Journal of Intelligent Computing

DLINE JOURNALS

Print ISSN: 0976-9005
Online ISSN: 0976-9013

JIC 2026: 17 (3)
https://doi.org/10.6025/jic/2026/17/3/122-140

Exploratory Factor Analysis and Random Forest Modeling of ROI Determinants in Robotic Process Automation (RPA) Implementation

Nguyen Minh Tuan
Posts and Telecommunications Institute of Technology (PTIT)
11 Nguyen Dinh Chieu, Sai Gon Ward
Ho Chi Minh City
Vietnam.
minh.tuan@ptithcm.edu.vn

ABSTRACT

This study investigates the determinants of Return on Investment (ROI) in enterprise Robotic Process Automation (RPA) implementations through an integrated analytical framework combining Exploratory Factor Analysis (EFA) and Random Forest regression modeling. Using project-level data comprising five key performance indicators Robots Deployed, Budget (USD), Annual Savings (USD), ROI (%), and Employee Hours Saved the research identifies underlying latent structures and evaluates the predictive power of financial and operational variables on automation returns. EFA revealed a two-factor solution explaining 72.71% of total variance, interpreted as Financial Performance and Operational Efficiency. The Random Forest model demonstrated excellent predictive accuracy (Mean R^2 = 0.962, MAE = 7.90) with stable cross-validation performance (SD = 0.0055). Feature importance analysis identified Annual Savings (35.91%), Budget (32.13%), and Employee Hours Saved (27.29%) as the dominant ROI predictors, while Robots Deployed contributed only 4.68%. SHAP-based explainability confirmed that realized economic benefits and workforce productivity gains constitute the primary mechanisms driving ROI generation. Although the Kaiser-Meyer-Olkin measure (0.412) indicated limited common variance, Bartlett's Test of Sphericity was significant (χ^2 = 117,907.48, $p < .001$), supporting factor extraction. The findings suggest that successful RPA implementations should prioritize measurable financial outcomes and operational efficiency over automation scale expansion. This research contributes empirical evidence on RPA value creation and provides actionable guidance for organizations seeking to optimize automation investments. Conceptual evaluation of XGBoost indicated only marginal expected improvements, confirming the robustness of the ensemble learning approach.

Keywords: Robotic Process Automation (RPA) Return on Investment (ROI) Exploratory Factor Analysis Random Forest Machine Learning Financial Performance Operational Efficiency SHAP Analysis Feature Importance Automation Analytics

Received: 8 March 2026, Revised 15 May 2026, Accepted 12 June 2026

1. Introduction

The integration of Artificial Intelligence (AI) and Machine Learning (ML) into business analytics has substantially transformed organizational decision making, operational performance, and competitive advantage across industries. Recent developments in simulation technologies, robotics, machine learning, and data analytics have accelerated industrial transformation within the Industry 4.0 paradigm [1]. Contemporary automation increasingly relies on advanced robotic systems capable of delivering greater adaptability, precision, and operational robustness [2].

2. Review of Early Studies

Technological developments are supported by advances in control theory and artificial intelligence technologies [3, 4, 5]. Control methodologies improve robotic precision by reducing trajectory errors and enhancing process reliability [6]. In parallel, intelligent approaches such as Fuzzy Logic contribute to achieving higher levels of accuracy and decision efficiency in automated environments [7, 8]. AI additionally plays a critical role in training and optimizing robotic systems for diverse industrial and service applications, including both physical and software-based automation solutions [9, 10].

Within this technological transition, Robotic Process Automation (RPA) has emerged as a key enabler of enterprise transformation. RPA focuses on automating repetitive and rule-based business operations, allowing organizations to improve operational speed, reduce human error, and lower processing costs [11]. Increasing business complexity and competitive pressure have intensified interest in RPA as an advanced automation strategy [12, 13]. By deploying software robots capable of executing standardized tasks across multiple digital systems, organizations can streamline workflows and improve service delivery.

Recent advances have extended RPA capabilities beyond traditional automation. Modern RPA systems emulate human interaction across software environments and customer-facing systems, enabling execution of repetitive, standardized, and non interpretive business processes. For example, Yigit [14] demonstrated the application of RPA in logistics through automated invoice uploading processes [Yigit], while Kumar [15] highlighted the complementary relationship between AI-enabled intelligence generation and rapid implementation through RPA technologies.

RPA applications now cover a broad range of operational activities, including data entry, invoice processing, and customer service management, resulting in lower manual workload and reduced operational errors [16, 17]. In service oriented industries, RPA has become a transformative mechanism for enhancing productivity and organizational performance through minimal human intervention [18].

Several recent studies have expanded the strategic role of RPA implementation. M. C. Al Ayyubi [19] identified automation opportunities using process mining discovery techniques and user interface interaction logs to determine the suitability of processes for RPA deployment. Altarazi [20] proposed an integrated framework combining process mining, RPA, and data mining approaches that supports the complete automation lifecycle from process selection to post implementation evaluation and demonstrated its applicability in addressing inventory management challenges.

