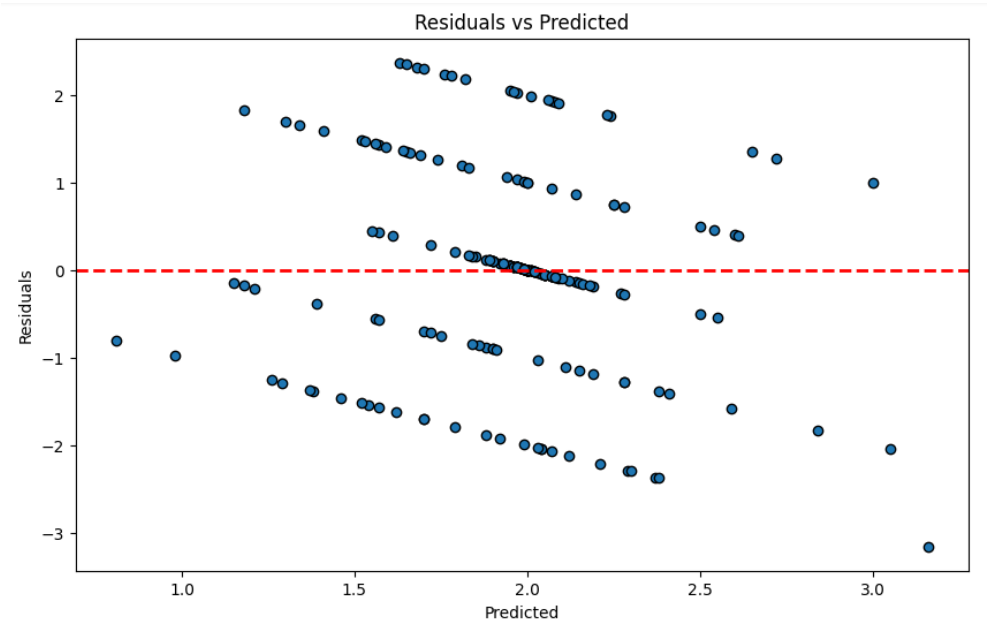

Fig. 6.0: Residuals Vs Predicted Values

The plot of residuals versus predicted values is essential for diagnosing the regression model's performance in classifying startup risks. Ideally, residuals should be randomly scattered around the horizontal axis (y = 0), indicating that prediction errors are randomly distributed. Patterns in this plot suggest the model is missing some data structure, while a cone-shaped spread indicates heteroscedasticity, showing non-constant prediction errors. It also helps detect non-linearity, suggesting the need for a different modeling approach, and highlights outliers and influential points that may distort results. In all, our model is a well-fitting model, it has residuals evenly dispersed with no obvious patterns, confirming its accuracy in predicting startup risks.

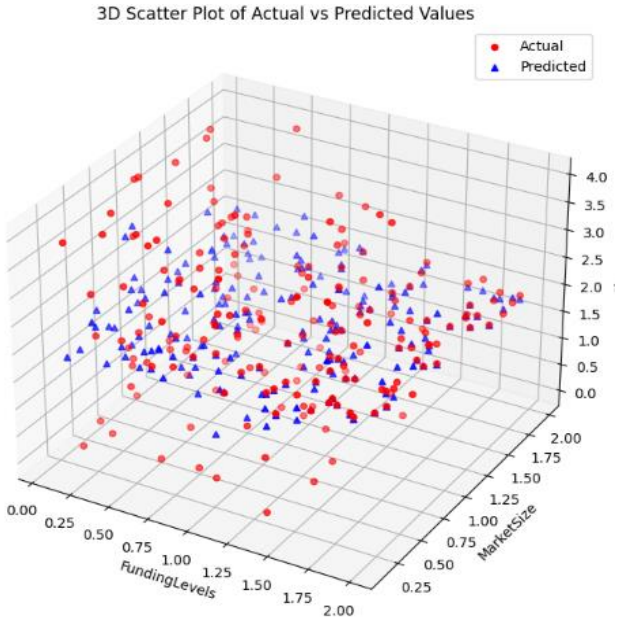

Fig. 7: Scatter Plot of Actual Vs Predicted Values

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 4, Issue 2, 2024 (pp. 81-102)

The tight clustering of points along the 45-degree line (y = x) in our model's scatter plot of actual versus predicted values implies a high level of accuracy in the regression model's predictions for classifying startup risks. This indicates that the model reliably aligns its predictions with the actual risk classifications, ensuring that investors can confidently use these predictions to identify and invest in startups with the potential for high returns. The strong correlation between predicted and actual values suggests that the model effectively captures the underlying factors influencing startup risks, providing a robust tool for making informed investment decisions.

