Startup Success Prediction with PCA-Enhanced Machine Learning Models

Youngkeun Choi

Abstract

This study evaluates the effectiveness of various machine learning algorithms in predicting startup success and explores the performance improvement achieved by applying Principal Component Analysis (PCA) to the models. By analyzing logistic regression, support vector classifier (SVC), XGBoost, and other supervised learning algorithms, the study demonstrates that PCA enhances the generalization performance of most models. Notably, Support Vector Classifier (SVC) showed an accuracy of 0.78, precision of 0.83, recall of 0.73, and F1 score of 0.74 without PCA, but performance significantly improved with PCA, recording an accuracy of 0.90, precision of 0.90, recall of 0.89, and F1 score of 0.89. Academically, this research contributes to the literature by examining how dimension reduction can boost the accuracy of machine learning models for startup success prediction, providing a valuable intersection of machine learning and venture capital studies. Practically, it offers investors AI-driven decision-making tools to enhance the precision of investment evaluations and better identify startups with high growth potential. Despite its contributions, this study is limited by the specific dataset used, suggesting that future research could explore various datasets and alternative dimension reduction techniques. Future studies could also assess real-time data application and incorporate deep learning models to improve predictive performance in startup success evaluation.

Keywords: Startup success prediction, Machine learning, Principal Component Analysis (PCA), Support Vector Classifier (SVC), Venture capital, Investment decision-making

Submitted: November 7, 2024 / Approved: December 18, 2024

1. Introduction

With the rise of startups and their growing economic impact, entrepreneurs, investors, and decision-makers increasingly require effective methods to analyze business data from various perspectives. However, identifying relevant factors influencing business volatility has become a challenging task due to ongoing technological advancements, competitive markets, and industry innovation. Recent studies have focused on factors such as mergers and acquisitions, financial determinants for business success, and investments essential for achieving IPO status (Ross et al., 2021). Nevertheless, these studies often examine only specific methods or limited factors, indicating certain research gaps.

While venture capital (VC) investment plays a critical role in the global economy, many investments consistently deliver low returns for investors (Mulcahy et al., 2012). A study analyzing annual returns for VC funds established since 1998 through June 2019 found that the top quartile of funds achieved an average return of 24.8%, while the bottom quartile reported an average return of just 0.5%, effectively incurring losses for limited partners when adjusted for inflation (Associates, 2020). Further insight into VC funds' poor performance reveals that from 2000 to 2010, VC returns were lower than S&P 500 returns (Guzy, 2010). Similarly, the Kauffman Foundation (Mulcahy et al., 2012) reported that between 1997 and 2012, VC funds returned less cash to investors than the capital initially raised. Among 30 VC funds with over $400 million in committed capital, only 4 outperformed the S&P 500. This disparity in success rates extends to individual investors, who, with limited investment choices, often invest small amounts in startups online, bypassing traditional financial intermediaries for minor equity stakes (Mollick, 2014). However, equity crowdfunding is high-risk (Vroomen & Desa, 2018) and often attracts lower-quality entrepreneurs linked to risky banks, resulting in high failure rates (Blaseg et al., 2021).

AI tools with potential to enhance return on investment (ROI) are likely to support VC investment decisions. AI has transformed decision-making in finance—improving credit evaluation, quantitative trading, risk management, fraud detection, and stock trading (Castleman, 2020)—and is expected to significantly impact financial services by enhancing decision accuracy (Ryll et al., 2020). Despite these advances, venture capitalists remain hesitant to adopt AI in investment decision-making, relying heavily on personal networks and subjective judgment (Weibl & Hess, 2019). In contrast, individual investors in equity crowdfunding tend to be influenced primarily by promotional content provided by startups, leading to more superficial assessments (Wang et al., 2020).

(1) Division of Business Administration, College of Business, Sangmyung University,
E-mail: penking1@smu.ac.kr

Existing research highlights the importance of selecting appropriate machine learning (ML) models based on business goals to effectively predict business outcomes (Gangwani & Zhu, 2024). For instance, unsupervised learning methods are commonly applied in financial stability analysis and product marketing, while supervised learning methods leverage diverse business features to predict success or survival in startups. Recent findings demonstrate that deep learning approaches are increasingly efficient for predicting business failure using financial and historical data.

Therefore, this study adopts supervised ML models to predict startup success, providing valuable support for investment decision-making by evaluating critical metrics. However, the data used to predict startup success often involves complex interrelationships between variables, potentially reducing predictive performance due to multicollinearity during the training process. To address this, the present study applies Principal Component Analysis (PCA) within supervised ML models to account for the unique characteristics of startup data, exploring how AI can optimize startup investment decisions by improving success prediction accuracy.

2. Related Work

In this section, we focus on explaining various machine learning models categorized as supervised or unsupervised learning, depending on the availability of company goals and features. Supervised learning models are further divided into two subgroups: regression and classification, both commonly used for predicting and forecasting business success. Based on label information, models are categorized into three types: binary classification, multi-class classification, and continuous variable prediction.

First, binary classification is frequently employed to predict business success, with many researchers using various machine learning algorithms. For instance, several studies have utilized machine learning algorithms to predict business outcomes for startups and SMEs (Pasayat et al., 2020). Logistic Regression (LR), Support Vector Machine (SVM), and Gradient Boosting are commonly used to predict the success of startups based on VC funding (Żbikowski & Antosiuk, 2021). The primary goal is to develop an unbiased predictive model that VCs and stakeholders can confidently use in real prediction scenarios. The target variable is labeled as "success" based on completing a second funding round, indicating a company's stability in generating sufficient future revenue. A previous study (Gangwani et al., 2023) developed new features to predict business success or failure, including acquisitions and IPOs, using the triangular relationship among investors, businesses, and markets. The study found that adding these relationship-based features improved prediction accuracy compared to using simple features alone. Another definition of a successful startup is based on company survival. McKenzie et al. (2017) used three machine learning approaches—SVM, LASSO (Least Absolute Shrinkage and Selection Operator), and Boosted Regressor—to predict survival probability based on startup revenue and profit. Böhm et al. (2017) proposed using clustering techniques and SVM to predict survival probability. Startups can also be evaluated based on innovation or project level, where innovation serves as a critical turning point for new achievements and economic growth. Kinne and Lenz (2021) proposed a method to predict startup success from a business innovation perspective by classifying data from surveys of multiple firms and using deep learning to capture product innovations on company websites for business outcome prediction. Guerzoni et al. (2019) argued that innovation improves a firm's survival probability, using seven supervised learning approaches, including classification, regression trees, logistic regression, Naive Bayes, and artificial neural networks (ANN), to predict survival rates from an innovation standpoint.

