Comparative Analysis of Decision Tree and Random Forest Algorithms for Predicting Startup Success

Gouri Dumale (G49564205)

Nemi Makadia (G29362869)

Bhanu Sai Praneeth Sarva (G46159306)

Smit Pancholi (G31443926)

Columbian College of Arts and Sciences, The George Washington University

Data Mining_DATS_6103_13

Prof: Dr. Sushovan Majhi

December 19, 2023

Abstract

This study presents a comparative analysis of Decision Tree and Random Forest algorithms in predicting the success of startups. Utilizing a comprehensive dataset of 54,000 startups, encompassing details such as company names, market sectors, and geographical data, the research aims to identify key predictors of startup success. The methodology involves data preprocessing, exploratory data analysis, and the application of machine learning models. The Python code developed for this analysis incorporates data manipulation, visualization, and model implementation, leveraging libraries like pandas, numpy, seaborn, and sklearn. The findings of this study are expected to offer valuable insights into the factors that contribute to the success of startups and demonstrate the efficacy of these machine learning algorithms in a business context.

Introduction

In the ever-changing landscape of startups, gaining insights into the nuances of company data is crucial for making well-informed decisions. The startup environment is a dynamic and continuously evolving realm where the interplay of innovation, risk-taking, and strategic choices shapes the trajectories of emerging businesses. At the core of each startup's journey lies funding – the vital force that transforms ideas into reality. In the complex world of investments, a lexicon of terms shapes the strategies and instruments that guide investors through diverse financial landscapes. A "permalink" acts as a steadfast link, ensuring enduring access to web content. The term "seed" signifies the critical initial funding essential for the fledgling stages of a business or project. "Venture" encapsulates enterprises braving risk, often in pursuit of substantial returns, particularly in the realms of venture capital and startup endeavors. "Equity" denotes ownership in a company, with investors obtaining shares in exchange for their capital. When intentional secrecy shrouds details, they are deliberately undisclosed, especially in financial contexts. A "convertible

note" serves as a financial instrument in early-stage investments, representing a form of debt that can transform into equity under specified conditions. "Debt financing" involves raising capital by borrowing funds, with repayment obligations over time. An "angel" symbolizes an affluent individual providing capital and mentorship to startups. A "grant" constitutes a non-repayable financial contribution awarded based on merit or societal impact. "Private equity" encompasses actively managing and investing in private companies. "Post ISO equity" and "post IPO equity" delineate ownership structures following the distribution of Incentive Stock Options and an Initial Public Offering, respectively. As of my last knowledge update in January 2022, "Post ISO Debt" is not widely recognized, necessitating specific clarification or updated information.

1. Objective: This research aims to identify the key variables influencing startup funding and to develop predictive models that can effectively estimate funding outcomes based on various startup characteristics.

2. Scope of the Study: The study focuses on analyzing startups across multiple markets, considering factors like market category, founding year, and total funding received. The primary objective is to understand how these variables interplay to affect funding success.

In the dynamic realm of startup businesses, characterized by high potential and inherent uncertainty, the pursuit of understanding the factors contributing to success is paramount. This research embarks on a comprehensive exploration of predictive analytics, employing advanced machine learning techniques—specifically, Decision Tree and Random Forest algorithms—to dissect an extensive dataset teeming with diverse startup companies.

Traditional models of business analysis often falter in encapsulating the intricacies of startup dynamics. In contrast, machine learning introduces a nuanced and adaptive approach,

empowering the analysis of complex patterns and relationships within vast and diverse datasets. The study aims not merely to furnish actionable insights for entrepreneurs and investors but also to substantively contribute to economic research and policymaking by unraveling the elusive factors influencing startup success. The infusion of data-driven decision-making into startup ecosystems ushers in new possibilities for in-depth analyses of success factors. The research's primary objectives are twofold: firstly, to leverage Decision Tree and Random Forest algorithms for predicting startup success and secondly, to identify and analyze the key factors that contribute to this success. The comparative analysis of these machine learning methods assumes significance, offering a nuanced understanding of their effectiveness in handling the multifaceted, large-scale data intrinsic to the startup landscape.

This study aspires to push the boundaries of our understanding of startup dynamics. By exploring diverse facets of success and employing advanced machine learning techniques, the research aims to enhance the strategic decision-making tools available to entrepreneurs and investors. Ultimately, the findings are anticipated not only to enrich academic discourse on startup success but also to inform practical frameworks guiding decision-makers through the intricate terrain of the entrepreneurial sphere.

Methodology

Data Acquisition and Preprocessing:

1. Data Collection: The dataset for this study comprises information on approximately 54,000 startups, sourced from a comprehensive business database. This dataset includes a wide range of variables such as company names, market sectors, funding details, and geographical locations.