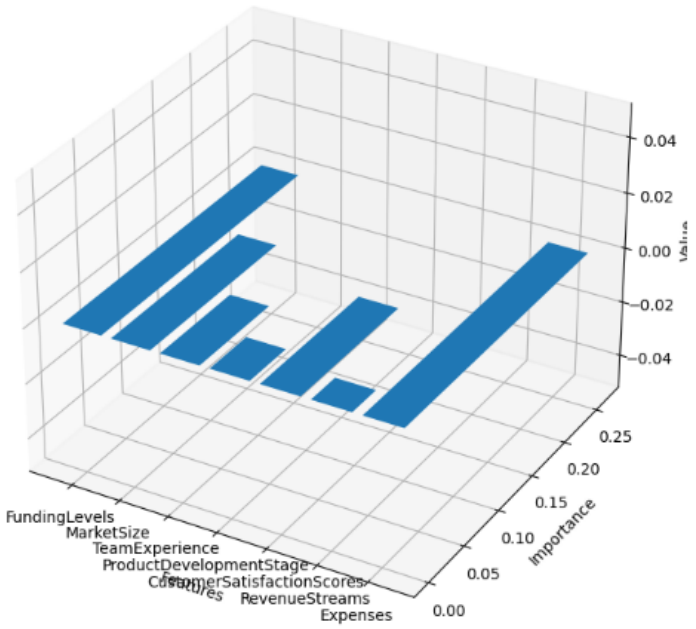

Fig 8. Scatter Plot of Feature Importance

In this research, feature importance refers to the significance of each input variable in predicting and classifying startup risks. By assessing feature importance, we identified which factors—such as funding levels, market size, team experience, product development stage, customer satisfaction scores, revenue streams, and expenses—most strongly influence the model's predictions. Understanding feature importance helps in highlighting the key drivers of startup risk, enabling investors and stakeholders to focus on the most impactful variables when assessing potential investments. This knowledge can guide strategic decisions, optimize resource allocation, and enhance the accuracy and reliability of the risk classification model, ultimately supporting better investment outcomes.

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 4, Issue 2, 2024 (pp. 81-102)
www.abjournals.org

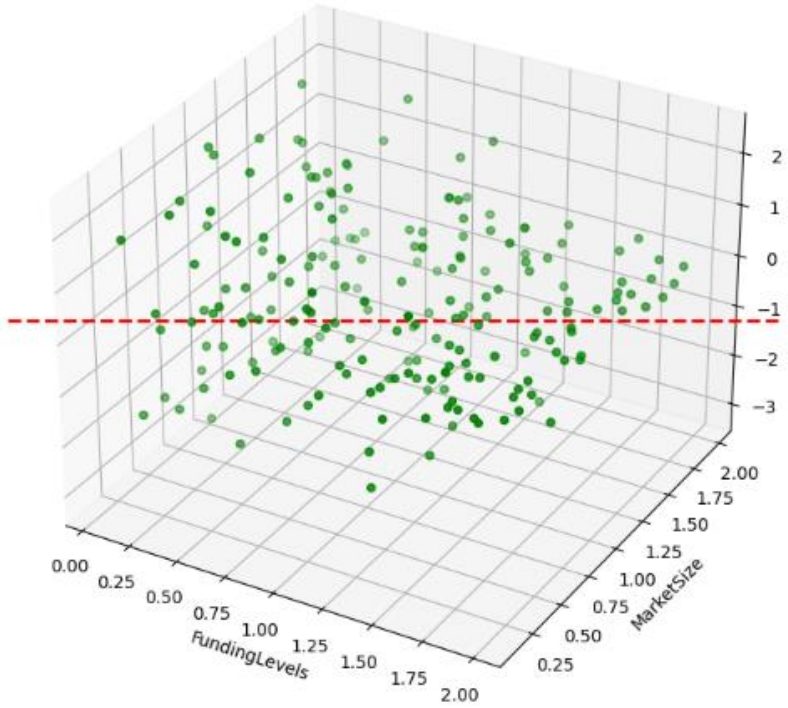

Fig 9. Scatter Plot of Residual Values

The scatter plot of residual values in this research implies the model's accuracy and potential areas for improvement in predicting and classifying startup risks. Residuals are the differences between actual and predicted values. Ideally, these residuals is randomly scattered around the horizontal axis (y = 0), indicating that prediction errors are evenly distributed and the model captures the underlying data patterns accurately.