Second, multi-class classification offers entrepreneurs and investors new perspectives in assessing business outcomes. Many companies ultimately fail due to insufficient funding, poor marketing strategies, or lack of competitiveness in similar markets, putting them at risk of bankruptcy. Multi-class algorithms help investors identify risks or bankruptcy likelihood and assess a company's current position relative to its revenue. These algorithms classify business outcomes into multiple categories, such as "risk," "failure," "survival," and "bankruptcy," beyond simply predicting success or failure. This allows investors to make more informed decisions based on a company's current status and investment potential. For example, Jones and Wang (2019) proposed the TreeNet method, based on gradient boosting, to predict bankruptcy risk for private companies, enabling risk analysis before a company files for bankruptcy. Arroyo et al. (2019) used time-sensitive analysis to predict company success by categorizing firms as acquisitions, funded, or IPOs. Using a sliding window to measure time, this approach provides predictions that guide investors on the appropriate timing for company growth investments, aiming for investment at the acquisition or IPO stage.

Finally, business growth measurements in business prediction models can be quantified. Evaluating growth effectively requires continuous assessment across various factors and variables. Growth rates vary by company, depending on the set target variable. For startups, growth may be measured by sales profit or units sold, while for SMEs, it may be based on revenue, employment, or customer service. External factors like market conditions, business environment, and product distribution should also be considered. Regression models in machine learning have proven to be powerful tools for business growth prediction.

Unsupervised machine learning models are also widely used for business data analysis, finding hidden patterns or discovering meaningful groups within a given dataset. One of the primary advantages of unsupervised learning is that it doesn't rely on labeled data. Unsupervised learning is broadly divided into four categories: clustering, association rule mining, outlier detection, and dimensionality reduction.

First, clustering helps predict business failure or survival by analyzing customers with similar behaviors, products with similar profiles, or companies with similar growth or failure histories. Various clustering methods are used in business data analysis, including k-means clustering, partition-based clustering, density-based clustering, hierarchical clustering, and model-based clustering. Among these, k-means clustering is most commonly used in business due to its transparency

and similarity-based approach. Density-based clustering is useful for grouping information and filtering noise (outliers) from data. A study used clustering, diffusion theory, and density estimation to analyze early sales data for predicting new product success (Garber et al., 2004). Key findings indicated that changes in data point density serve as early warning signals for business status. Business Model DNA (Böhm et al., 2017) describes business process characteristics, like the human genome, using a combined approach of SVM and k-means clustering to identify similar clusters based on different growth types (e.g., slow growth, fast growth). This approach improves prediction accuracy compared to previous studies. Shah and Murtaza (2000) demonstrated a bankruptcy prediction method using neural networks combined with clustering techniques. The neural network architecture involved three layers: the first layer clustered companies based on financial ratios, the second layer used time-series data to learn financial trends, and the third layer included two neurons to classify companies as bankrupt or non-bankrupt.

Second, in the finance sector of the business domain, association rules reveal relationships between business operations and financial status. Specific rules applied to financial data can identify combinations of business operations at risk of bankruptcy (Martin et al., 2011). The Apriori algorithm, combined with a financial domain ontology, identifies a company's strengths and weaknesses, such as accounting health or total debt, aiding strategic planning and decision-making. Overall, association rule mining offers valuable insights into customer behavior, product portfolios, and financial analysis, enabling companies to make data-driven decisions on product placement, pricing, operations, and promotional strategies, ultimately driving sales and business success.

Third, outliers in the business world often signify crucial risks or values. In banking and credit card sectors, outlier detection models are used to identify irregularities and predict risks within the business domain. Outliers vary in definition, but local outliers, in particular, are useful for identifying samples that differ from others within a specific region. The Local Outlier Factor (LOF) compares the local density of data points with that of neighboring points. A previous study (Chen et al., 2007) employed LOF in banking to detect inconsistencies or fraud, contributing to reliability and customer satisfaction. Overall, outlier detection provides companies with valuable insights, improving operations and risk management. This enables early risk detection and necessary actions to mitigate the impacts of such risks, leading to better decision-making and ultimately enhancing business strategy.