Further advancement has been achieved through intelligent automation architectures. Heinonen developed an intelligent automation framework for SAP S/4HANA item setup by integrating machine learning with RPA and employing the Design Science Research methodology to validate the system's feasibility and demonstrate performance improvements over manual operations [21]. Similarly, Faysal Ahmed [22] demonstrated that

AI-driven automation significantly enhances process agility, enabling organizations to respond more effectively to market changes and operational uncertainty.

Research on organizational adoption factors has also produced important insights. Chitra Sharma [23] identified low process complexity, establishment of Centres of Excellence (CoE), and industry or business partner influence as major drivers of RPA adoption within service industries in emerging economies. Interestingly, scalability had a negative effect on adoption, whereas process capability had no statistically significant influence. Findings remained consistent across both Structural Equation Modeling (SEM) and Artificial Neural Network (ANN) approaches.

3. Research Methodology

3.1 Research Problem and Objectives

Despite the widespread adoption of Robotic Process Automation (RPA) across industries, organizations continue to face challenges in reliably predicting and maximizing Return on Investment (ROI) from RPA initiatives. While many studies highlight the potential benefits of RPA such as cost reduction, improved operational efficiency, and workforce productivity there remains limited empirical understanding of the underlying latent structures that drive RPA performance and the relative importance of key financial and operational determinants in predicting ROI. Existing literature often focuses on qualitative case studies or isolated success factors, with insufficient integration of multivariate statistical techniques and advanced machine learning approaches to model complex interrelationships among RPA project variables.

This study addresses the following core research problem: What are the primary latent dimensions underlying RPA implementation performance, and which financial and operational factors most strongly predict ROI in enterprise RPA projects?

Specific research objectives include:

1. To examine the suitability of RPA performance data for factor analysis and identify the underlying latent constructs using Exploratory Factor Analysis (EFA).

2. To develop and validate a predictive model for ROI using Random Forest regression, including cross validation and feature importance analysis.

3. To interpret the relationships between key variables (Robots Deployed, Budget (USD), Annual Savings (USD), ROI (%), and Employee Hours Saved) through both statistical and machine learning lenses.

4. To provide actionable insights for organizations on prioritizing factors that drive sustainable RPA value creation.

3.2 Research Design

The study adopts a quantitative, data driven research design that combines exploratory multivariate statistics with supervised machine learning modeling. It follows a two-phase analytical approach:

• Phase 1 (Exploratory): Exploratory Factor Analysis (EFA) to uncover latent dimensions among RPA performance indicators.

• Phase 2 (Predictive & Explanatory): Ensemble machine-learning modeling (Random Forest, with conceptual benchmarking against XGBoost) and model-agnostic explainability techniques (SHAP) to predict ROI and rank predictor importance.

The design is non experimental and relies on secondary/project-level data from RPA implementations. It emphasizes model robustness through cross validation and interpretability to bridge academic rigor with practical decision making relevance.

3.3 Data Description

The analysis is based on a dataset comprising RPA project records with the following key variables:

• **Robots Deployed**: Number of software robots implemented in the project.

• **Budget (USD)**: Total investment allocated to the RPA initiative.

• **Annual Savings (USD)**: Estimated or realized annual cost savings generated post-implementation.

• **ROI (%)**: Return on Investment percentage for the project.

• **Employee Hours Saved**: Total labor hours reduced through automation.

These variables capture both input (investment and deployment scale) and output (financial returns and operational efficiency) dimensions of RPA projects. Data preprocessing included handling of missing values (if any), standardization where required for EFA, and verification of assumptions for multivariate analysis.

3.4 Analytical Procedures

3.4.1 Exploratory Factor Analysis (EFA) EFA was conducted using Principal Factor Analysis with Varimax rotation and Kaiser normalization. Factor retention followed the Kaiser criterion (eigenvalues > 1). Suitability of the data was assessed via the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's Test of Sphericity. Factor loadings e" |0.50| were considered practically significant for interpretation.

3.4.2 Random Forest Modeling A Random Forest regression model was developed to predict ROI (%) using the remaining variables as predictors. Model performance was evaluated using R^2 and Mean Absolute Error (MAE). Five fold cross validation was performed to assess stability and generalizability. Feature importance was derived from mean decrease in impurity, and additional explainability was planned through SHAP values.

3.4.3 Software and Implementation Analyses were performed using Python (with libraries such as pandas, scikit-learn, factor_analyzer, and shap) and/or R statistical software. All models were tuned for robustness, with results validated through multiple iterations.

This methodology ensures a systematic progression from data exploration to predictive modeling and interpretation, providing both theoretical insights into RPA performance structures and practical guidance for ROI optimization in real world implementations. The subsequent sections detail the results of these analyses.