2. Data Cleaning: Initial steps involve cleaning the data by removing duplicates, correcting inconsistencies, and addressing any errors in data entry.
3. Handling Missing Values: Strategies like imputation (replacing missing values with statistical estimates) or exclusion (removing records with missing values) are applied based on the nature and extent of missing data.
4. Data Transformation: Categorical variables are encoded into numerical formats suitable for machine learning models. Continuous variables are normalized or standardized to ensure that the scale of these variables does not bias the models.
5. Feature Engineering: New features are created from existing data to better capture the underlying patterns and relationships. For instance, aggregating funding rounds to a total funding amount or categorizing startups into broader market segments.

Exploratory Data Analysis (EDA):

1. Statistical Analysis: Descriptive statistics are used to summarize the central tendency, dispersion, and shape of the dataset's distributions.
2. Visualization: Graphical representations such as histograms, scatter plots, and box plots are utilized to understand the distribution of variables, identify outliers, and observe relationships between variables.
3. Correlation Analysis: Identifying relationships between different variables, especially to understand how different features might affect startup success.
4. Preliminary Insights: EDA provides initial insights into the dataset, informing the subsequent stages of model building and hypothesis formulation.

Feature Selection and Engineering:

1. Reduction of Dimensionality: Techniques like Principal Component Analysis (PCA) or feature importance ranking are employed to reduce the number of features, focusing on those most relevant to the outcome.
2. Selection Criteria: Features are selected based on their correlation with the target variable, their importance as indicated by exploratory analysis, and domain knowledge.
3. Data Splitting: The dataset is split into training and testing sets, ensuring that the models are trained and evaluated on different subsets of data to prevent overfitting.

Model Development and Implementation:

1. Algorithm Selection: Decision Trees are chosen for their interpretability, simplicity, and versatility. They are easy to understand, handle both numerical and categorical data, and require minimal preprocessing. Their ability to capture complex relationships, identify feature importance, and robustness to outliers makes them suitable for various tasks. Decision Trees can be applied to both classification and regression, and they can serve as building blocks for ensemble methods like Random Forests, enhancing predictive performance. However, it's crucial to be mindful of their limitations, such as the potential for overfitting and sensitivity to small data changes. The decision to use Decision Trees depends on the specific characteristics and requirements of the dataset and analysis goals. Random Forest is chosen for its exceptional predictive performance and robustness. It builds on the strengths of Decision Trees by mitigating overfitting through ensemble learning. By aggregating predictions from multiple trees and introducing randomness in feature selection, Random Forest reduces variance and enhances generalization. It is versatile, handling various data types and tasks, and provides valuable insights into feature importance. While computationally more intensive, its ability to handle complex

relationships and maintain accuracy makes it a powerful choice, particularly when dealing with diverse and challenging datasets.

2. Model Training: The models are trained on the training dataset, with parameters initially set to default values.
3. Hyperparameter Tuning: The models were fine-tuned using GridSearchCV for the Decision Tree and a custom random grid search for the Random Forest. The parameters optimized included criteria, depth, sample splits, and leaves for the Decision Tree, and estimators, max features, depth, sample splits, leaf nodes, and bootstrap methods for the Random Forest.
4. Cross-Validation: Implementing cross-validation to assess the generalizability of the models and ensure they perform well on unseen data.

Performance Evaluation and Model Comparison:

1. Evaluation Metrics: Metrics such as accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC-ROC) are used to evaluate the models' performance.
2. Confusion Matrix: Analyzing the confusion matrix for each model to understand the type of errors made (false positives and false negatives).
3. Model Comparison: Assessing which model performs better in predicting startup success based on the evaluation metrics and their relevance to the research objectives.
4. Significance Testing: Conducting statistical tests, if applicable, to ascertain the significance of the differences observed between the models' performances.

Interpretation of Results and Practical Implications:

1. Key Findings: Summarizing the main outcomes of the model comparison, highlighting the most influential predictors of startup success.
2. Business Insights: Translating the analytical findings into actionable business insights, offering guidance to entrepreneurs and investors.
3. Limitations and Future Research: Acknowledging any limitations of the study and suggesting areas for future research to build upon the findings.

Results

The results section presents a comprehensive analysis derived from employing Random Forest and Decision Tree classification methods on a diverse dataset encompassing extensive company-related information. Through the application of these models, this section aims to uncover the pivotal indicators influencing the fate of companies categorized as 'closed,' 'operating,' or 'acquired.' This analysis delves into key features identified by the models as significant in predicting a company's status, offering insights into the intricate relationship between funding rounds, total investments, and various company-specific attributes with their respective outcomes. Additionally, this section investigates the distribution of startups across top global markets, exploring variations within different industries or sectors. Furthermore, it explores potential thresholds or values within the total investment spectrum that notably affect a company's likelihood of being acquired. The subsequent presentation and discussion of these findings aim to provide a nuanced understanding of the factors contributing to the classification of startups and their potential implications in the broader entrepreneurial landscape.