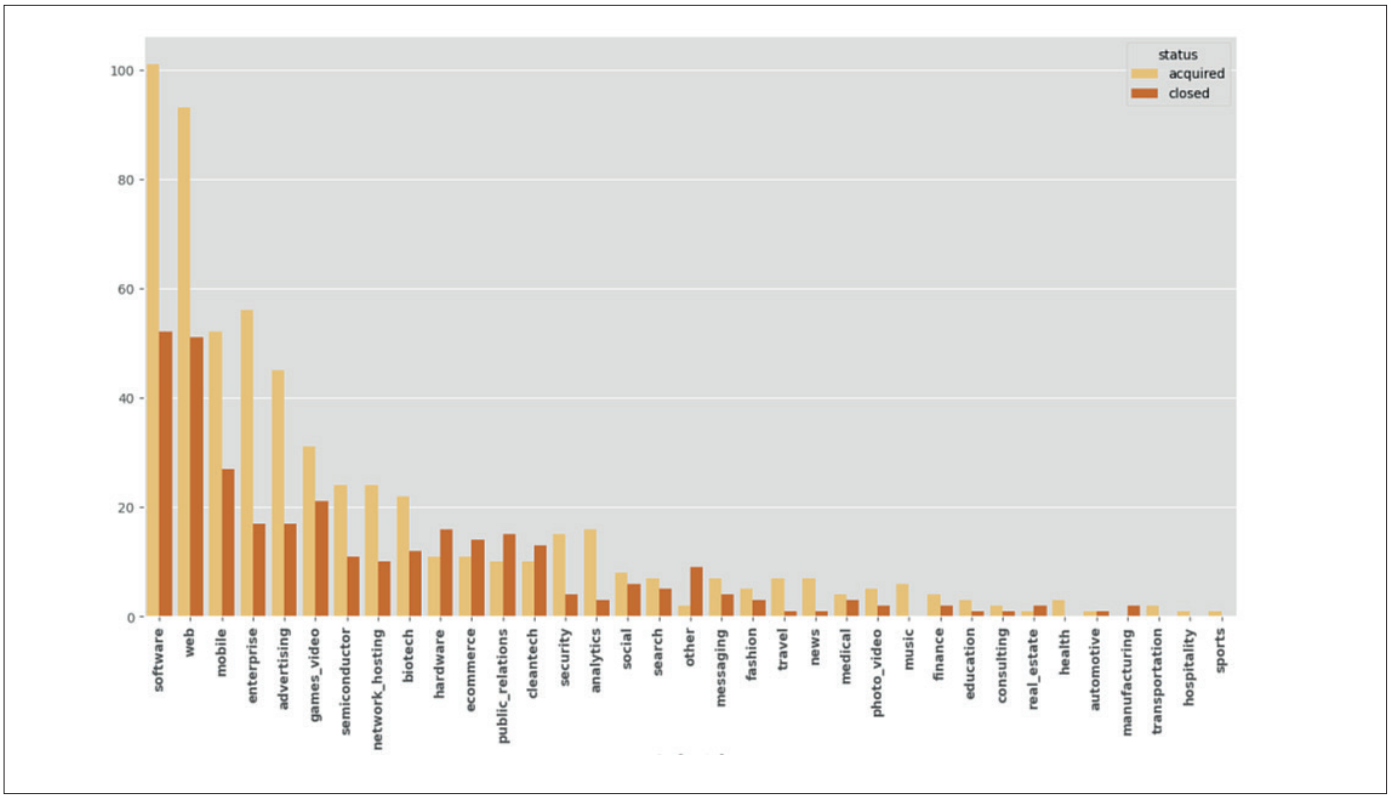
Finally, dimensionality reduction is popular in various data analysis fields and is also used in business prediction. This technique simplifies complex datasets by reducing the number of features (dimensions) while preserving essential information. Wang and Wu (2017) proposed a two-stage ensemble approach, using feature selection to remove redundant data and improve business failure prediction. The final subset included carefully selected financial metrics containing information about healthy and failing companies. Three manifold learning algorithms—ISOMAP, Linear Embedding (LE), and Locally Linear Embedding (LLE)—were applied to select various feature subsets, comparing their performance with PCA to improve model performance. Another study (Tsai, 2009) aimed to predict business bankruptcy using financial ratios. This study used PCA to reduce data dimensionality and identify critical financial ratios for bankruptcy prediction, showing that dimensionality reduction improves model accuracy by explaining 91% of the total variance with five principal components. Recent studies (Sivasankar et al., 2017; Rtayli & Enneya, 2019) have utilized feature selection and extraction (PCA and LDA) techniques for credit risk or fraud detection. Rtayli and Enneya (2019) combined Random Forest with feature filtering to detect credit card fraud, using the Gini index to calculate feature importance scores in financial datasets and constructing decision trees to determine final classes. These studies emphasize the importance of dimensionality reduction in business prediction by identifying the most crucial variables for the target variable and enhancing prediction model accuracy.

3. Methodology

3.1 Dataeset

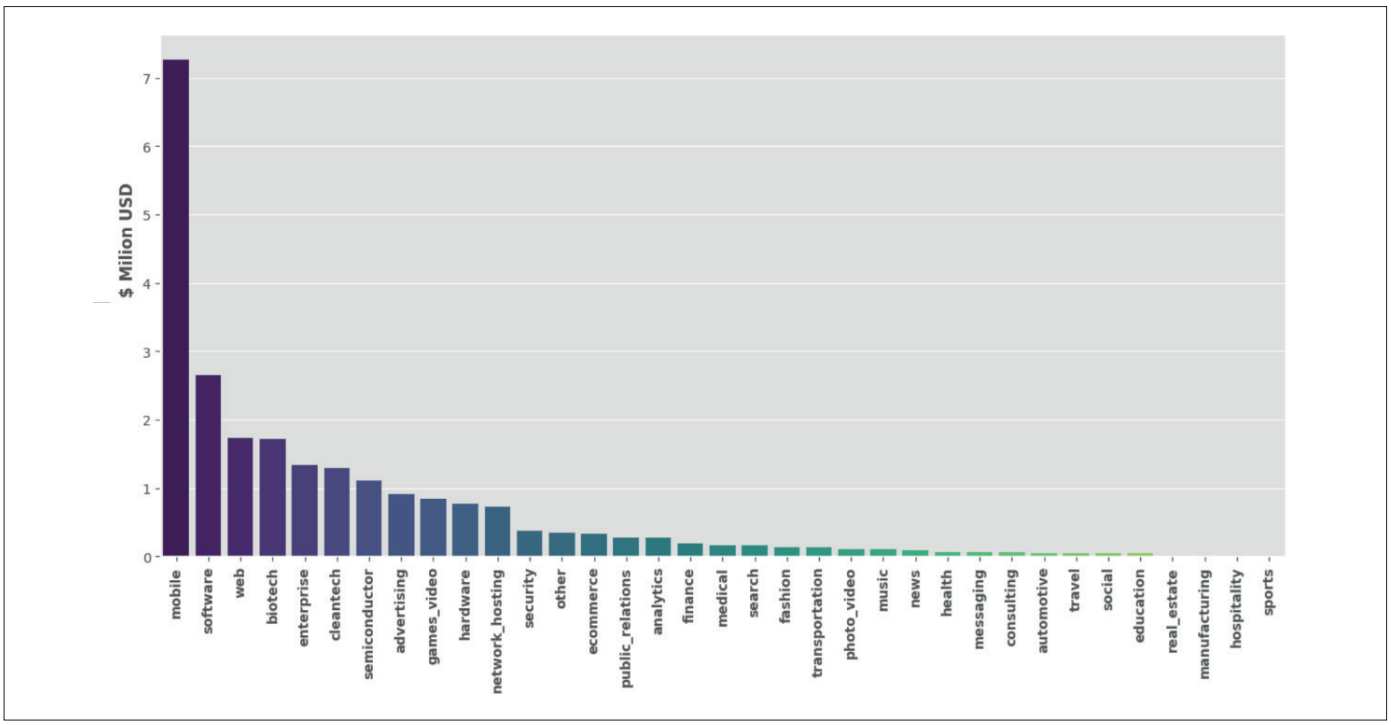
First, the data required to test the supervised machine learning models with PCA Appliedin predicting startup success needs to be collected. The dataset obtained from Kaggle contains information essential for predicting startup success (Zbikowski et al., 2021) which is startup data in the United States contains 923 rows and 48 features, including quantitative and categorical attributes, such as age at first and last funding year, relationships, funding rounds, total funding (USD), milestones, state, industry type, presence of venture capital (VC), angel investors, and funding rounds (Round A, B, C, D). The target variable "status" categorizes startups as "acquired" or "closed," representing the ultimate measure of startup success or failure, respectively. The dataset's potential to predict startup success allows investors and decision-makers to gain a competitive advantage by identifying high-growth prospects and fostering a thriving entrepreneurial ecosystem. The dataset was used in data sprint #5 at DPhi, and acknowledgments go to Ramkishan Panthena, a Machine Learning Engineer at GMO, for providing this dataset. The data is processed using machine learning to determine the distribution of successful and unsuccessful (closed) startup data based on the categories presented in Figure 1.