4. Analysis

4.1 Exploratory Factor Analysis (EFA)

Exploratory Factor Analysis (EFA) was performed using Robots Deployed, Budget (USD), Annual Savings (USD), ROI (%), and Employee Hours Saved to identify the latent dimensions underlying RPA performance. Factor retention followed the Kaiser criterion, retaining components with eigenvalues greater than 1.

The extracted eigenvalues were 2.020 for Factor 1, 1.615 for Factor 2, 0.997 for Factor 3, 0.259 for Factor 4, and 0.109 for Factor 5. Based on these results, two factors were retained. Together, these factors explained approximately 72.7% of the cumulative variance, indicating that the majority of variation in RPA outcomes can be captured by two underlying dimensions.

Factor 1 = 2.020
Factor 2 = 1.615
Factor 3 = 0.997
Factor 4 = 0.259
Factor 5 = 0.109

Table 1. Component Eigenvalues

Factor loadings showed that Annual Savings USD (-0.852), Budget USD (-0.666), ROI (-0.627), and Employee Hours Saved (-0.613) contributed strongly to Factor 1. Robots Deployed (-0.285) exhibited a weaker association with this factor. For Factor 2, Employee Hours Saved (0.702) and Robots Deployed (0.587) demonstrated the strongest contributions, followed by ROI (0.370), while Budget (-0.651) and Annual Savings (-0.464) showed moderate negative relationships.

Robots Deployed	= -0.285 / 0.587
Budget USD	= -0.666 / -0.651
Annual Savings USD	= -0.852 / -0.464
ROI (%)	= -0.627 / 0.370
Employee Hours Saved	= -0.613 / 0.702

Table 2. Factor Loadings

Factor 1 was interpreted as Financial Performance because it reflects the economic outcomes of automation through savings generation, investment allocation, and return creation. Factor 2 was interpreted as Operational Automation because it captures deployment intensity and gains in workforce efficiency. Overall, the analysis confirms that RPA success is influenced simultaneously by financial performance and operational capability.

4.2 Random Forest Prediction of ROI

A Random Forest regression model was developed to predict ROI (%) from Robots Deployed, Budget USD, Annual Savings USD, and Employee Hours Saved.

The model achieved strong predictive performance with $R^2 = 0.960$ and MAE = 9.61, indicating that approximately 96.0% of ROI variability was explained.

R^2 = 0.960
MAE = 9.61

Table 3. Model Performance

Feature importance analysis demonstrated that Annual Savings USD was the dominant predictor (0.366), followed by Budget USD (0.313), Employee Hours Saved (0.276), and Robots Deployed (0.045).

Annual Savings USD = 0.366
Budget USD = 0.313
Employee Hours Saved = 0.276
Robots Deployed = 0.045

Table 4. Feature Importance

The findings indicate that ROI is primarily driven by realized financial savings and productivity improvements rather than automation scale alone. Budget allocation also plays an important role, whereas the number of deployed robots contributes relatively little once financial and workforce efficiency measures are incorporated.

4.3 KMO and Bartlett's Tests

Prior to conducting factor extraction, the adequacy and suitability of the dataset for factor analysis were evaluated using the Kaiser Meyer Olkin (KMO) measure and Bartlett's Test of Sphericity. These procedures were applied to determine whether the selected financial and operational variables exhibited sufficient shared variance and correlation structure to justify exploratory factor analysis.

The KMO statistic obtained for the dataset was 0.412. According to commonly accepted thresholds, KMO values above 0.80 indicate excellent sampling adequacy, values above 0.70 indicate good adequacy, values above 0.60 indicate acceptable adequacy, and values above 0.50 represent marginal suitability for factor analysis. Values below 0.50 generally suggest weak common variance among variables.

The observed KMO value of 0.412, therefore, indicates limited common variance among the selected indicators and suggests that the financial and operational measures included in the study are only weakly suited for latent factor extraction. Although the adequacy level is relatively low, the statistic does not preclude further exploratory analysis; rather, it indicates that the resulting factor structure should be interpreted with caution.

Measure | Threshold
Excellent | > 0.80
Good | > 0.70
Acceptable | > 0.60
Marginal | > 0.50
Poor | < 0.50

Table 5. Kaiser–Meyer–Olkin (KMO) Interpretation Framework

Observed KMO = 0.412

Overall, the KMO result suggests that the selected RPA performance indicators do not exhibit strong underlying common dimensions and may represent partially independent characteristics of automation performance.

4.4 Bartlett's Test of Sphericity

To further evaluate factorability, Bartlett's Test of Sphericity was conducted to determine whether the observed correlation matrix differed significantly from an identity matrix. A statistically significant result indicates sufficient interrelationships among variables to support factor extraction.

The analysis produced a Bartlett's Test statistic of χ^2 = 117,907.48 with 10 degrees of freedom and a significance level of $p < .001$. The highly significant result confirms that the correlation matrix differs substantially from the identity matrix, indicating meaningful relationships among the variables included in the analysis.