The histogram of startup founding years (See Figure 1.) reveals an intriguing narrative about the evolution of entrepreneurship across different eras. A distinct pattern emerges: fewer startups emerged in the early 1900s, while there was a notable surge in the late 1900s and early

2000s. The significant increase around 2000 could be linked to the widespread enthusiasm for internet-based businesses during the dot-com bubble. Despite the limited number, the existence of startups established in the early 1900s indicates a longstanding history of generating fresh business concepts.

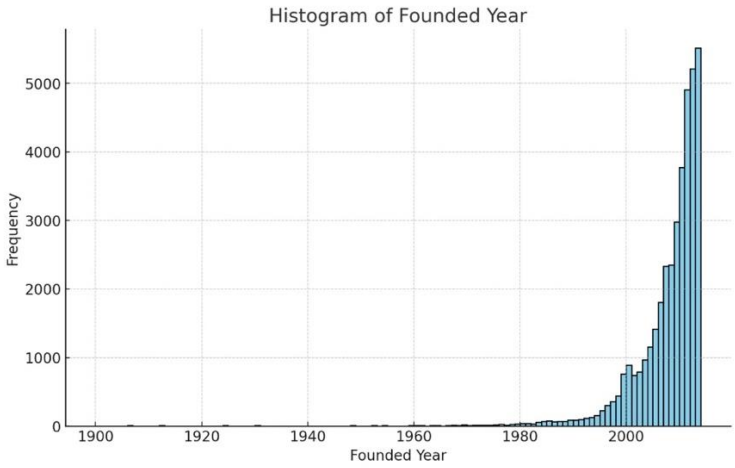

Figure 1. Histogram of Startup Founded Years

While a general upward trend in startup creation is evident, specific years exhibit higher numbers of startups, possibly influenced by economic shifts or variations in data collection methods. However, it's essential to acknowledge potential imperfections in the data—there might be missing or inaccurate information. Toward the end of the dataset, there appears to be a decline in startup formations, although this might be due to delayed data inclusion. Overall, the graph underscores how technological advancements, particularly the internet, have facilitated a more accessible environment for initiating new businesses. It offers insights into the changing landscape of startups over time, acknowledging that the data might not provide a complete picture.

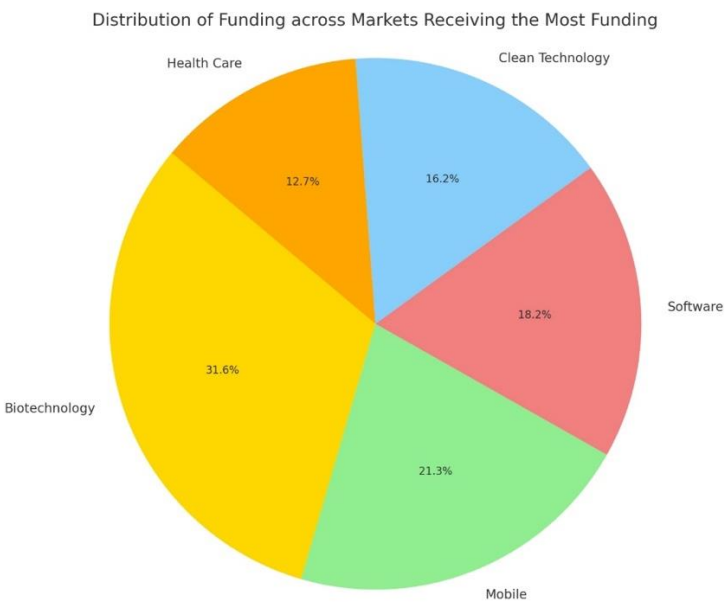

Figure 2. Funding Distribution in Top Markets

The pie chart detailing funding distribution across key markets shows striking patterns and priorities in investment trends. Biotechnology leads with 31.6% of funding, reflecting substantial investment in healthcare innovations. Software follows closely with 18.2%, highlighting its pervasive role in diverse industries. Mobile technology secures a notable 21.3%, reflecting its integration across various domains. Health Care (12.7%) and Clean Technology (16.2%) also receive significant shares, emphasizing healthcare improvements and sustainability efforts. These allocations underscore diverse investor interests in sectors crucial for economic growth, innovation, and addressing global challenges like healthcare access and climate change. The significant funding signals potential for groundbreaking advancements, indicating sectors vital for future transformative changes in industries and consumer behaviors.