Figure 1. Startup categories



From Figure 1, it can be seen that the software category has the highest number of successes followed by startups in the web and other categories. The smallest number of successes is the sports, hospitality, and other categories. Next, we will look at the data based on startup funding presented in Figure 2.

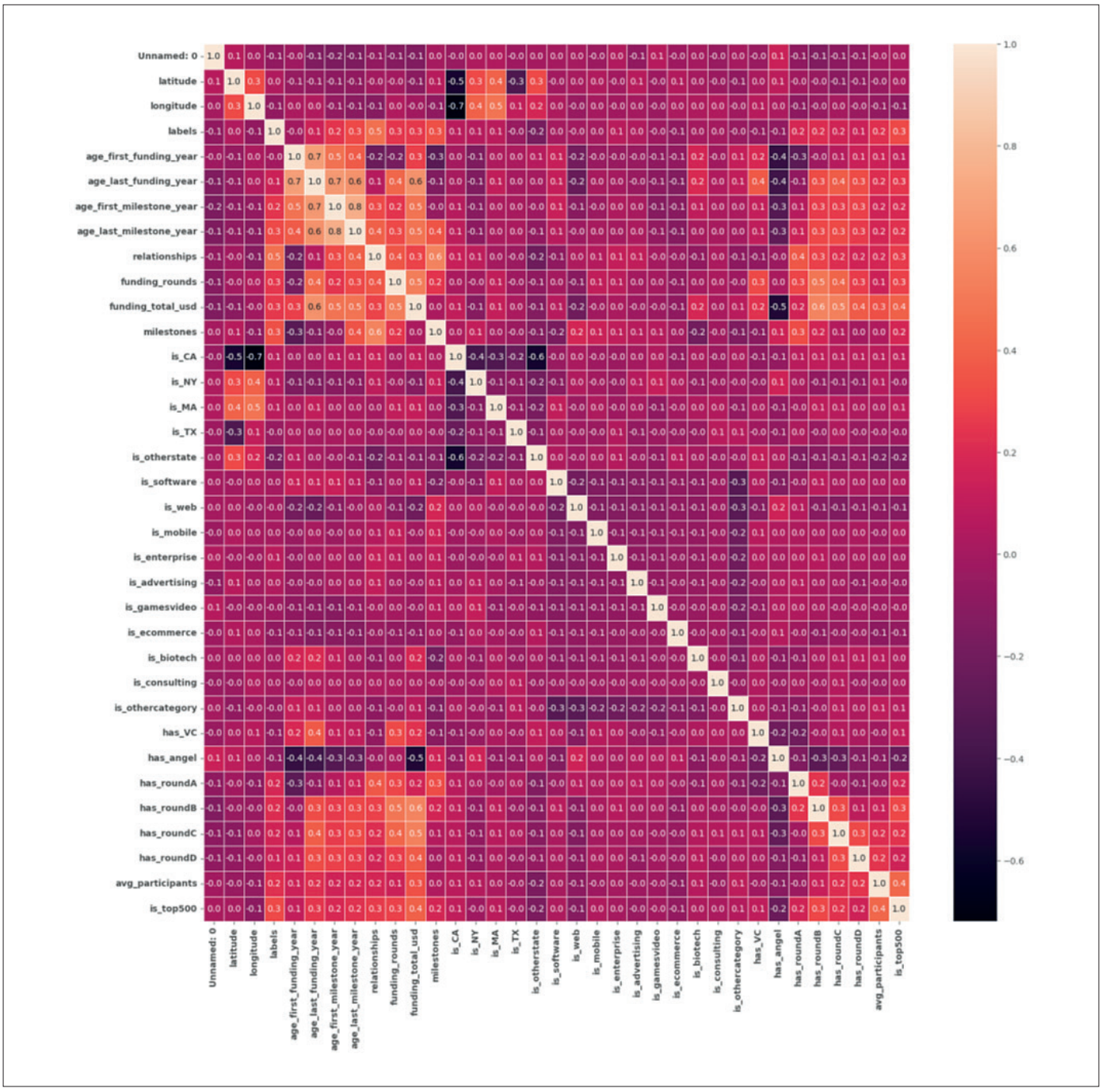
Figure 2. Starup financing



3.2 Data preprocessing

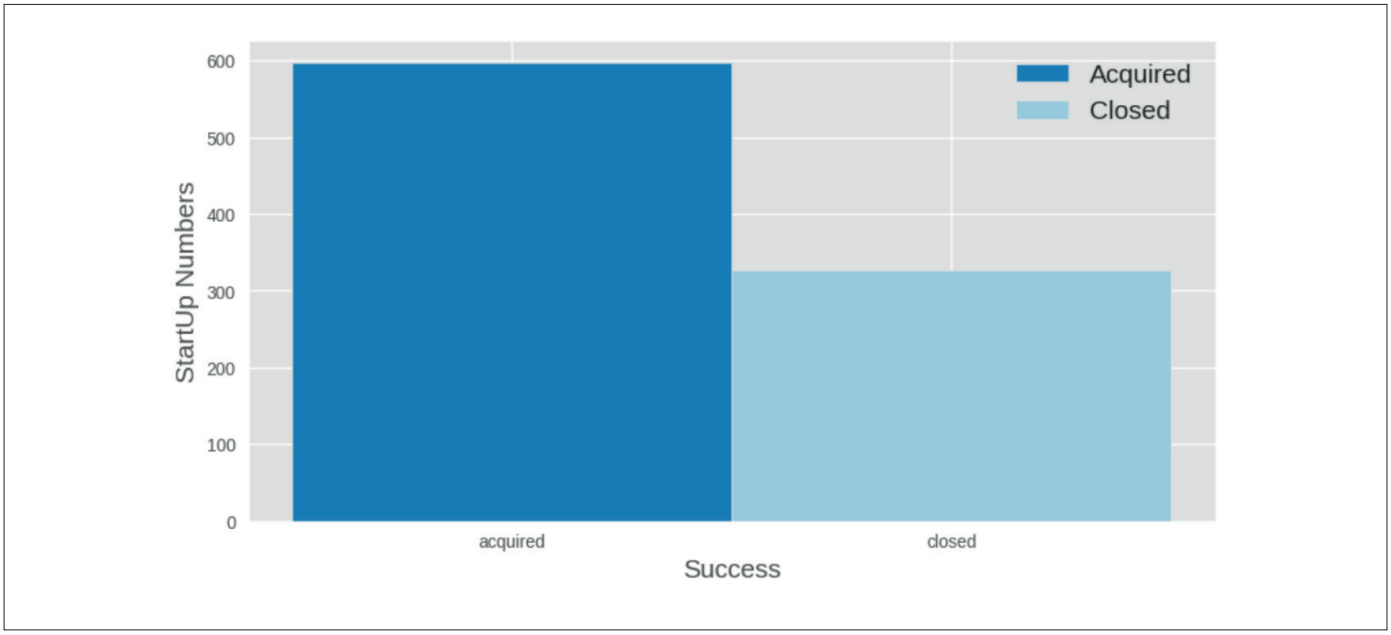
Once the data has been collected, the next step is to perform data preprocessing. This includes processing missing data, removing irrelevant data, and data normalization. Data preprocessing should also consider the problem of unbalanced data. Furthermore, to aid in data preprocessing, we analyze the correlation matrix of each attribute to identify relationships between variables. The results of this analysis are presented in Figure 3. The attributes with a correlation value close to 1 are selected as training data, resulting in 36 attributes used for the training process.

