Prediction of Startup Profitability Based on Logistic Regression and Random Forest

Abstract – Assessing startup profitability is crucial for investors and entrepreneurs. This study applies Logistic Regression and Random Forest to classify startups as profitable or unprofitable based on financial and market factors, including market share, funding rounds, startup age, revenue, funding amount and employee count. It explores data preprocessing, feature selection, and model evaluation, with key findings and limitations discussed in the conclusion.

Keywords: Startup Profitability, Machine Learning, Financial Indicators, Predictive Analytics

I. INTRODUCTION

Aim

The project aims to evaluate the effectiveness of machine learning models in predicting startup profitability based on key factors, including funding rounds, funding amount, revenue, valuation, employee count, and market share.

Research Question

How accurately can machine learning models predict startup profitability based on funding, market share, and other business factors?

Problem statement

Startup profitability refers to a startup's ability to generate sufficient revenue to cover operational costs and achieve profit. However, predicting profitability in the competitive and uncertain startup landscape remains challenging. While financial metrics such as funding, revenue, and market share serve as key indicators, traditional evaluations often incorporate subjective factors, including the entrepreneur's background, industry experience, and market intuition, These subjective assessments introduce bias and inconsistency (Zacharakis and Meyer, 1998)

Biases in investment decisions shape a startup's trajectory, often favouring charismatic founders over financially solid but unconventional ventures (Welter, Holcomb and McIlwraith, 2023). As a result, resource allocation becomes inconsistent, with some startups struggling despite strong fundamentals while others receive funding without sustainable growth prospects. These inefficiencies undermine investor confidence, disrupt funding cycles, and distort competition by prioritising perceived potential over actual performance (ibid.).

Hypothesis

Machine learning models can accurately predict startup profitability based on financial and market-related features.

Objectives

1. Identify the key variables influencing startup profitability predictions.
2. Develop and assess two machine learning models to classify startups as profitable or unprofitable.
3. Conduct Exploratory Data Analysis (EDA) to uncover patterns, correlations, and distributions.
4. Uilise Python libraries such as Pandas, Seaborn, and Matplotlib to visualise trends and present findings effectively.
5. Evaluate model performance using AUC as the primary metric, supplemented by accuracy, recall, and F1-score for a comprehensive analysis.
6. Analyse the practical implications for investors and entrepreneurs, proposing refinements to enhance predictive accuracy.

II. LITERATURE REVIEW

Startup Failure Rates and the Need for Predictive Models

The high failure rate of startups poses a major challenge for investors and entrepreneurs. Ünal (2019) reports that up to 90% of startups fail within their first year, with fewer than 40% of survivors lasting beyond five years. Böhm et al. (2017) found failure rates ranging from 50% to 83%, underscoring the difficulties in scaling and sustaining operations. Similarly, Giardino et al. (2015) observed that 60% of startups fail within five years, with 75% of funded ventures ultimately collapsing. These findings highlight the uncertainty of startup survival and the necessity of data-driven predictive models.

Machine Learning in Financial-Based Prediction

To address these challenges, researchers have increasingly adopted Machine Learning (ML) to assess startup success probability. These models leverage financial indicators such as funding amounts, funding rounds, and profitability to

predict a startup's growth potential (Cirjevskis, 2018; Martinez, 2019; Spiegel et al., 2015). By offering data-driven insights, ML models help investors make more informed decisions, reducing reliance on subjective judgment.

Common Machine Learning Models for Profitability Prediction

ML techniques such as Logistic Regression, Random Forest, and Support Vector Machines (SVM) have been widely applied to startup profitability prediction. Logistic Regression models achieve accuracy rates between 71% to 92%, depending on dataset characteristics (Bento, 2017; Martinez, 2019; Ünal, 2019). Random Forest models demonstrate strong predictive performance, with accuracy rates ranging from 83% to 94% (Bento, 2017; Ünal, 2019; Krishna, Agrawal and Choudhary, 2016). SVM models, often used for profitability analysis, achieve accuracy between 66% and 83% (Böhm et al., 2017).