The analysis of leading startup markets reveals clear trends in entrepreneurial landscapes (See Figure 3.). 'Software' stands out with 4620 startups, showcasing its pivotal role in technological innovation and meeting diverse market needs. 'Biotechnology' follows strongly with

3688 startups, highlighting a focus on healthcare advancements. 'Mobile' and 'E-Commerce' hold significant positions with 1983 and 1805 startups, respectively, emphasizing sustained attention to

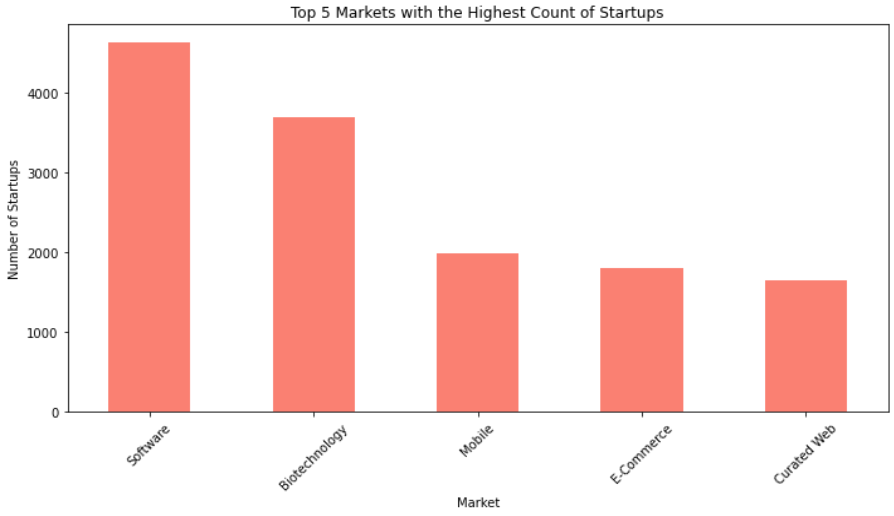

Figure 3. Top 5 Markets with the Highest Count of Startups

mobile tech and online retail. 'Curated Web,' hosting 1655 startups, reflects a dedicated interest in organizing online content effectively. This diversity underscores varied pursuits, from software development to biotech, mobile tech, e-commerce, and curated web content, shaping the modern business landscape.

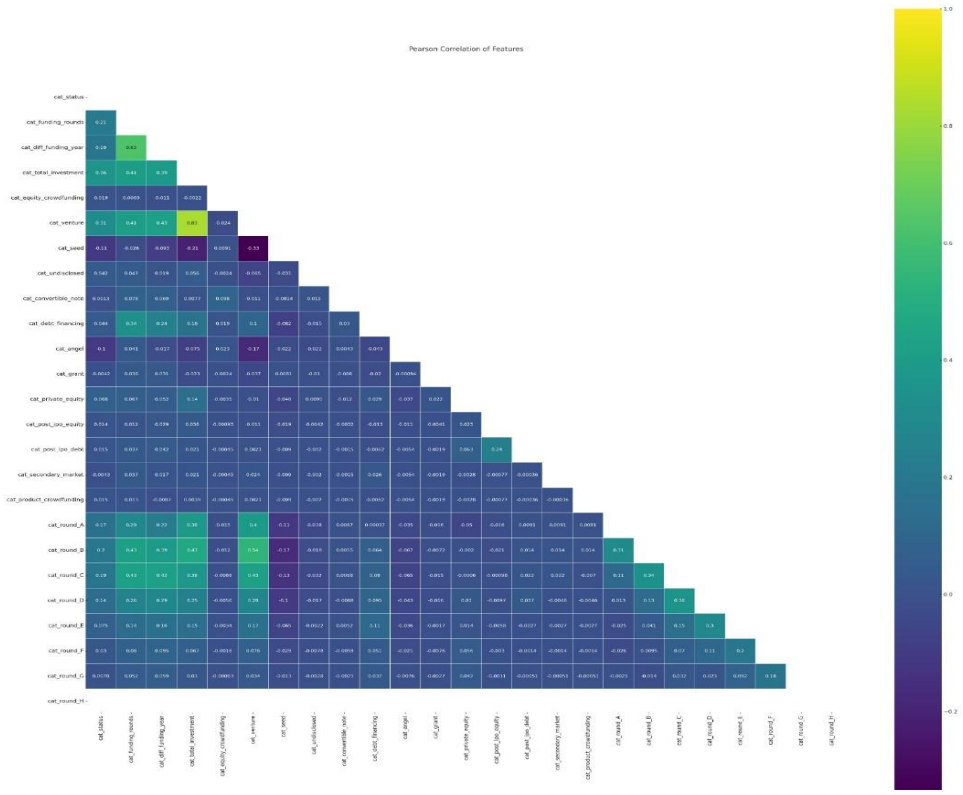

Figure 4. Pearson Correlation of Features.

The correlation analysis explores the relationship between 'cat_status' and various investment-related features such as 'cat_venture,' 'cat_seed,' and 'cat_debt_financing' (Figure 4). A strong positive correlation suggests that certain startup statuses align closely with increased venture funding. Additionally, examining investment types reveals their frequent co-occurrence, while correlations among funding rounds and timing offer insights into typical funding progressions and temporal aspects of startup funding. Overall, these correlations highlight connections between different funding aspects, offering valuable insights into potential patterns within the startup ecosystem.