Figure 3. Attributes correlations



Attribute correlation provides an overview of the relationship between variables and also the contribution of input features to the output. The value of the correlation coefficient ranges from 0 to 1. A value of 0 indicates no correlation, while a value of 1 describes a full correlation. Good selection means that the selected input variables have a small correlation. The features selected in this study are good candidates for investigating ML models. The target distribution used to display the number of successful and unsuccessful startup data is presented in Figure 4.

Figure 4. Target distributions



Referring to Figure 4, the distribution of targets in successful startups is higher than that of unsuccessful ones. It shows that the success rate is higher than that of unsuccessful startups.

The next step is to divide the data into two parts, namely training data and testing data. Training data uses 70% of existing data while testing data uses 30% of existing data that will be used to test the performance of the algorithm. This splitting technique will affect the result of the model [23].

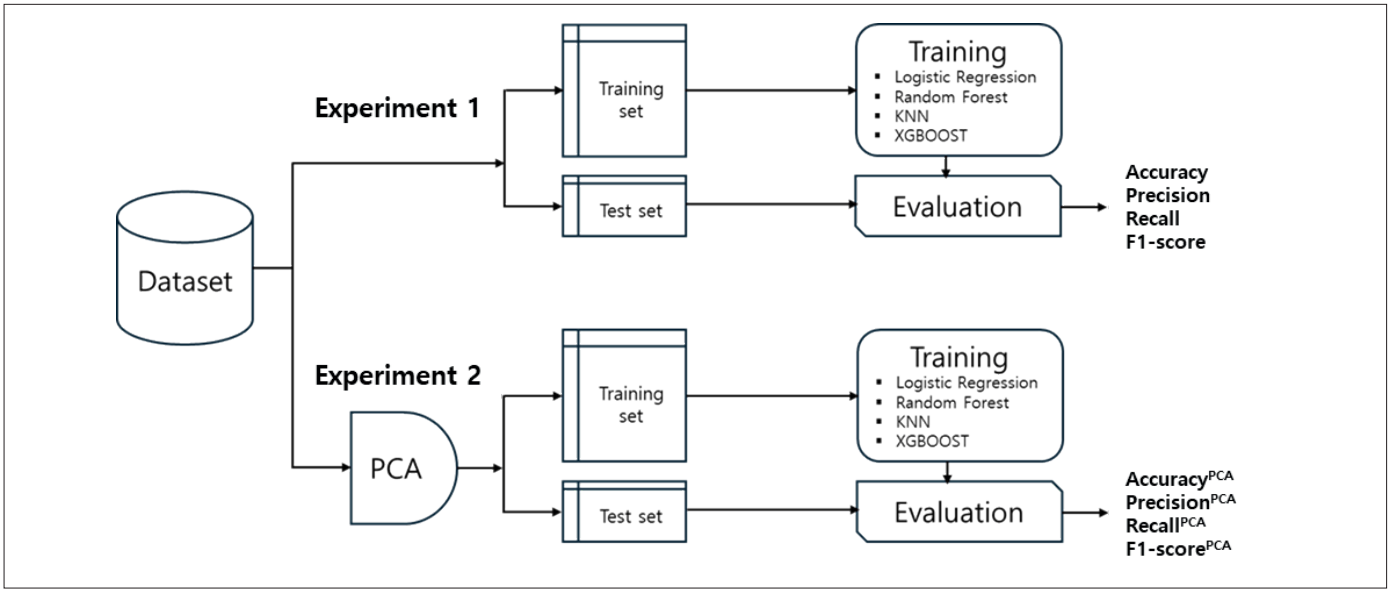
3.3 Proposed experimental model

This study proposes the following experimental models to predict startup success. First, in the initial experiment, the dataset undergoes preprocessing before being split into training and test datasets. The training dataset is then used to train various supervised machine learning algorithms, including Logistic Regression, Random Forest, KNN, and XGBoost. Each trained algorithm is subsequently tested on the test dataset, and metrics such as accuracy, precision, recall, and F1-score are calculated for each model.

In the second experiment, the dataset undergoes preprocessing, followed by the application of PCA to generate new factor variables. The PCA-transformed data is then split into training and test datasets. Similar to the first experiment, the training dataset is used to train Logistic Regression, Random Forest, KNN, and XGBoost models. Each trained algorithm is then tested on the test dataset, and performance metrics, specifically accuracyPCA, precisionPCA, recallPCA, and F1-scorePCA, are calculated.

Through this approach, the study aims not only to identify the most suitable supervised machine learning model for predicting startup success but also to determine whether the proposed PCA application improves predictive performance by removing multicollinearity among variables in the supervised learning models.

Figure 5. Proposed experimental model



3.4 Performance indices

Confusion matrix is a tool that evaluates classification models using matching between actual classes and predicted classes. In this study, positive and negative are specified to suit the purpose of the experimental models proposed in this study, as shown in Table 1 below.

Table 1. Confusion matrix

		Actual	
		Positive (Buy)	**Negative (Hold)**
Predicted	Positive (Buy)	TP (True Positive)	FP (False Positve)
	Negative (Hold)	FN (False Negative)	TN (True Negative)

This study uses four evaluation indicators: accuracy, precision, recall, and F1-score to measure the purchase decision performance of the experimental modelproposed in this study as shown in Table 2.