Challenges in Predicting Startup Profitability with Machine Learning

Despite advancements in ML-based predictions, challenges persist. Issues such as data quality, model transparency, and the absence of a standardised definition of startup success hinder predictive accuracy (Żbikowski and Antosiuk, 2021). Addressing these concerns will be crucial for improving ML-driven startup evaluation models.

While previous research has explored startup success in broad terms, this study specifically examines profitability as a measurable financial outcome. By focusing on the ability of startups to generate profit, this research aims to provide a more precise evaluation of the financial factors influencing their viability.

III. DATA MANAGEMENT

External Libraries

NumPy: performs fast numerical calculations and optimises data processing efficiency

Pandas: used for data manipulation, cleaning and feature engineering

Matplotlib and Seaborn: visualise data distributions, correlations, and comparisons

Scikit-learn: trains ML models and handles feature scaling, train-test splitting and model evaluation

Data Source and Description

The "Startup Growth & Funding Trends" (Ashar, 2025) dataset from Kaggle contains 500 startups and 12 features, covering key financial and market-related metrics relevant to startup performance analysis.

	Startup Name	Industry	Funding Rounds	Funding Amount (M USD)	Valuation (M USD)	Revenue (M USD)	Employees	Market Share (%)	Profitable	Year Founded	Region	Exit Status
0	Startup_1	IoT	1	101.09	844.75	67.87	1460	5.20	0	2006	Europe	Private
1	Startup_2	EdTech	1	247.62	3310.83	75.65	3280	8.10	1	2003	South America	Private
2	Startup_3	EdTech	1	109.24	1056.37	84.21	4933	2.61	1	1995	South America	Private
3	Startup_4	Gaming	5	10.70	101.80	47.08	1030	3.13	0	2003	South America	Private
4	Startup_5	IoT	4	249.28	850.11	50.25	1905	4.09	0	1997	Europe	Acquired

Figure 1. A Sample of Data Extracted from the Original Dataset

Since Startup Name serves only as an identifier, it was excluded to remove redundancy and improve the dataset's analytical value. The remaining features include categorical, numerical, and binary variables, making the dataset suitable for classification (e.g., predicting profitability) and regression (e.g., estimating valuation or revenue.

Dealing with Missing and Duplicated Data

The original dataset contained no missing or duplicated values, eliminating the need for any data imputation or cleaning in this step.

Univariate Analysis

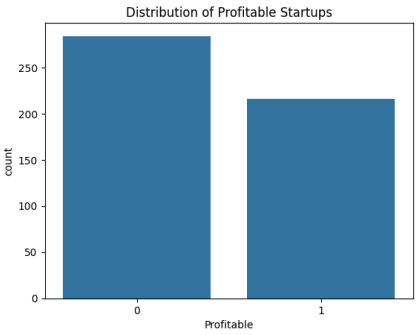

Figure 2. Distribution of Profitable Startups

EDA commenced with an assessment of the target variable, Profitability, using a countplot to visualise the distribution of profitable and unprofitable startups.

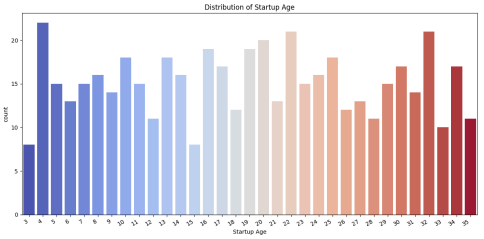

Figure 3. Distribution of Startup Age

Next, feature distributions were analysed. The discrete feature, Startup Age, was visualised using a countplot to observe their spread. The Startup Age distribution is uneven, with peaks at specific ages, suggesting that startups emerge in certain economic cycles, while those 25+ years old exhibit higher survival rates.