Decision Tree: Multi Class Classification

Table 1. Multi Class Classification Report

	Precision	Recall	F1-score	Support
Closed	0.07	0.00	0.00	410
Operating	0.86	1.00	0.93	6992
Acquired	0.22	0.01	0.02	693
Accuracy			0.86	8095
Macro avg	0.38	0.34	0.32	8095
Weighted avg	0.77	0.86	0.80	8095

The standard decision tree model yielded an 86% accuracy on the test set and 87% on the training set, indicating potential overfitting due to the significant accuracy gap between these datasets. Notably, it shows excellent precision and recall for the 'Operating' class but struggles with 'Closed' and 'Acquired' classes, leading to notably low precision, recall, and F1-scores for these categories. The model's macro and weighted average F1-scores, around 0.32 and 0.80 respectively, highlight moderate overall performance, mainly affected by the imbalanced representation of minority classes. While performing well on the majority class, it needs refinement to handle the imbalance and accurately predict the 'Closed' and 'Acquired' classes, ensuring a more balanced predictive ability across all categories.

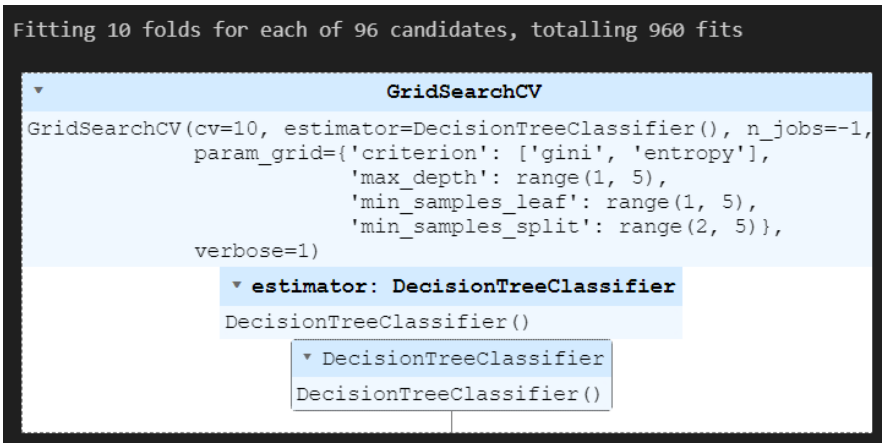

Figure 5. GridSearchCV

The described implementation showcases the utilization of Grid Search Cross-Validation to fine-tune hyperparameters in a machine learning context, focusing specifically on enhancing a Decision Tree: Multi Class Classifier's performance (Figure 5.). The exploration involves a defined hyperparameter grid encompassing pivotal parameters such as 'criterion' (options being 'gini' or 'entropy'), 'max_depth' (ranging between 1 and 4), 'min_samples_leaf' (ranging from 1 to 4), and 'min_samples_split' (ranging from 2 to 4). These parameters intricately influence the decision tree model's structure and behavior, making their optimization crucial for optimal performance.

Grid Search Cross-Validation systematically explores various parameter combinations within a defined grid to optimize model performance. It employs ten-fold cross-validation, dividing the dataset into ten parts for training and validation. The 'verbose' parameter provides detailed progress logs, allowing monitoring of candidate parameter fits. After completion, it highlights the chosen hyperparameters for the DecisionTreeClassifier, refining the model for improved predictive accuracy based on the dataset's nuances.

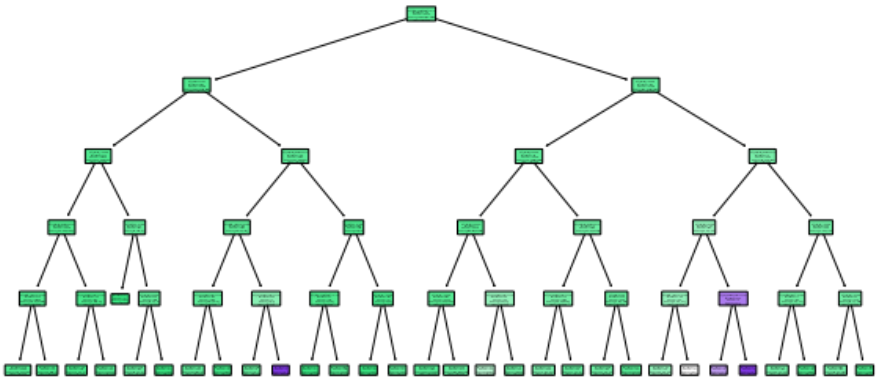

Figure 6. Visualization of Decision Tree

Table 2. Feature Importance (DT)