Table 2. Evaluation indices for performance

	Calculation	Explanation
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	Accuracy is the ratio of the sum of the number of cases that are actually Buy and the prediction is also Buy, and the number of cases that are actually Buy but the prediction is Hold, among all identified samples. It is the simplest indicator to evaluate a classification model, but it has the disadvantage of being difficult to evaluate datasets with unbalanced classes.
Precision	$\frac{TP}{TP+FP}$	Precision is the ratio of samples that are actually Buy among those judged as Buy by the prediction. Precision indicates how accurate the result detected as Buy is.
Recall	$\frac{TP}{TP+FN}$	Recall is the ratio of predicted buys among samples that are actual buys. Recall indicates how accurately the model predicts the actual Positive class.
F1-score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$	F1-score is used to simultaneously consider the precision and recall performance of the model. Fi-score is a value between 0 and 1, and the closer it is to 1, the better the prediction performance.

Results

Table 3 presents a comparative analysis of performance evaluation metrics for various machine learning algorithms. Each algorithm was assessed based on accuracy, precision, recall, and F1-score. Logistic Regression demonstrated an accuracy of 0.80, precision of 0.81, recall of 0.77, and an F1-score of 0.78, indicating overall high performance. Notably, the high precision suggests that the model made a significant number of correct predictions. Random Forest maintained consistent values across accuracy, precision, and recall at 0.76, with an F1-score of

0.75. While it exhibits stable performance, it is slightly inferior to that of Logistic Regression. K-Nearest Neighbors (KNN) achieved an accuracy of 0.78, precision of 0.80, recall of 0.74, and an F1-score of 0.75, indicating relatively high precision among the evaluated models. Support Vector Classifier (SVC) showed an accuracy of 0.78, precision of 0.83, recall of 0.73, and an F1-score of 0.74. The exceptionally high precision suggests a lower rate of false positives, which may be advantageous in specific contexts. XGBoost recorded an accuracy of 0.79, precision of 0.79, recall of 0.76, and an F1-score of 0.77, displaying balanced performance metrics comparable to those of Logistic Regression. From this comparison, it is evident that Logistic Regression and XGBoost exhibit relatively superior performance. Additionally, the high precision of SVC may be particularly beneficial in scenarios where minimizing false positives is critical.

Table 3. Experiment 1

	accuracy	precision	recall	f1-score
Logistic Regression	0.80	0.81	0.77	0.78
Random Forest	0.76	0.76	0.76	0.75
KNN	0.78	0.80	0.74	0.75
SVC	0.78	0.83	0.73	0.74
XGBOOST	0.79	0.79	0.76	0.77

The reasons for the relatively superior performance of Logistic Regression and XGBoost, as well as the potential advantage of high precision in SVC under specific conditions, are as follows. Logistic Regression, though a relatively simple linear model, proves highly effective when the data is linearly separable. This model is less susceptible to overfitting, resulting in strong generalization performance, fast training, and ease of interpretation. Therefore, Logistic Regression demonstrates stable performance in terms of accuracy and F1 score. XGBoost, an ensemble learning method based on boosting, combines individual weak learners to create a powerful model. This model delivers robust performance and handles data non-linearity and complex interactions effectively, leading to high performance. Additionally, the model adapts well to data characteristics, resulting in evenly high precision, recall, and F1 scores.

The advantage of high precision in SVC is as follows. Precision refers to the proportion of true positive cases among those predicted as positive by the model. This metric is particularly important in scenarios sensitive to false positives, such as fraud detection, medical diagnostics, and spam filtering, where the costs or damage resulting from misclassification as positive can be significant. SVC recorded a precision of 0.83, indicating that it effectively reduced false positives, achieving high accuracy in positive case predictions. SVC maintains high precision by optimizing the decision boundary through support vectors, making it particularly effective for high-dimensional data or complex boundaries.

Table 4 presents a comparison of various machine learning algorithms' performance following the application of Principal Component Analysis (PCA). Each model's performance was evaluated based on accuracy, precision, recall, and F1 score. Logistic Regression with PCA achieved an accuracy of 0.88, precision of 0.87, recall of 0.88, and F1 score of 0.88, showing overall high performance. This indicates that Logistic Regression effectively classifies data even after PCA application. Random Forest with PCA, with an accuracy of 0.77, precision of 0.77, recall of 0.72, and F1 score of 0.73, showed relatively lower performance compared to other models. This may indicate some information loss during PCA application, potentially affecting the model's performance. K-Nearest Neighbors (KNN) with PCA achieved balanced performance, recording an accuracy of 0.81, precision of 0.81, recall of 0.77, and F1 score of 0.78, showing higher performance than Random Forest. Support Vector Classifier (SVC) with PCA achieved the highest performance among the models in the table, with an accuracy of 0.90, precision of 0.90, recall of 0.89, and F1 score of 0.89. Notably, SVC maintained high precision and recall even after dimensionality reduction through PCA, demonstrating effective classification. XGBoost with PCA recorded an accuracy of 0.84, precision of 0.83, recall of 0.81, and F1 score of 0.82, achieving consistently high performance similar to Logistic Regression. In summary, SVC with PCA achieved the highest performance, with Logistic Regression with PCA and XGBoost with PCA also demonstrating strong results. These findings indicate that SVC is an effective model even after PCA application reduces data dimensionality.

Table 4. Experiment 2

	accuracy	precision	recall	f1-score
Logistic Regression with PCA	0.88	0.87	0.88	0.88
Random Forest with PCA	0.77	0.77	0.72	0.73
KNN with PCA	0.81	0.81	0.77	0.78
SVC with PCA	0.90	0.90	0.89	0.89
XGBOOST with PCA	0.84	0.83	0.81	0.82

The superior performance of SVC with PCA can be attributed to the following characteristics. First, Principal Component Analysis (PCA) is a dimensionality reduction technique that preserves most of the information while reducing data dimensions, allowing high-dimensional data to be represented efficiently. High-dimensional data often risks model overfitting; however, by selecting only the principal components that capture essential information, the model becomes simpler and noise is reduced. This leads to a more generalized performance for SVC. Second, Support Vector Classifier (SVC) is an algorithm that identifies the optimal decision boundary (margin) through support vectors. Particularly in high-dimensional spaces, SVC defines decision boundaries precisely, yielding high performance. Even after dimensionality reduction with PCA, crucial features remain, enabling SVC to find an optimal decision boundary and maintain high accuracy and precision. SVC is especially advantageous for data requiring nonlinear boundaries and can identify complex boundaries between data points, resulting in high classification performance with balanced precision and recall, leading to an elevated F1 score. Third, by removing variables related to unnecessary noise and retaining only significant features, PCA effectively reflects data variance, reducing overfitting and enhancing model generalization. SVC, by selecting optimal support vectors in this simplified data, is able to maintain stable, high performance. In conclusion, the combination of dimensionality reduction effects of PCA and the optimal boundary formation capability of SVC contributes to the model's high performance.