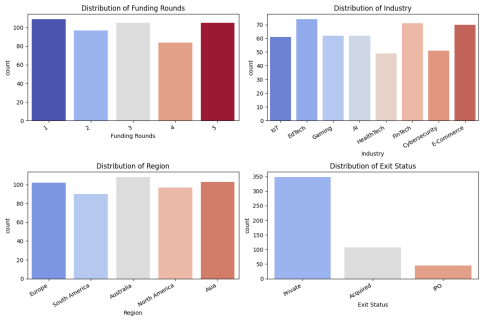

Figure 4. Distribution of Funding Rounds, Industry, Region and Exit Status

Countplots were used to analyse categorical features, including Funding Rounds, Industry, Region, and Exit Status, to identify variations across startup categories.

Funding Rounds (1-5) is treated as categorical. Early-stage funding dominates, mid-stage rounds are well-represented and only the most successful startups reach Round 5. Industry distribution shows EdTech, FinTech, and E-Commerce leading, while HealthTech and Cybersecurity are underrepresented, potentially affecting model generalisation. Regional distribution is Australia-heavy, followed by Europe and Asia, with South America least presented, posing a risk of regional bias. Exit Status is highly skewed, with most startups private, few acquired, and IPOs rare, requiring oversampling for better model balance.

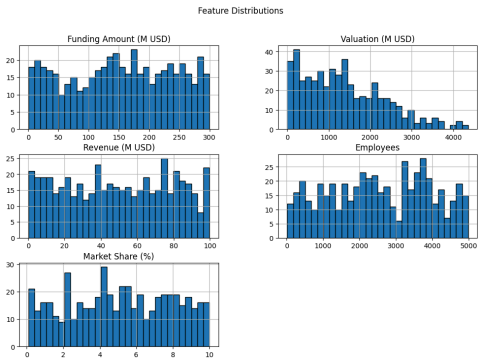

Figure 5. Continuous Numeric Features' Distribution

Continuous numerical variables – Funding Amount, Valuation, Revenue, Employees, and Market Share – were analysed using histograms. Valuation exhibited right-skewness, suggesting the presence of extreme values affecting its distribution.

Funding Amount, Revenue, and Employees exhibited moderate spread (std ≈ 55-59% of mean), necessitating standardisation for comparability. Valuation showed moderate right skewness (skewness = 0.69), requiring a log transformation to mitigate extreme values before standardisation. Market Share, already bounded between 0% and 10%, does not require scaling.

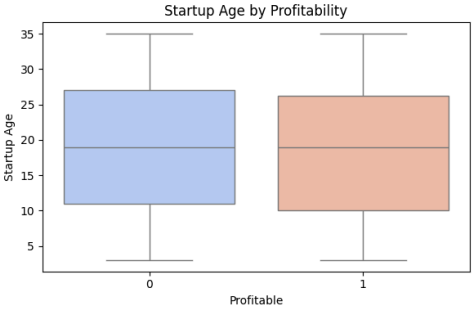

Figure 6. Outlier Detection: Startup Age Distribution by Profitability

Startup Age exhibits no extreme outliers, with all values contained within the whiskers.

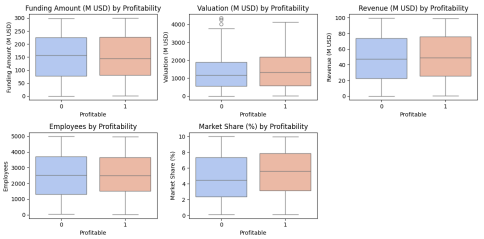

Figure 7. Outlier Detection: Profitability vs. Numerical Features

Outlier analysis using boxplots confirmed significant outliers in Valuation, necessitating transformation for normalisation and improved model performance.

Bivariate Analysis

Bivariate analysis was conducted to examine the relationships between profitability and other features.

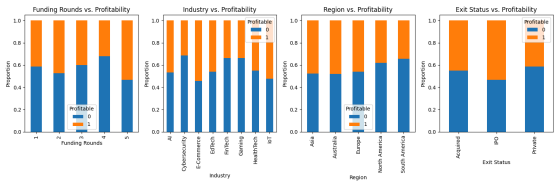

Figure 8. Proportional Distribution of Profitability across Key Categorical Features

Categorical features – Funding Rounds, Industry, Region, and Exit Status – were analysed against profitability using normalised crosstabulation to identify variations across categories.