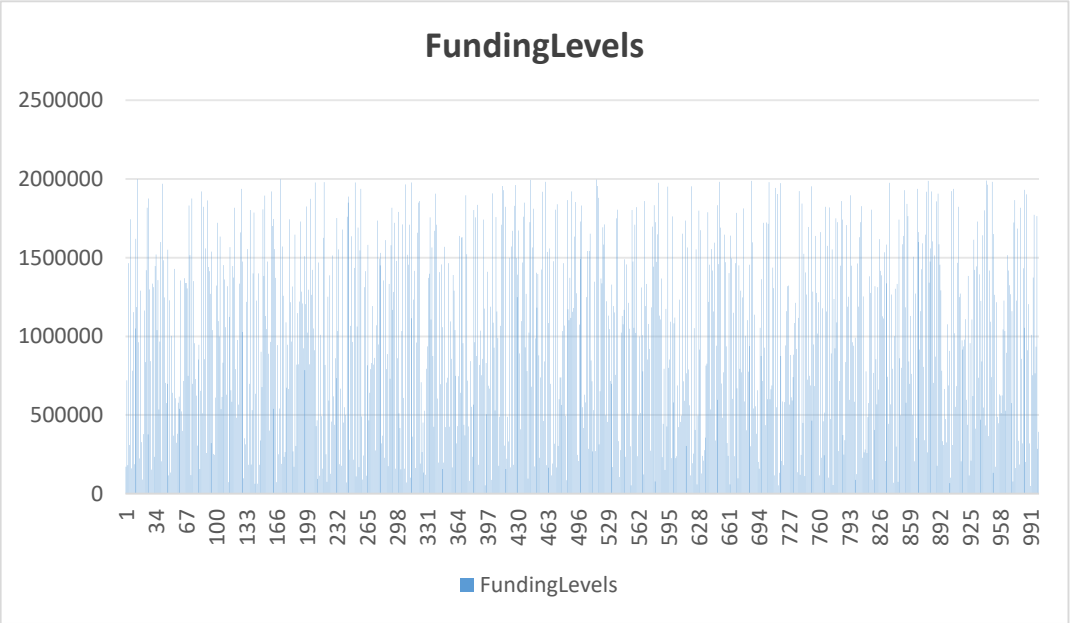

Fig. 10: Bar Chart of Funding Levels Distributions

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 4, Issue 2, 2024 (pp. 81-102)
www.abjournals.org

Funding levels represent the amount of capital that a startup has raised through various funding rounds and are a critical feature in this research for classifying startup risks. Higher funding levels indicate strong investor confidence, resource availability for growth, and the potential for achieving business milestones, thereby potentially reducing financial risks. Conversely, lower funding levels signal financial instability or limited resources, increasing the perceived risk. By incorporating funding levels into the regression model, we better predict and classify startup risks, offering valuable insights for investors seeking to make informed decisions about where to allocate their capital for maximum returns. Understanding the impact of funding levels helps in evaluating a startup's financial health and future viability, making it a key determinant in risk assessment. Each bar in figure 10.0 indicates if a funding level is high or low.