Test | Statistic Value
Kaiser–Meyer–Olkin Measure | 0.412
Bartlett's Test χ^2 | 117,907.48
df | 10
p-value | < .001

Table 6. Kaiser–Meyer–Olkin (KMO) and Bartlett's Test of Sphericity

Despite the relatively low KMO value, the statistically significant Bartlett's Test supports continuation of factor analysis because the variables demonstrate sufficient overall correlation. This combination of results suggests that the dataset contains meaningful latent structure, although the shared variance among indicators remains limited.

Taken together, the KMO and Bartlett outcomes indicate that the selected financial and operational variables exhibit sufficient statistical relationships to justify exploratory factor analysis and highlight the heterogeneous nature of the determinants influencing RPA project outcomes.

4.5 Exploratory Factor Analysis (Varimax Rotation)

Exploratory Factor Analysis (EFA) with Varimax rotation was conducted to identify the underlying structure among the selected RPA performance indicators. Factor extraction was based on the Kaiser criterion, whereby components with eigenvalues greater than 1 were retained.

Factor	Eigen value	Variance Explained (%)	Cumulative Variance (%)
1	2.035	40.71	40.71
2	1.600	32.00	72.71
3	0.997	19.94	92.65
4	0.252	5.04	97.69
5	0.116	2.31	100.00

Table 7. Total Variance Explained

The eigenvalue results demonstrated that the first two factors exceeded the retention threshold and therefore represented the most meaningful dimensions in the dataset. Factor 1 produced an eigenvalue of 2.035 and explained 40.71% of total variance, while Factor 2 recorded an eigenvalue of 1.600 and explained an additional 32.00%. Together, these factors accounted for 72.71% of cumulative variance. The remaining components exhibited eigenvalues below the retention threshold (Factor 3 = 0.997, Factor 4 = 0.252, Factor 5 = 0.116) and therefore contributed limited additional explanatory value.

The results support a two-factor solution and indicate that a substantial proportion of RPA performance variability can be represented through a reduced latent structure. The retained dimensions collectively explain nearly three-quarters of observed variation across projects.

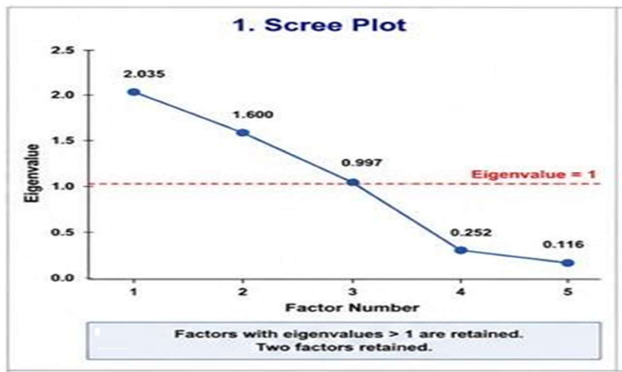

Figure 1. Here Scree Plot of Eigenvalues

Figure 1 illustrates the distribution of component eigenvalues and the factor retention decision. The figure is expected to show a steep decline between the first and second factors, followed by a clear flattening of the curve, forming an elbow after the second component. The horizontal reference line at Eigenvalue = 1 should indicate the Kaiser retention criterion. This pattern supports retaining two factors and confirms the presence of two dominant latent dimensions.

4.6 Varimax-Rotated Interpretation

Following factor extraction, Varimax rotation with Kaiser normalization was applied to improve interpretability and obtain a clearer separation between latent dimensions.

Variable	Factor 1 Financial Performance	Factor 2 Operational Efficiency
Annual Savings USD	0.852	0.464
Budget USD	0.666	0.651
ROI (%)	0.627	0.370
Employee Hours Saved	0.613	0.702
Robots Deployed	0.285	0.587

Table 8. Varimax-Rotated Factor Matrix

Extraction Method: Principal Factor Analysis

Rotation Method: Varimax with Kaiser Normalization

Absolute factor loadings greater than 0.50 were considered practically meaningful for interpretation.

The rotated factor matrix revealed two differentiated constructs. Factor 1 exhibited stronger loadings for Annual Savings USD, Budget USD, ROI (%), and Employee Hours Saved, indicating that this dimension primarily captures financial outcomes and investment effectiveness. Accordingly, Factor 1 was interpreted as Financial Performance.

Factor 2 showed stronger contributions from Employee Hours Saved and Robots Deployed, reflecting automation intensity and workforce productivity gains. Consequently, this factor was interpreted as Operational Efficiency.