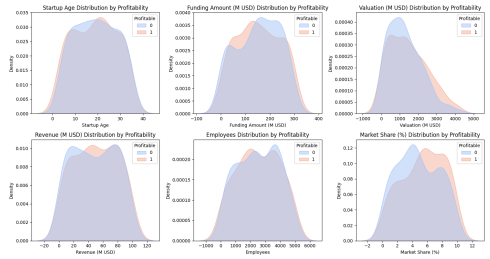

Figure 9. KDE plots for Numerical Features

Kernel Density Estimation (KDE) plots were used to examine the distribution of Startup Age, Funding Amount, Valuation, Revenue, Employees, and Market Share across profitable and unprofitable startups.

Revenue and Market Share emerged as the strongest predictors of profitability. While Funding Amount and Valuation influence outcomes, their impact depends on fund utilisation rather than absolute figures. Startup Age plays a role, with 5-20 years being the most profitable range. Funding Rounds also affect profitability, with startups in Round 2 and 5 performing better. Additionally, a large workforce (>5000 employees) does not guarantee profitability, emphasising the importance of operational efficiency.

Feature Engineering

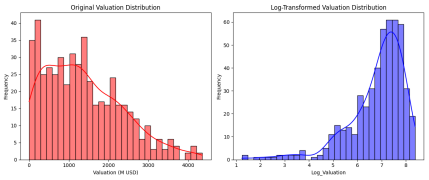

Figure 10. Log Transformation of Valuation

To address right skewness and outliers in Valuation, a log transformation was applied to normalise its distribution. The original column was replaced with Log_Valuation for improved interpretability.

Categorical features were encoded for machine learning compatibility: Funding Rounds (ordinal) was label-encoded, while Industry, Region, and Exit Status (nominal) were one-hot encoded to facilitate analysis.

Feature Selection

Feature selection was performed to identify the most relevant predictors while simplifying the model. A correlation matrix was generated and visualised using a heatmap to examine relationships between features.

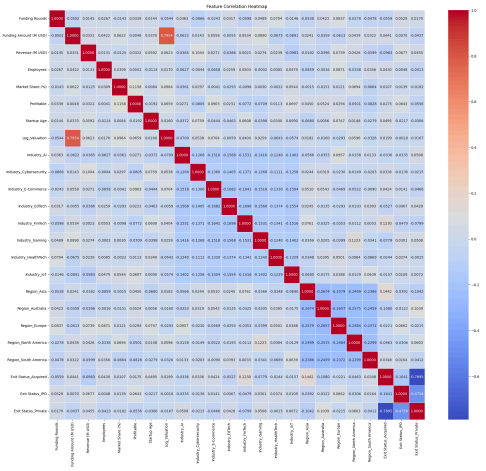

Figure 11. Correlation Matrix

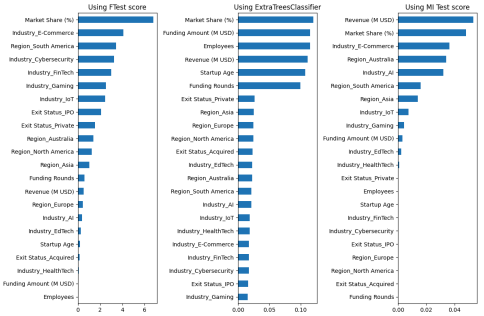

Figure 12. Feature Importance Tests

Feature importance was assessed using F-Test (ANOVA), ExtraTreesClassifier, and Mutual Information (MI Test), with results visualised in a horizontal bar chart.

As shown in Figure 11 (Correlation Matrix) and Figure 12 (Feature Importance Tests), Funding Amount, Industry, Region, and Exit Status were removed due to their correlation with profitability, low importance scores, and risks of overfitting or data leakage. This refinement improved model efficiency while maintaining predictive accuracy.