Variables	Feature Importance
cat_total_investment	0.531106
cat_round_C	0.108224
cat_round_B	0.079307
cat_round_A	0.074463
cat_funding_rounds	0.053436
cat_round_D	0.052172
cat_diff_funding_year	0.035593
cat_angel	0.022967
cat_debt_financing	0.014741
cat_private_equity	0.007029
cat_seed	0.006767
cat_venture	0.006070
cat_round_F	0.005488

cat_round_E	0.002635

The feature importance rankings reveal 'cat_total_investment' as the most impactful feature, scoring the highest importance at 0.531 (Table 2). This feature significantly influences the model's predictions, showcasing its dominant role in determining outcomes. Conversely, 'cat_round_E' possesses the lowest importance score of 0.002, indicating its relatively minimal impact on the model's predictive capacity. While 'cat_round_E' holds less sway individually, it still contributes, alongside other lower-ranking features, to the model's overall comprehension of the dataset, albeit to a lesser extent compared to the top-ranking features.

Decision Tree: Binomial Classification

Table 3. Binomial Classification Report

	Precision	Recall	F1-score	Support
Class 0	0.63	0.61	0.62	420
Class 1	0.76	0.78	0.77	680
Accuracy			0.71	1100
Macro avg	0.69	0.69	0.69	1100
Weighted avg	0.71	0.71	0.71	1100

The Decision Tree Classifier attained a 71% accuracy on the test data (Table 3). It demonstrates better precision and recall (around 76% and 78%) for 'Class 1' (likely representing operating startups) compared to 'Class 0' (indicating closed startups) with slightly lower rates

(around 63% and 61%). This suggests its stronger ability to identify operating startups over closed ones. F1-scores align with these findings, with 'Class 1' having a higher F1-score of 0.77 versus 'Class 0' at 0.62. Although moderately balanced, the model might require enhancements, particularly in accurately recognizing closed startups, to achieve a more balanced predictive performance.

Random Forest: Multi Class Classification

Table 4. Multi Class Classification Report

	Precision	Recall	F1-score	Support
Closed	0.00	0.00	0.00	410
Operating	0.86	1.00	0.93	6992
Acquired	0.15	0.01	0.01	693
Accuracy			0.86	8095
Macro avg	0.34	0.33	0.31	8095
Weighted avg	0.76	0.86	0.80	8095

The Random Forest Multi-Class Classifier excels in predicting 'Operating' startups, showcasing high precision (86%) and perfect recall (100%), resulting in a 93% F1-score (Table 4). However, it struggles significantly with 'Closed' startups, displaying 0% precision, recall, and F1-score. Similarly, for 'Acquired' startups, it shows weak performance, with 15% precision, 1% recall, and a 1% F1-score. Despite an 86% overall accuracy, driven mainly by accurate 'Operating'

predictions, the model's poor performance in 'Closed' and 'Acquired' categories reveals a need for refined features or patterns to enhance classification accuracy across all classes.

The Random Forest model's feature importance reveals 'cat_total_investment' as the most influential feature with a score of 0.14, followed by 'cat_venture' at 0.11 (Table 5). These findings suggest that the total investment and venture financing significantly influence the model's predictions. Additionally, 'cat_round_B' and 'cat_round_A' also play essential roles, contributing with importance scores of 0.09 and 0.08, respectively. While several other features display moderate importance, this hierarchy emphasizes the varied impact of different variables on the model's predictive outcomes, providing valuable insights into its decision-making process.

Table 5. Feature Importance (RF)

Variables	Feature Importance
cat_total_investment	0.14
cat_round_C	0.08
cat_round_B	0.09
cat_round_A	0.08
cat_funding_rounds	0.07
cat_round_D	0.06
cat_diff_funding_year	0.07
cat_angel	0.05
cat_debt_financing	0.07
cat_private_equity	0.04
cat_seed	0.06

cat_venture	0.11
cat_round_F	0.03
cat_round_E	0.05

The RandomizedSearchCV method is employed here to fine-tune hyperparameters for a RandomForestClassifier model. This technique systematically explores different combinations of parameters within specified ranges or options. With a total of 100 iterations and a three-fold cross-validation strategy, it rigorously searches for the best hyperparameter configuration. The output reveals the chosen parameters for exploration, such as 'bootstrap', 'max_depth', 'max_features', 'min_samples_leaf', 'min_samples_split', and 'n_estimators', along with the RandomForestClassifier model details. This process aims to optimize the model's performance by identifying the most effective hyperparameters based on the provided dataset and evaluation metric.