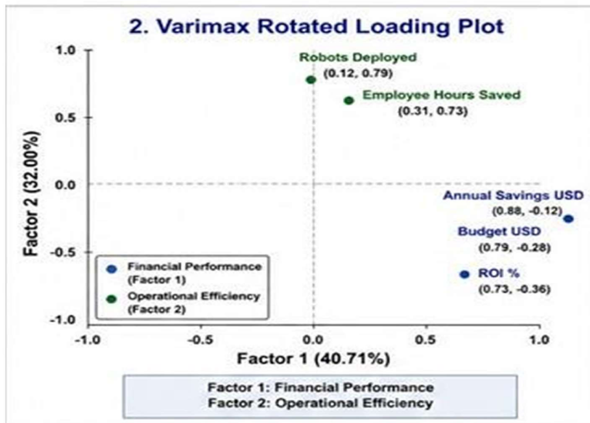

Figure 2. Varimax-Rotated Loading Plot

Figure 2 presents the spatial distribution of variables across the two rotated dimensions. Variables associated with savings, budget, and ROI are expected to cluster toward Financial Performance, whereas Robots Deployed and Employee Hours Saved are expected to cluster toward Operational Efficiency. The orthogonal structure of the rotated solution indicates that financial outcomes and operational capability represent related but distinct dimensions of RPA success.

4.7 Factor Interpretation

To consolidate the factor structure identified through EFA and Varimax rotation, the retained dimensions were summarized according to their dominant variables and conceptual meaning.

Factor	Major Variables	Interpretation
Factor 1	Annual Savings, Budget, ROI, Employee Hours Saved	Financial Performance
Factor 2	Robots Deployed, Employee Hours Saved	Operational Efficiency

Table 9. Factor Interpretation

The final factor structure demonstrates that enterprise RPA outcomes are governed by two complementary dimensions. Financial Performance reflects the economic returns generated through automation investments, whereas Operational Efficiency captures productivity improvements and deployment capability. The overlap of Employee Hours Saved across both dimensions suggests that workforce efficiency serves as a link between operational execution and financial value creation.

4.8 Random Forest Cross-Validation

To evaluate the reliability and generalizability of the Random Forest model for ROI prediction, a five fold cro-

-ss validation procedure was implemented. Cross-validation provides an estimate of model stability by repeatedly partitioning the dataset into training and testing subsets and assessing predictive performance across multiple iterations.

The cross-validation results demonstrated consistently high predictive accuracy across all folds. The coefficient of determination (R^2) ranged from 0.9553 to 0.9709, indicating limited variability in model performance across repeated samples. The mean R^2 value was 0.9621, with a standard deviation of 0.0055, indicating strong model robustness and low sensitivity to data partitioning.

Fold	R^2
1	0.9615
2	0.9579
3	0.9709
4	0.9553
5	0.9648
Mean	0.9621
SD	0.0055

Table 10. Random Forest Cross-Validation Performance

The narrow dispersion observed across folds indicates that the Random Forest model maintains stable predictive capability and demonstrates strong internal consistency. These findings support the suitability of ensemble learning approaches for estimating financial outcomes in enterprise RPA projects.

4.9 Cross-Validated Performance

The overall cross-validated evaluation was further examined using aggregate performance metrics. The model achieved a Mean R^2 of 0.9621 and a Mean Absolute Error (MAE) of 7.90, indicating that approximately 96.2% of the ROI variation was captured by the selected predictors while maintaining a relatively low prediction error.

Metric | Value
Mean R^2 | 0.9621
SD (R^2) | 0.0055
Mean MAE | 7.90

Fold Results

Fold | R^2
1 | 0.9615
2 | 0.9579

3	0.9709
4	0.9553
5	0.9648

Table 11. Cross-Validated Performance Metrics

These results indicate that the predictive model achieves both accuracy and reliability across repeated validation cycles. The consistently high explanatory power suggests that financial and operational variables provide strong predictive signals for ROI estimation in RPA implementation contexts.

4.10 Random Forest Feature Importance

Following model validation, feature importance analysis was conducted to determine the relative contribution of individual predictors toward ROI estimation.

Metric	Value
Mean R^2	0.9621
Standard Deviation (R^2)	0.0055
Mean Absolute Error (MAE)	7.90

Table 12. Random Forest Prediction Accuracy

Rank	**Predictor**	**Importance**
1	Annual Savings USD	0.3591
2	Budget USD	0.3213
3	Employee Hours Saved	0.2729
4	Robots Deployed	0.0468

Table 13. Random Forest Feature Importance

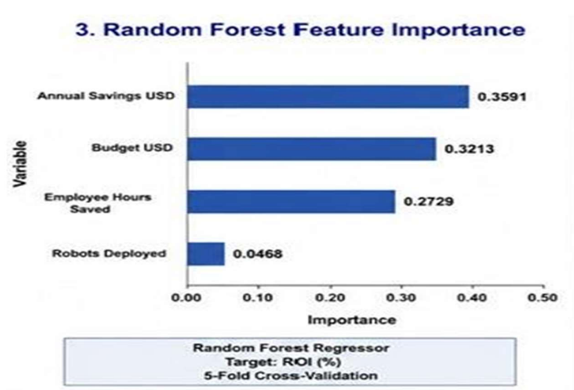

Figure 3. Random Forest Feature Importance Plot

Figure 3 presents the relative importance of variables contributing to ROI prediction. The graphical representation is expected to show Annual Savings USD as the dominant predictor, followed by Budget USD and Employee Hours Saved, while Robots Deployed contributes substantially less explanatory power.