IV. METHODOLOGIES

This study employed a supervised learning approach, using labelled data to classify startups as profitable or unprofitable. Unsupervised learning was unsuitable due to the lack of predefined labels, and regression was inappropriate for discrete classification.

Model performance was evaluated using AUC score and accuracy, assessing classification effectiveness across thresholds and overall correctness. AUC is beneficial for class imbalance, while accuracy provides a general measure of correctness. Additionally, precision, recall, and F1-score were examined for a comprehensive evaluation, ensuring the model is both accurate and reliable for investment decisions.

Logistic Regression and Random Forest were selected for their complementary strengths in profitability prediction.

Logistic Regression, a linear model, offers interpretability and computational efficiency, making it suitable for linear relationships and feature importance analysis. Random Forest, an ensemble method, captures non-linear relationships, handles high-dimensional data, and reduces overfitting, making it effective for modelling complex financial and market interactions. These models ensure a balanced approach, combining interpretability and predictive performance for robust profitability classification.

V. ANALYSIS, TESTING, RESULTS

After pre-processing and feature selection, two classification algorithms were applied using Scikit-learn's default hyperparameters, while Logistic Regression was adjusted to max_iter = 1000 due to iteration limits Profitability was the dependent variable (y), with six features in (X). The dataset was split 80:20 for training and testing, and five-fold cross-validation was used to evaluate the accuracy and AUC score.

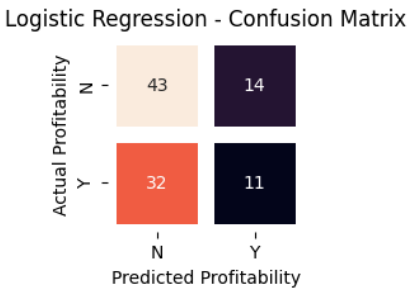

Figure 11. Confusion Matrix for Logistic Regression

Logistic Regression achieved a training accuracy of 0.5925 and a test accuracy of 0.5400, with AUC scores of 0.5927 (training) and 0.5051 (test), indicating underfitting. Cross-validation yielded a mean accuracy of 0.5600 (±0.0483) and a mean AUC of 0.5490 (±0.0557), confirming weak predictive performance. The confusion matrix (Figure 11) shows a high false negative rate (74.4%), with only 11 of 43 profitable startups correctly classified. Precision (0.44), recall (0.26), and F1-score (0.32) for the profitable class further indicate poor model sensitivity in detecting profitable startups.

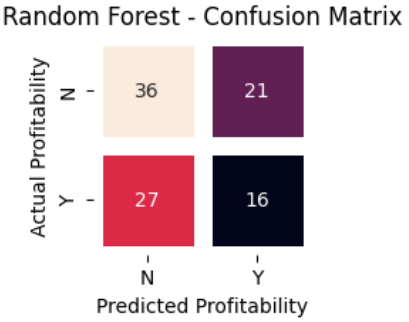

Figure 13. Confusion Matrix for Random Forest

Random Forest exhibited perfect training accuracy (1.0000) but a test accuracy of 0.5200, with AUC scores dropping from 1.0000 (training) to 0.5204 (test), indicating overfitting. Cross-validation yielded a mean accuracy of 0.5875 (±0.0326) and a mean AUC of 0.5789 (±0.0507), suggesting the model struggles to generalise. The confusion matrix (Figure 12) shows that 16 of 43 profitable startups were correctly classified, with a false negative rate of 62.8%. Precision (0.43), recall (0.37), and F1-score (0.40) for the profitable class suggest slightly better performance than Logistic Regression but still poor overall classification reliability.

Both models exhibited limited predictive power. Logistic Regression is underfitted, failing to capture meaningful patterns, while Random Forest is overfitted, leading to poor generalisation. Both models struggled to correctly classify profitable startups, with low recall and F1-scores, high false negatives, and inconsistent cross-validation results. These findings highlight the need for richer financial data, improved feature selection, and advanced modelling techniques to enhance performance.