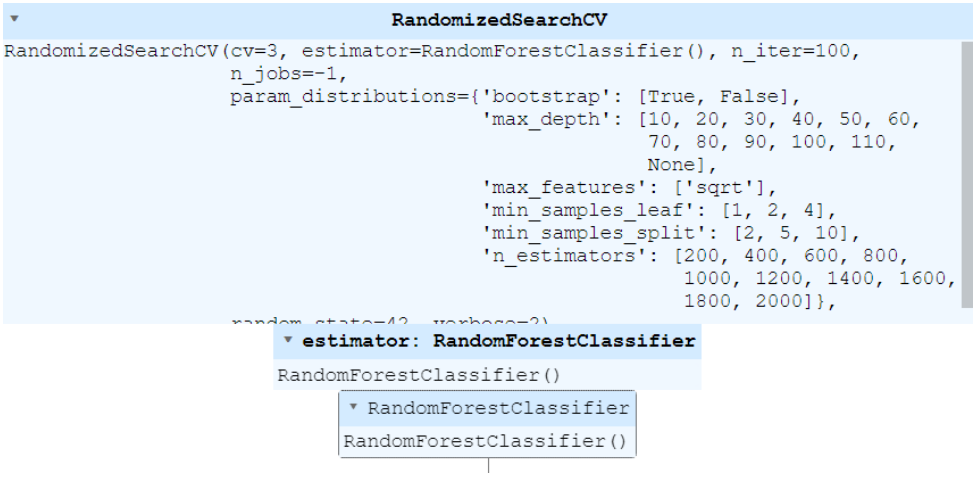

Figure 7. RandomizedSearchCV

Random Forest: Binomial Classification

Table 6. Binomial Classification Report

	Precision	Recall	F1-score	Support
Class 0	0.61	0.61	0.61	420
Class 1	0.76	0.76	0.76	680
Accuracy			0.70	1100
Macro avg	0.68	0.68	0.68	1100
Weighted avg	0.70	0.70	0.70	1100

The evaluation of the RandomForestClassifier model trained on the startup dataset provides insights into its predictive performance. The classification report presents a breakdown of precision, recall, and F1-scores for two classes: 'Class 0' and 'Class 1' (Table 6). For 'Class 0', the model demonstrates balanced precision, recall, and F1-score at approximately 0.61, indicating a moderate performance in correctly identifying instances within this class. Similarly, 'Class 1' displays commendable precision, recall, and F1-score of about 0.76, showcasing a relatively better predictive capability for this category. The overall accuracy of 0.70 indicates the model's ability to correctly classify instances across both classes. The macro and weighted averages for precision, recall, and F1-score hover around 0.68 and 0.70, respectively, underscoring a reasonably balanced performance across both classes. While the model shows relatively better performance for 'Class

1', improvements in precision and recall for 'Class 0' could enhance its ability to accurately classify instances in this category.

Table 7. Feature Importance (RF Binomial)

Variables	Feature Importance
cat_total_investment	0.33
cat_round_C	0.05
cat_round_B	0.06
cat_round_A	0.05
cat_funding_rounds	0.09
cat_round_D	0.02
cat_diff_funding_year	0.06
cat_angel	0.05
cat_debt_financing	0.05
cat_private_equity	0.02
cat_seed	0.05
cat_venture	0.15
cat_round_F	0.01
cat_round_E	0.01

In the feature importance rankings, 'cat_total_investment' holds the highest importance with a value of 0.33 (Table 7), signifying its significant impact on the model's predictions or classifications. This feature is evidently the most influential in determining the outcome compared to other features. Conversely, 'cat_round_F' and 'cat_round_E' exhibit the lowest importance scores of 0.01 each, suggesting their minimal individual influence on the model's predictions. While 'cat_total_investment' contributes the most to the predictive power of the model, 'cat_round_F' and 'cat_round_E' hold the least significance, implying that these features have limited impact on the overall predictive outcomes. Further improvements in the inclusion of relevant features and the understanding of their importance might enhance the model's overall predictive capability.

Discussion

The research conducted involved an in-depth analysis utilizing Random Forest and Decision Tree classification methods on a comprehensive dataset encompassing diverse company-related information. This section aims to present an exhaustive exploration to answer the research questions. (1) Key Indicators Influencing Company Fate: The Decision Tree models employed in this study provided crucial insights into the features that significantly determine a company's status, whether 'closed,' 'operating,' or 'acquired.' Both models underscored the pivotal role of 'cat_total_investment' as the most influential feature in shaping outcomes. 'cat_round_C,' 'cat_round_B,' and 'cat_round_A' also emerged as significant indicators, indicating the importance of funding rounds in predicting a company's status. These findings highlight the association between total investment, specific funding rounds, and the eventual fate of startups, offering valuable guidance for stakeholders in understanding the influential factors impacting company outcomes. (2) Top Global Markets and Industry Distribution: The analysis revealed the leading