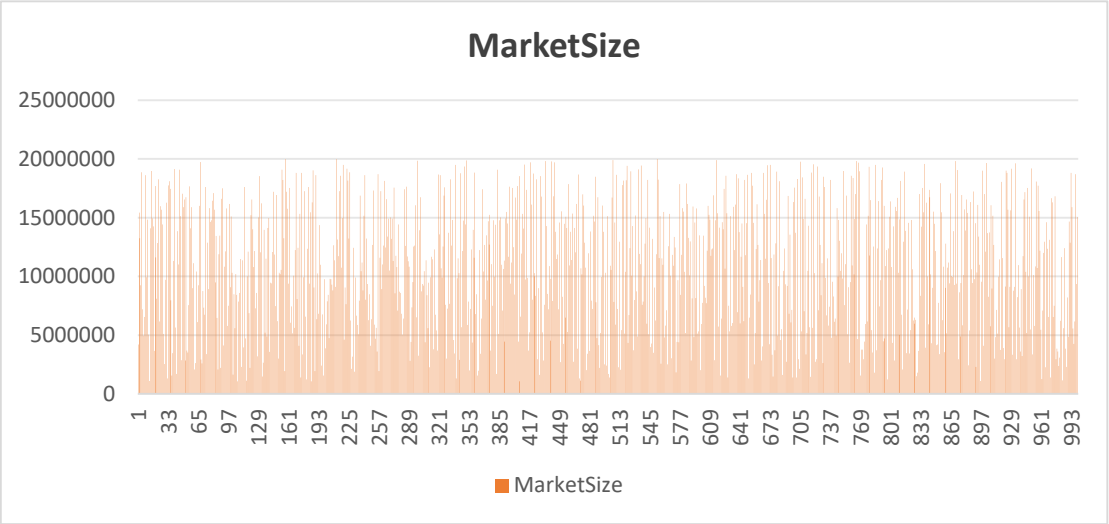

Fig. 11: Bar Chart of Market Size Distributions

Market size represents the total potential revenue available within a specific market and is a crucial feature in this research for classifying startup risks. A larger market size indicates greater opportunities for a startup to capture significant market share, scale operations, and achieve substantial growth, thereby potentially reducing market risks. Conversely, a smaller market size might limit growth potential and increase competition, heightening the risk for investors. By including market size in the regression model, we can more accurately predict and classify startup risks, offering investors valuable insights into the market potential and scalability of different startups. Understanding market size helps investors gauge the startup's potential for revenue generation and market penetration, making it an essential factor in risk assessment and strategic decision-making.

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 4, Issue 2, 2024 (pp. 81-102)

www.abjournals.org

DISCUSSION

In this research, we aimed to classify startup risks to enhance high investment returns using Random Forest Regression. Startups present a unique set of challenges and opportunities, and accurately assessing their risk profiles can significantly impact investment outcomes. Our study utilized a comprehensive dataset containing various features that are critical to a startup's success or failure, including funding levels, market size, expenses, team experience, product development stage, customer satisfaction scores, and revenue streams. We employed the Random Forest Regression model due to its robustness and ability to handle complex, non-linear relationships among features. This methodology allowed us to effectively capture the patterns in the data and make accurate predictions about startup risks.

The performance metrics of our model were highly encouraging, with a Mean Squared Error (MSE) of 0.255 and an R-squared value of 0.9515, indicating that the model explains approximately 95.15% of the variance in startup risks. Additionally, the Mean Absolute Error (MAE) was 0.782, the Mean Squared Logarithmic Error (MSLE) was 0.219, and the Explained Variance Score was 0.915, further confirming the model's accuracy and reliability. The analysis revealed that expenses and funding levels were the most significant factors influencing startup risk classification, underscoring the importance of financial health and resource allocation in determining a startup's potential for success. Our findings showed that 48.2% of the startups in the dataset exhibited no significant risks, while the rest were distributed across strategic, financial, operational, and market risks. These insights provide a valuable framework for investors to identify and mitigate potential risks, enabling them to make more informed and strategic investment decisions. In all, our research demonstrates the efficacy of using Random Forest Regression for risk classification in startups and offers practical recommendations for investors seeking to maximize their returns.

IMPLICATION TO RESEARCH AND PRACTICE

The implications of this research are significant for investors, startup founders, and financial analysts. By utilizing Random Forest Regression to classify startup risks, the study provides a powerful and reliable tool for predicting potential challenges that startups may face. This allows investors to make more informed decisions, optimize their resource allocation, and strategically invest in startups with the highest potential for high returns. For startup founders, understanding the key factors that influence their risk profile, such as funding levels and expenses, can help them better prepare and mitigate potential risks. Financial analysts can leverage the insights from this model to advise their clients more effectively, ensuring that investment portfolios are balanced and targeted towards ventures with favorable risk-return profiles. Ultimately, this research enhances the precision of risk assessment in the dynamic startup ecosystem, promoting more strategic and data-driven investment practices.

CONCLUSION

Our research focused on using Random Forest Regression to classify startup risks for high investment returns. The model's performance metrics indicate robust predictive capabilities, with a Mean Squared Error (MSE) of 0.255, an R-squared value of 0.9515, and an Explained

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

Variance Score of 0.915, suggesting that the model explains a significant portion of the variance and predicts startup risks accurately. The Mean Absolute Error (MAE) of 0.782 and the Mean Squared Logarithmic Error (MSLE) of 0.219 further support the model's reliability. The analysis revealed that 12.4% of startups faced Strategic Risks, 12.6% Financial Risks, 13.1% Operational Risks, 13.7% Market Risks, while 48.2% had no significant risks. Key features influencing the risk classification included Expenses, Funding Levels, Market Size, Customer Satisfaction Score, Team Experience, Product Development State, and Revenue Streams, with Expenses and Funding Levels being the most critical. Figures depicting actual versus predicted values and residuals versus predicted values demonstrated a tight clustering along the 45-degree line and random residual distribution, respectively, confirming the model's accuracy and lack of significant biases. Feature importance analysis highlighted crucial factors such as funding levels and market size, which significantly impact the risk profile of startups.Our Random Forest Regression model is effective in classifying startup risks, providing investors with a reliable tool to identify high-potential startups and make informed investment decisions. The insights gained from feature importance can guide strategic resource allocation and improve risk assessment accuracy, supporting better investment outcomes.