Hyperparameter tuning was conducted for Logistic Regression and Random Forest to identify optimal configurations to enhance model performance. For Logistic Regression, adjustments focused on regularisation strength, C, to control model complexity, solver selection for optimisation efficiency, and max_iter (increased to 1000) to ensure convergence. For Random Forest, tuning involved n_estimators (number of trees) for stability, max_depth to prevent overfitting, and min_samples_split and min_samples_leaf to balance complexity and generalisation.

Performance was assessed against baseline results, measuring improvements in accuracy, AUC, precision, recall, and F1-score. The objective was to reduce underfitting in Logistic Regression and overfitting in Random Forest, ensuring more reliable profitability predictions.

VI. RESULTS

During model refinement, models were iteratively trained with varying hyperparameters to improve performance, aiming to mitigate underfitting and overfitting for better generalisation. While hyperparameter tuning successfully reduced overfitting in Random Forest, Logistic Regression remained underfitted.

The evaluate_model function was employed after each training cycle to assess model effectiveness, computing training and test accuracy, AUC scores, and classification metrics. Cross-validation was utilised to ensure predictive stability across different data subsets.

AUC Score (ROC-AUC) effectively measures the model's ability to distinguish profitable from unprofitable startups. F1-score is crucial for handling class imbalances by balancing precision and recall. Cross-validation assesses model stability, identifying potential overfitting despite high AUC. Accuracy, precision, and recall provide deeper insights into classification errors and overall predictive reliability.

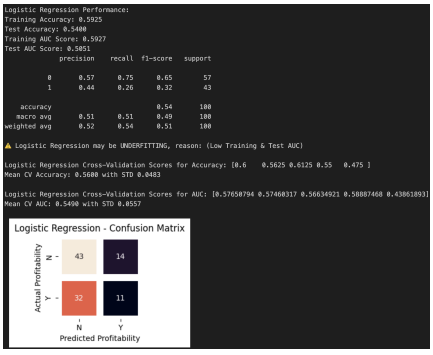

Figure 14. Logistic Regression Model Evaluation

For Logistic Regression, the default hyperparameters with max_iter=1000, and random_state=42 yielded the best performance, indicating that further adjustments did not significantly improve model generalisation.

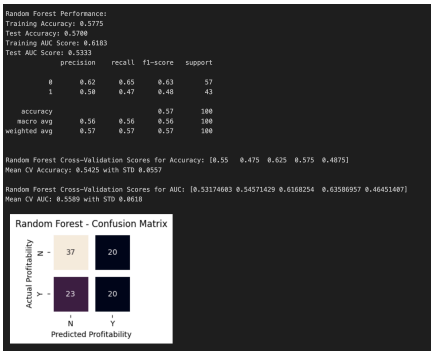

Figure 15. Random Forest Model Evaluation

The Random Forest model was configured with n_estimators=3, max_depth=3, min_samples_split=88, min_samples_leaf=58, random_state=42. This setup aimed to reduce overfitting by limiting tree complexity and enforcing a higher threshold for splits and leaf nodes.

Random Forest performs better overall, while Logistic Regression remains more conservative in profitability prediction.

VII. CONCLUSION

This study examined whether machine learning models could predict startup profitability using financial and operational factors. Logistic Regression and Random Forest achieved moderate accuracy, but their predictive power was limited due to dataset constraints. While Random Forest performed slightly better, its AUC score (0.5333) indicates weak classification ability, highlighting the limitations of machine learning alone in profitability prediction.

The primary constraints arise from missing key business indicators, including historical financial trends, cost structures, and customer-related metrics, restricting the model's ability to differentiate between genuinely profitable startups and those with high funding or market share. Operational efficiency metrics were also lacking, further limiting predictive accuracy.