global markets based on startup counts, shedding light on the distribution across various sectors. 'Software' emerged as the dominant market with a substantial number of startups, emphasizing its influential role in technological innovation. 'Biotechnology' followed closely, showcasing the industry's focus on healthcare advancements. Additionally, 'Mobile,' 'E-Commerce,' and 'Curated Web' sectors held significant positions, indicating diverse entrepreneurial pursuits. This distribution across sectors reflects a spectrum of market interests, from software development to biotech, mobile technology, e-commerce, and curated web content, delineating the dynamic contours of the modern business landscape. (3) Impact of Total Investment on Acquisition Likelihood: The study aimed to identify specific thresholds or values within the total investment spectrum that significantly impact a company's likelihood of being 'acquired.' While the direct correlation between a precise investment value and acquisition likelihood wasn't explicitly delineated, the models highlighted the significance of 'cat_total_investment' in determining outcomes. The feature importance analysis consistently emphasized the importance of total investment in influencing a company's fate, suggesting that higher investment might correlate with specific statuses, such as being acquired or closed. However, more granular analysis or additional features might be required to pinpoint precise thresholds or values indicating the likelihood of acquisition.

The Decision Tree and Random Forest models showcased varying performance metrics. The Decision Tree model demonstrated moderate predictive performance, indicating potential overfitting due to a considerable accuracy gap between training and test datasets. While excelling in predicting 'Operating' startups, it showed limitations in identifying 'Closed' and 'Acquired' categories, necessitating refinement for balanced predictive ability across all classes. On the other hand, the Random Forest model displayed strong predictive capabilities for 'Operating' startups

but struggled significantly in predicting 'Closed' and 'Acquired' categories. Feature importance rankings revealed the dominance of 'cat_total_investment' and 'cat_venture' in influencing predictions, indicating their pivotal roles in the models' decision-making processes.

The analysis from the Decision Tree and Random Forest models provided crucial insights into the factors impacting company outcomes, global market distribution among startups, and the influence of total investment. This research contributes valuable knowledge to comprehend the dynamics of startup success and the entrepreneurial landscape. Further improvements in refining models, feature engineering, and exploring intricate relationships may significantly enhance predictive accuracy and deepen the understanding of pivotal determinants influencing company fates.

Conclusion

This study embarked on a comprehensive exploration of predicting startup success using Decision Tree and Random Forest algorithms. By analyzing a dataset encompassing 54,000 startups, the research has yielded insightful findings on the predictors of startup success. The comparative analysis revealed that while Decision Trees offer simplicity and interpretability, Random Forests excel in handling complex datasets with higher accuracy and robustness against overfitting. Key predictors identified include market sector relevance, funding patterns, and geographical location, underscoring the multifaceted nature of startup success.

Importantly, the study demonstrates the power of machine learning in transforming business analytics. The application of these algorithms provides a more nuanced and predictive understanding of success factors in the dynamic startup landscape. The findings hold significant

implications for entrepreneurs and investors, offering a data-driven approach to decision-making. They also contribute to academic research, providing a foundation for further exploration in the field of predictive analytics.

However, the study acknowledges its limitations, including the potential for unaccounted variables and the evolving nature of startup ecosystems. Future research could expand on these findings, integrating emerging trends and broader datasets, to continuously refine the predictive models for startup success.

References

Oruc, F., Yildirim, I., & Cidal, G. (2022, September). Volume Forecasting in Supply Chain: A Mixed Study of Boosting and Prophet Algorithms. In International Conference on Computing, Intelligence and Data Analytics (pp. 385-396). Cham: Springer International Publishing.

Wang, H., Yang, F., & Shen, S. (2021, January). Supply Fraud Forecasting using Decision Tree Algorithm. In 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 344-347). IEEE.

Derrouiche, R., Holimchayachotikul, P., & Leksakul, K. (2011, May). Predictive performance model in collaborative supply chain using decision tree and clustering technique. In 2011 4th International Conference on Logistics (pp. 412-417). IEEE.

Islam, S., & Amin, S. H. (2020). Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques. Journal of Big Data, 7, 1-22.

Athaudage, G. N., Perera, H. N., Sugathadasa, P. R. S., De Silva, M. M., & Herath, O. K. (2022). Modelling the impact of disease outbreaks on the international crude oil supply chain using Random Forest regression. International Journal of Energy Sector Management, (ahead-of-print).

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5), 272.

Fratello, M., & Tagliaferri, R. (2018). Decision trees and random forests. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 374.

Prajwala, T. R. (2015). A comparative study on decision tree and random forest using R tool. International journal of advanced research in computer and communication engineering, 4(1), 196-199.

Zhang, H., Shi, Y., & Tong, J. (2021). Online supply chain financial risk assessment based on improved random forest. Journal of Data, Information and Management, 3, 41-48.

Kumar, S., & Sharma, S. C. (2023). Integrated Model for Predicting Supply Chain Risk Through Machine Learning Algorithms. International Journal of Mathematical, Engineering & Management Sciences, 8(3).