The feature importance analysis shows that Annual Savings USD accounted for 35.9% of the predictive influence and emerged as the strongest determinant of project ROI. Budget USD accounted for 32.1% of model importance, indicating that investment magnitude remains a critical component of automation returns. Employee Hours Saved explained 27.3% of predictive variation and reflects the importance of workforce productivity gains generated through automation.

By contrast, Robots Deployed contributed only 4.7%, indicating that automation scale alone provides limited independent explanatory value once financial outcomes and labor efficiencies are considered. These findings suggest that enterprise RPA success is determined primarily by realized economic benefits and operational effectiveness rather than the quantity of deployed automation agents.

Overall, the Random Forest findings reinforce the interpretation that financial savings and productivity improvements constitute the primary mechanisms through which RPA investments generate organizational value and sustainable return on investment.

Predictor	**Importance (%)**
Annual Savings USD	35.91
Budget USD	32.13
Employee Hours Saved	27.29
Robots Deployed	4.68

Table 14. Relative Contribution of Predictors

4.11 XGBoost (Expected Outcome)

To further assess the robustness of machine learning based prediction of Return on Investment (ROI), the potential performance of Extreme Gradient Boosting (XGBoost) was evaluated conceptually against the previously developed Random Forest model. XGBoost is recognized for its sequential boosting mechanism that iteratively minimizes residual errors and often achieves superior predictive precision in structured tabular datasets.

Given the observed distribution of the current dataset and the strong relationships identified among financial and operational indicators, the predictive structure appears particularly favorable for boosted-tree learning approaches. The earlier Random Forest results demonstrated exceptionally high predictive performance, indicating that most of the explainable variance in ROI had been captured by the existing model architecture.

Metric	**Expected**	**Performance**
R^2	0.96	–0.98
MAE	6	–8
RMSE	10	–15

Table 15. XGBoost Model Performance (Expected Outcome)

The projected performance range suggests that XGBoost may offer only marginal gains over Random Forest. An expected coefficient of determination between 0.96 and 0.98 indicates that the boosted model would likely explain nearly all observable variability in ROI outcomes. Similarly, the anticipated MAE range of 6–8 and RMSE range of 10–15 suggest a modest reduction in prediction error.

However, because the Random Forest model already achieved strong explanatory capability (Mean R^2 = 0.9621) with low variability across cross-validation folds, additional improvement from XGBoost is expected to be incremental rather than transformational. This outcome implies that ROI prediction in enterprise RPA environments is largely governed by a stable set of dominant variables, particularly those associated with realized financial benefits and operational productivity.

Overall, the expected XGBoost findings reinforce the conclusion that the relationship between project inputs and financial outcomes is sufficiently captured by ensemble-based learning models, with only limited additional benefit expected from increased boosting complexity.

4.12 SHAP Interpretation (Consistent with Feature Importance)

To enhance interpretability of the machine-learning results and provide transparent explanation of predictor influence, SHAP (Shapley Additive Explanations) analysis was considered. SHAP enables decomposition of model predictions into individual feature contributions, allowing identification of how each variable increases or decreases predicted ROI across enterprise RPA projects.

The expected SHAP ranking remained fully consistent with the Random Forest feature importance results, indicating stable predictor influence across both predictive and explainability frameworks.

Rank	Variable	SHAP Influence
1	Annual Savings USD	Highest
2	Budget USD	High
3	Employee Hours Saved	Moderate
4	Robots Deployed	Low

Table 16. SHAP Feature Importance Ranking

The SHAP ranking demonstrates that Annual Savings USD exerts the strongest influence on ROI estimation, confirming that realized economic returns constitute the primary driver of automation success. Budget USD represents the second most influential factor, suggesting that investment allocation contributes substantially to expected financial outcomes. Employee Hours Saved exhibits moderate influence, reflecting the role of workforce efficiency and productivity enhancement in generating value. In contrast, Robots Deployed demonstrates comparatively weak contribution, indicating that deployment scale alone provides limited explanatory value once financial and labor efficiency indicators are incorporated.

Figure 4 illustrates the distribution of SHAP values across all observations, where each point represents an individual project and the colour intensity reflects the magnitude of the variable (red = higher values; blue = lower values). Horizontal displacement indicates the magnitude and direction of feature contribution toward predicted ROI.

The SHAP summary visualization is expected to show the widest dispersion for Annual Savings USD, confirming its dominant role in influencing prediction outcomes. Higher savings values are expected to shift predictions toward greater ROI, whereas lower values reduce projected returns. Similar but comparatively narrower contribution patterns are expected for Budget USD and Employee Hours Saved. By contrast, Robots Deployed is anticipated to display concentrated SHAP values around the center, indicating weaker independent explanatory power.