Table 5 shows the performance variations of various machine learning algorithms based on the application of Principal Component Analysis (PCA). Performance evaluation metrics include accuracy, precision, recall, and F1 score. Logistic Regression without PCA recorded an accuracy of 0.80, precision of 0.81, recall of 0.77, and F1 score of 0.78. With PCA applied, performance improved across all metrics to an accuracy of 0.88, precision of 0.87, recall of 0.88, and F1 score of 0.88. This demonstrates that reducing noise through PCA enhances the model's generalization performance. Random Forest, without PCA, achieved approximately 0.76 on all metrics, and while applying PCA resulted in minor changes, recall and F1 score slightly decreased, resulting in an accuracy of 0.77, precision of 0.77, recall of 0.72, and F1 score of 0.73. This indicates that PCA has minimal impact on Random Forest and may sometimes cause information loss. K-Nearest Neighbors (KNN) recorded an accuracy of 0.78, precision of 0.80, recall of 0.74, and F1 score of 0.75 without PCA, and showed a slight improvement post-PCA, achieving an accuracy of 0.81, precision of 0.81, recall of 0.77, and F1 score of 0.78, indicating a positive effect of dimensionality reduction on KNN performance. Support Vector Classifier (SVC) showed an accuracy of 0.78, precision of 0.83, recall of 0.73, and F1 score of 0.74 without PCA, but performance significantly improved with PCA, recording an accuracy of 0.90, precision of 0.90, recall of 0.89, and F1 score of 0.89. This reflects SVC's capability to handle high-dimensional data and suggests that removing unnecessary variables through PCA enables SVC to achieve higher performance. XGBoost recorded balanced performance before PCA, with an accuracy of 0.79, precision of 0.79, recall of 0.76, and F1 score of 0.77. It maintained similar performance after PCA (accuracy of 0.84, precision of 0.83, recall of 0.81, and F1 score of 0.82), showing consistent performance even without PCA, with some improvement. In summary, SVC and Logistic Regression exhibit the greatest performance improvements post-PCA, with SVC with PCA achieving the highest performance overall. On the other hand, Random Forest shows minimal performance changes with or without PCA. The effectiveness of PCA varies by model, proving particularly beneficial for models like SVC.

Table 5. Comparision

	accuracy	precision	recall	f1-score
Logistic Regression	0.80	0.81	0.77	0.78
Logistic Regression with PCA	0.88	0.87	0.88	0.88
Random Forest	0.76	0.76	0.76	0.75
Random Forest with PCA	0.77	0.77	0.72	0.73
KNN	0.78	0.80	0.74	0.75
KNN with PCA	0.81	0.81	0.77	0.78
SVC	0.78	0.83	0.73	0.74
SVC with PCA	0.90	0.90	0.89	0.89
XGBOOST	0.79	0.79	0.76	0.77
XGBOOST with PCA	0.84	0.83	0.81	0.82

The reason SVC with PCA achieves the highest performance, while Random Forest shows minimal performance change before and after PCA application, stems from the characteristics of each algorithm and the way PCA functions. First, Support Vector Classifier (SVC) is a model that forms an optimal decision boundary in high-dimensional space, demonstrating high performance even when data has complex nonlinear structures. Since PCA reduces high-dimensional data to a lower dimension while preserving key information, SVC can find a meaningful classification boundary through optimal support vectors even in dimensionally reduced data. When PCA reduces noise and irrelevant variance, SVC can establish a decision boundary in a more concise feature space, maximizing generalization performance. As a result, SVC can significantly improve its performance by eliminating unnecessary information through PCA. Second, Random Forest is an

ensemble model composed of multiple decision trees that predict by considering various feature combinations. Each tree is trained on a random subset of the features, which means that even if the original data contains noise, this noise tends to be automatically offset during the ensemble process. This characteristic explains why dimensionality reduction through PCA does not lead to significant performance gains. Additionally, while Random Forest captures nonlinear relationships well, each tree operates independently, so PCA's summary of overall data variance does not substantially affect the trees' splitting criteria. Consequently, Random Forest's performance tends to remain stable regardless of PCA application. Third, SVC requires the essential features of the entire dataset to create a decision boundary, and with PCA retaining only the principal features, SVC can form an even more accurate boundary. In contrast, Random Forest uses a variety of subsets to form decision trees, resulting in minimal performance differences between utilizing the entire dataset and the dimensionally reduced data from PCA. In summary, SVC can maximize its performance using data that preserves only the essential high-dimensional features, while Random Forest, due to its intrinsic mechanism of combining diverse features to offset noise, is less influenced by PCA.

5. Conclusion

This study evaluated the predictive performance of various machine learning models in forecasting startup success and analyzed the impact of Principal Component Analysis (PCA) on improving model performance. The results demonstrated that applying PCA enhanced the performance of several representative supervised learning models, including Logistic Regression, SVC (Support Vector Classifier), and XGBoost, with the most significant improvement observed in the SVC model. Specifically, SVC achieved the best results in accuracy, precision, recall, and F1 score after PCA was applied, indicating that PCA helps improve the generalization ability of the model by retaining key features from high-dimensional data.

The academic contributions of this study are twofold. First, it empirically demonstrates that combining machine learning models with preprocessing techniques, such as PCA, can improve the performance of startup success prediction. This approach, which has not been extensively explored in previous research, provides a detailed examination of how dimensionality reduction techniques impact the effectiveness of machine learning models. Second, this study highlights the role of PCA in addressing multicollinearity in high-dimensional data, offering an effective method to prevent overfitting and enhance predictive performance in the context of startup success prediction.

From a practical perspective, the findings suggest that AI-based decision-support tools can significantly improve decision-making in startup investment. Investors can leverage the SVC model with PCA to make more accurate predictions and select startups with higher potential for success based on refined forecasts. Moreover, the study emphasizes the importance of integrating real-time dynamic data processing and various dimensionality reduction techniques to enhance predictive accuracy in real-world applications.

However, this study has several limitations. First, the analysis is based on a specific dataset, which may limit the generalizability of the results to other industries or regions. Second, while PCA was the primary dimensionality reduction technique used, a comparison with other techniques, such as Linear Discriminant Analysis (LDA), was not conducted, leaving a gap in the comprehensive evaluation of dimensionality reduction methods.