Reformulating the problem as a regression task to predict profit margins rather than binary profitability could provide a more granular

financial assessment. Advanced models such as Gradient Boosting (XGBoost, LightGBM) and Deep Learning could improve predictive performance by capturing complex non-linear relationships. Integrating external financial data, industry trends, and investor sentiment analysis could further enhance accuracy.

These findings have practical implications for investors, entrepreneurs, and financial institutions. Enhanced profitability prediction models could support venture capital decision-making, aid startup financial planning, and assist banks in risk assessment. Future research should incorporate explainable AI techniques (e.g., SHAP, LIME) to improve model transparency and refine feature selection for more reliable investment insights.

VIII. REFERENCES

Conference Paper / Journals / Thesis

Bento, F. R. da S. R. (2017). Predicting startup-success with machine learning. Master's dissertation. University of Nova Lisboa. Available at: https://run.unl.pt/bitstream/10362/33785/1/TGI0132.pdf [Accessed 8 Feb 2025]

Böhm, M., Weking, J., Fortunat, F., Müller, S. and Welpe, I. (2017). The Business Model DNA: Towards an Approach for Predicting Business Model Success. In: 13th International Conference on Wirtschaftsinformatik, pp. 1006-1020. Available at: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1102&context=wi2017 [Accessed 8 Feb 2025]

Cirjevskis, A. (2018). Exploration of qualitative success factors of innovative e-business startups: Blue Ocean strategy versus dynamic capabilities. *International Journal of Business and Emerging Markets*, June. Available at: https://doi.org/10.1504/IJBEX.2017.087755 [Accessed 8 Feb 2025]

Giardino, C., Bajwa, S. S. and Wang, X. (2015). Key Challenges in Early-Stage Software Startups. In: International Conference on Agile Software Development, pp.52-63. Available at: https://doi.org/10.1007/978-3-319-18612-2_5 [Accessed 8 Feb 2025]

Krishna, A., Agrawal, A. and Choudhary, A. (2016). Predicting the outcome of startups: Less failure, more success. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, pp. 798-805. Available at: https://doi.org/10.1109/ICDMW.2016.0118 [Accessed 8 Feb 2025]

Martinez, D. C. (2019). Startup Success Prediction in The Dutch Startup Ecosystem. Master's Thesis. Delft University of Technology. Available at: https://resolver.tudelft.nl/uuid:1adc2972-db09-4583-b2da-05fd4e462941 [Accessed 8 Feb 2025]

Spiegel, O., Abbassi, P., Zylka, M. P. Schlagwein, D., Fischbach, K. and Schoder, D. (2015). Business model development, founders' social capital, and the success of early-stage internet start-ups: A mixed-method study. *Information Systems Journal*. Available at: https://doi.org/10.1111/isj.12073 [Accessed 8 Feb 2025]

Ünal, C. (2019). Searching for a Unicorn: A Machine Learning Approach Towards Startup Success Prediction. Master's Thesis. Humboldt Universität, Berlin. Available at: https://hdl.handle.net/10419/230798 [Accessed 8 Feb 2025]

Welter, C., Holcomb, T. R., and McIlwraith, J. (2023). The inefficiencies of venture capital funding. *Journal of Business Venturing Insights*, 19: e00392. Available at: https://doi.org/10.1016/j.jbvi.2023.e00392 [Accessed 8 Feb 2025]

Zacharakis, A. L. and Meyer, D. G. (1998). A lack of insight: Do venture capitalists really understand their own decision process?, *Journal of Business Venturing*, 13(1), pp. 57-76. Available at: https://doi.org/10.1016/S0883-9026(97)00004-9 [Accessed 8 Feb 2025]

Żbikowski, K. and Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data, *Information Processing & Management*, 58(4): 102555. pp. 1-18. Available at: https://doi.org/10.1016/j.ipm.2021.102555 [Accessed 8 Feb 2025]

Web References

Ashar, S. (2025). Startup Growth & Funding Trends. Kaggle. Available at: https://www.kaggle.com/datasets/samayashar/startup-growth-and-funding-trends [Accessed 26 Feb 2025]