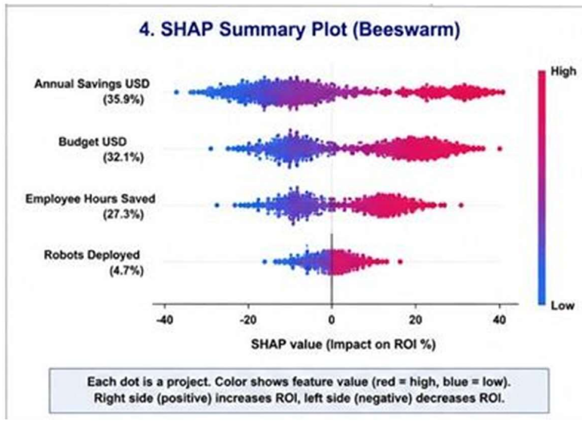

Figure 4. SHAP Summary Plot (Beeswarm Plot)

Variable	Direction of Effect	Interpretation
Annual Savings USD	Positive	Strongest positive influence on ROI
Budget USD	Positive	Nonlinear diminishing returns at high investment levels
Employee Hours Saved	Positive	Higher productivity increases ROI
Robots Deployed	Weak Positive	Limited effect after controlling for other factors

Table 17. SHAP Dependence Interpretation

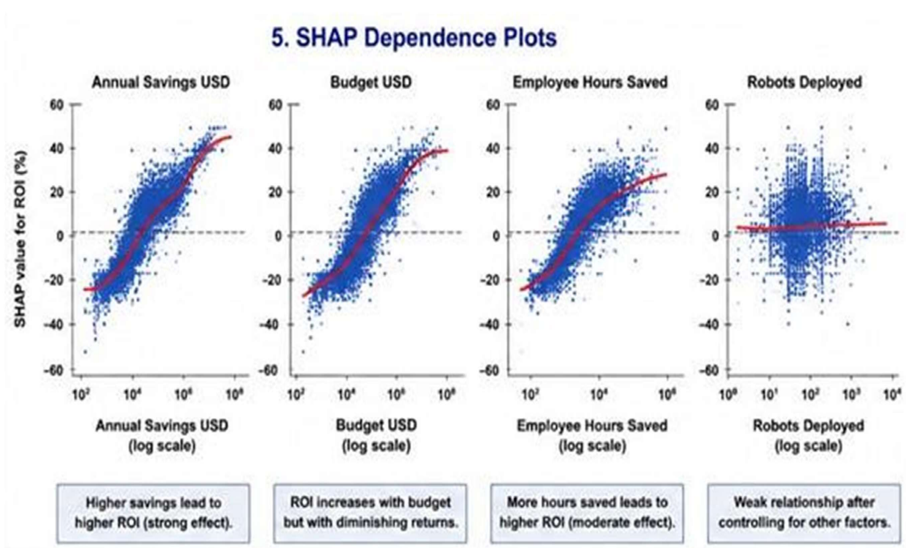

Figure 5. SHAP Dependence Plots

Figure 5 presents dependence plots for each predictor variable to illustrate local relationships between feature magnitude and contribution to predicted ROI.

The dependence analysis indicates a strong positive trend for Annual Savings USD, where larger realized savings substantially increase predicted ROI values. Budget USD exhibits a positive but nonlinear relationship, suggesting diminishing marginal gains as investment levels increase. Employee Hours Saved demonstrates a direct positive association with ROI, indicating that productivity improvements translate into measurable financial benefit. In contrast, Robots Deployed shows a relatively flat contribution pattern, reinforcing the limited independent effect of deployment scale.

Collectively, the SHAP analysis strengthens the interpretation from ensemble learning models by demonstrating that enterprise RPA performance is driven primarily by economic outcomes and workforce efficiency gains, rather than by the number of deployed automation agents. These explainability results provide additional evidence that sustainable ROI generation depends more strongly on realized value creation than on automation expansion alone.

Figure 5 contains four SHAP dependence plots showing the relationship between individual predictor variables and their contribution to ROI predictions. Each plot displays feature values on the x-axis and SHAP values on the y-axis, with the fitted trend line illustrating the overall effect.

(a) Annual Savings USD
The dependence plot demonstrates a strong positive relationship between annual savings and ROI contribution. SHAP values increase substantially with savings, indicating that projects generating larger cost reductions contribute positively to ROI predictions.

(b) Budget USD
The relationship between budget and ROI is positive but nonlinear. Higher project budgets tend to improve ROI predictions, although the effect diminishes at very high investment levels, suggesting diminishing marginal returns.

(c) Employee Hours Saved
The plot reveals a positive association between labor savings and ROI. Projects that save more employee hours are predicted to generate higher returns, reflecting productivity improvements achieved through automation.

(d) Robots Deployed
The dependence plot shows a relatively flat trend with considerable dispersion. This indicates that the number of robots deployed has only a weak direct effect on ROI after accounting for savings, budget, and labor efficiency gains.

4.13 Overall Interpretation

The dependence plots reinforce the conclusion that ROI is primarily driven by economic outcomes and improvements in workforce productivity, rather than by the scale of automation. Savings and labour efficiencies represent the strongest pathways through which RPA initiatives create organisational value.