Future research should focus on validating the generalizability of the model using more diverse datasets and comparing the impact of different dimensionality reduction techniques on startup success prediction. Additionally, exploring real-time dynamic startup data could lead to the development of real-time predictive models, which could be valuable for investors. Further comparison with advanced machine learning techniques, including deep learning models, could provide deeper insights into predictive performance and contribute to enhancing the accuracy of investment decision-making.

In conclusion, this study demonstrates the potential of combining machine learning models with PCA to improve the prediction of startup success, offering valuable insights for both academic research and practical applications in AI-based investment decision-making.

References

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A survey on churn analysis in various business domains. IEEE Access, 8, 220816–220839.

Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of machine learning performance for decision support in venture capital investments. IEEE Access, 7, 124233–124243.

Associates, C. (2020). US venture capital index and selected benchmark statistics.

Blaseg, D., Cumming, D., & Koetter, M. (2021). Equity crowdfunding: high-quality or low-quality entrepreneurs? Entrepreneurship Theory and Practice, 45, 505–530.

Böhm, M., Weking, J., Fortunat, F., Müller, S., Welpe, I., & Krcmar, H. (2017). The business model DNA: Towards an approach for predicting business model success.

Brem, A., Giones, F., & Werle, M. (2023). The AI digital revolution in innovation: A conceptual framework of artificial intelligence technologies for the management of innovation. IEEE Transactions on Engineering Management, 70(2), 770–776.

Castleman, R. (2020). Five ways artificial intelligence is transforming finance.

Chen, M.-C., Wang, R.-J., & Chen, A.-P. (2007). An empirical study for the detection of corporate financial anomaly using outlier mining techniques. In International Conference on Convergence Information Technology (pp. 612–617).

Gangwani, D., & Zhu, X. (2024). Modeling and prediction of business success: A survey. Artificial Intelligence Review, 57(2), 44.

Gangwani, D., Zhu, X., & Furht, B. (2023). Exploring investor-business-market interplay for business success prediction. Journal of Big Data, 10(1), 1–28.

Garber, T., Goldenberg, J., Libai, B., & Muller, E. (2004). From density to destiny: Using spatial dimension of sales data for early prediction of new product success. Marketing Science, 23(3), 419–428.

Guerzoni, M., Nava, C. R., & Nuccio, M. (2019). The survival of start-ups in time of crisis: A machine learning approach to measure innovation. arXiv preprint arXiv:1911.01073.

Guzy, M. C. (2010). Venture capital returns and public market performance (Ph.D. thesis). University of Florida.

Javaid, M., Haleem, A., & Singh, R. P. (2023). A study on ChatGPT for industry 4.0: Background, potentials, challenges, and eventualities. Journal of Economic and Technological Studies, 1, 127–143.

Jones, S., & Wang, T. (2019). Predicting private company failure: A multi-class analysis. Journal of International Financial Markets, Institutions, and Money, 61, 161–188.

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. PLoS ONE, 16(4), e0249071.

Li, H., Yu, B. X. B., Li, G., & Gao, H. (2023). Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews. Tourism Management, 96, 104707.

Martin, A., Manjula, M., & Venkatesan, D. V. P. (2011). A business intelligence model to predict bankruptcy using financial domain ontology with association rule mining algorithm. arXiv preprint arXiv:1109.1087.

McKenzie, D., David, J., & Sansone. (2017). Man vs. machine in predicting successful entrepreneurs: Evidence from a business plan competition in Nigeria.

Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. Journal of Business Venturing, 29, 1–16.

Mulcahy, D., Weeks, B., & Bradley, H. S. (2012). We have met the enemy and he is us: Lessons from twenty years of the Kauffman Foundation's investments in venture capital funds and the triumph of hope over experience. Available at SSRN 2053258.

Pasayat, A. K., Bhowmick, B., & Roy, R. (2020). Factors responsible for the success of a start-up: A meta-analytic approach. IEEE Transactions on Engineering Management, 70, 342–352.

Qu, Y., Quan, P., Lei, M., & Shi, Y. (2019). Review of bankruptcy prediction using machine learning and deep learning techniques. Procedia Computer Science, 162, 895–899.

Raj, R., Singh, A., Kumar, V., & Verma, P. (2023). Analyzing the potential benefits and use cases of ChatGPT as a tool for improving the efficiency and effectiveness of business operations. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3(3), 100140.

Ross, G., Das, S., Sciro, D., & Raza, H. (2021). CapitalVX: A machine learning model for startup selection & exit prediction. Journal of Financial Data Science, 7, 94–114.

Rtayli, N., & Enneya, N. (2019). Credit card risk detection based on feature-filter and fraud identification. In 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS) (pp. 1–8). IEEE.

Ryll, L., Barton, M. E., Zhang, B. Z., McWaters, R. J., Schizas, E., Hao, R., Bear, K., Preziuso, M., Seger, E., Wardrop, R., Rau, P. R., Debata, P., Rowan, P., Adams, N., Gray, M., & Yerolemou, N. (2020). Transforming paradigms: A global AI in financial services survey.

Shah, J. R., & Murtaza, M. B. (2000). A neural network-based clustering procedure for bankruptcy prediction. American Business Review, 18(2), 80.

Sivasankar, E., Selvi, C., & Mala, C. (2017). A study of dimensionality reduction techniques with machine learning methods for credit risk prediction. In Behera, H. S., & Mohapatra, D. P. (Eds.), Computational Intelligence in Data Mining. Springer, Singapore, pp. 65–76.

Tsai, C.-F. (2009). Feature selection in bankruptcy prediction. Knowledge-Based Systems, 22(2), 120–127.

Vroomen, P., & Desa, S. (2018). Rates of return for crowdfunding portfolios: Theoretical derivation and implications. Venture Capital, 20, 261–283.

Wang, L., & Wu, C. (2017). Business failure prediction based on two-stage selective ensemble with manifold learning algorithm & kernel-based fuzzy self-organizing map. Knowledge-Based Systems, 121, 99–110.

Wang, W., Chen, W., Zhu, K., & Wang, H. (2020). Emphasizing the entrepreneur or the idea? The impact of text content emphasis on investment decisions in crowdfunding. Decision Support Systems, 136.

Weibl, J., & Hess, T. (2019). Finding the next unicorn: When big data meets venture capital.

Zbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. Information Processing & Management, 58, 102555.

Zhou, F., Fu, L., Li, Z., & Xu, J. (2022). The recurrence of financial distress: A survival analysis. International Journal of Forecasting, 38(3), 1100–1115.