FUTURE RESEARCH

Future research should focus on integrating additional data sources, such as real-time market trends and social media sentiment analysis, to enhance the model's predictive accuracy and adaptability. It should also explore the application of advanced machine learning techniques, like deep learning, to capture more complex patterns and relationships in the data. Additionally, investigating the impact of macroeconomic factors on startup risks and returns can provide a more comprehensive risk assessment framework.

REFERENCES

Anietie Ekong, Abasiama Silas, Saviour Inyang (2022). A Machine Learning Approach for Prediction of Students' Admissibility for Post-Secondary Education using Artificial Neural Network. Int. J. Comput. Appl, 184, 44-49.44-49

Anietie Ekong, Blessing Ekong and Anthony Edet (2022), Supervised Machine Learning Model for EffectiveClassification of Patients with Covid-19 Symptoms Based on Bayesian Belief Network, Researchers Journal ofScience and Technology(2022),2, pp-27-33.

Anthony Edet, Uduakobong Udonna, Immaculata Attih, and Anietie Uwah (2024). Security Framework for Detection of Denial of Service (DoS) Attack on Virtual Private Networks for Efficient Data Transmission. Research Journal of Pure Science and Technology, 7(1),71-81. DOI: 10.56201/rjpst.v7.no1.2024.pg71.81

Ebong, O., Edet, A., Uwah, A., & Udoetor, N. (2024). Comprehensive Impact Assessment of Intrusion Detection and Mitigation Strategies Using Support Vector Machine Classification. Research Journal of Pure Science and Technology, 7,(2), 50-69.

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1

Edet, A. E. and Ansa, G. O. (2023). Machine learning enabled system for intelligent classification of host-based intrusion severity. Global Journal of Engineering and Technology Advances,16(03), 041–050.

Edet, A., Ekong, B. and Attih, I. (2024). Machine Learning Enabled System for Health Impact Assessment of Soft Drink Consumption Using Ensemble Learning Technique. International Journal Of Computer Science And Mathematical Theory,10(1):79-101, DOI: 10.56201/ijcsmt.v10.no1.2024.pg79.101

Ekong, A., James, G., Ekpe, G., Edet, A., & Dominic, E. A (2024). Model For The Classification Of Bladder State Based On Bayesian Network. International Journal of Engineering and Artificial Intelligence, 5 ,(2) 33–47

Ekong, B., Edet, A., Udonna, U., Uwah, A.,and Udoetor, N. (2024), Machine Learning Model for Adverse Drug Reaction Detection Based on Naive Bayes and XGBoost Algorithm. British Journal of Computer, Networking and Information Technology 7(2), 97-114.DOI: 10.52589/BJCNIT-35MFFBC6

Ekong, B., Ekong, O., Silas, A., Edet, A., & William, B. (2023). Machine Learning Approach for Classification of Sickle Cell Anemia in Teenagers Based on Bayesian Network. Journal of Information Systems and Informatics, 5(4), 1793-1808. https://doi.org/10.51519/journalisi.v5i4.629.

Inah Omoronyia, Ubong Etuk, Peter Inglis (2019). A privacy awareness system for software design. International Journal of Software Engineering and Knowledge Engineering, 29, (10), 1557-1604.

Mbang, u. b., Effiom, e. i., Ebri, w., Anzor, e. c., & Ekaetor, e. (2023). Entrepreneurial Education and Professional Exclusivity of Nigerian University Students in Cross River State.International Journal of Innovative Science and Research Technology,8,(2),688-693.

S. Inyang and I. Umoren (2023). From Text to Insights: NLP-Driven Classification of Infectious Diseases Based on Ecological Risk Factors. Journal of Innovation Information Technology and Application (JINITA), vol. 5, no. 2, pp. 154-165,

S. Inyang and I. Umoren (2023). Semantic-Based Natural Language Processing for Classification of Infectious Diseases Based on Ecological Factors. International Journal of Innovative Research in Sciences and Engineering

Uwah, A. and Edet, A. (2024).Customized Web Application for Addressing Language ModelMisalignment through Reinforcement Learning from HumanFeedback. World Journal of Innovation And Modern Technology,8,(1), 62-71. DOI: 10.56201/wjimt.v8.no1.2024.pg62.71.

Article DOI: 10.52589/AJSTE-UHDGSWQ1
DOI URL: https://doi.org/10.52589/AJSTE-UHDGSWQ1