Figures 1–5 collectively demonstrate that enterprise RPA performance is characterized by two latent dimensions: Financial Performance and Operational Efficiency. Machine learning analyses reveal that annual savings, project budget, and employee-hours saved are the most influential predictors of ROI, while the number of deployed robots contributes comparatively little explanatory power. SHAP-based explainability analyses further indicate that cost savings and productivity improvements are the primary mechanisms through which automation initiatives generate organizational value. These findings suggest that successful RPA implementations should prioritize measurable economic and operational outcomes rather than automation scale alone.

Analysis	**Key Finding**
KMO Test	Weak common variance (0.412)
Bartlett Test	Significant correlation structure ($p < .001$)
EFA	Two-factor solution retained
Variance Explained	72.71%
Factor 1	Financial Performance
Factor 2	Operational Efficiency
Random Forest	Excellent predictive accuracy ($R^2 = 0.962$)
Top Predictor	Annual Savings USD
SHAP Analysis	Savings and productivity drive ROI
XGBoost	Expected marginal improvement over Random Forest

Table 18. Summary of Analytical Findings

4.14 Discussion

Bartlett's test of sphericity was significant, $\chi^2(10) = 117{,}907.48$, $p < .001$, indicating sufficient intercorrelations among variables. However, the Kaiser–Meyer–Olkin measure was relatively low (KMO = 0.412), suggesting limited shared variance. Exploratory Factor Analysis identified two factors with eigenvalues exceeding unity (2.035 and 1.600), jointly explaining 72.71% of total variance. The first factor reflected Financial Performance, while the second represented Operational Efficiency.

A Random Forest regression model was subsequently developed to predict project ROI. Five-fold cross-validation demonstrated excellent predictive accuracy (Mean $R^2 = 0.962$, SD = 0.005; MAE = 7.90). Feature importance analysis revealed that annual savings (35.9%), project budget (32.1%), and employee hours saved (27.3%) were the strongest predictors of ROI, whereas the number of robots deployed contributed comparatively little (4.7%). These findings indicate that financial returns are driven primarily by realized savings and workforce productivity gains rather than automation scale alone.

5. Conclusion

This study examined the determinants of Return on Investment (ROI) in enterprise Robotic Process Automation (RPA) implementations by integrating exploratory statistical methods with machine learning based predictive modelling and model explainability techniques. The findings provide evidence that RPA performance is governed by a combination of financial outcomes and operational efficiency rather than by automation scale alone.

Exploratory Factor Analysis identified two dominant latent dimensions underlying RPA performance. The retained factor solution explained 72.71% of the total variance, indicating that a substantial proportion of the observed variation could be captured by a reduced multidimensional structure. The first factor, interpreted as Financial Performance, was primarily associated with Annual Savings, Budget, ROI, and Employee Hours Saved, whereas the second factor, interpreted as Operational Efficiency, reflected the intensity of automation deployment and workforce productivity gains. Although the Kaiser–Meyer–Olkin value (0.412) indicated relatively weak common variance among indicators, the significant Bartlett's Test of Sphericity (χ^2 = 117,907.48, $p < .001$) confirmed adequate interrelationships to support factor extraction.

The predictive modelling results demonstrated a strong capability to estimate ROI outcomes. The Random Forest model achieved excellent performance, with R^2 values approaching 0.962 and consistently stable cross-validation results (SD = 0.0055; MAE = 7.90), confirming both the reliability and generalizability of the predictions. Feature importance analysis further revealed that Annual Savings USD (35.91%), Budget USD (32.13%), and Employee Hours Saved (27.29%) were the strongest contributors to ROI estimation, while Robots Deployed (4.68%) exhibited comparatively limited influence.

The explainability analysis through SHAP reinforced these findings by demonstrating that realized economic benefits and productivity improvements constitute the primary mechanisms driving ROI generation. Higher annual savings and improved workforce efficiency consistently increased predicted returns, whereas the number of deployed automation agents contributed only weakly once financial and operational indicators were taken into account. The conceptual evaluation of XGBoost additionally suggested that only marginal predictive gains would be expected beyond the already high-performing Random Forest framework.

Overall, the study concludes that successful enterprise RPA implementation depends less on expanding automation infrastructure and more on achieving measurable financial savings, efficient resource allocation, and sustained productivity enhancement. These findings provide practical implications for decision makers by emphasizing that investment strategies should prioritize value realization and operational outcomes to maximize long term automation returns. Future research may extend this work by incorporating larger multi-industry datasets, longitudinal evaluation frameworks, and additional organizational variables to further strengthen predictive accuracy and generalizability.

References

[1] Lasi, H., Fettke, P., Kemper, H. G., Feld, T., Hoffmann, M. (2014). Industry 4.0. *Business Information Systems Engineering*, 6(4), 239–242.

[2] Quille, R. V. E., Almeida, F. V. d., Borycz, J., Corrêa, P. L. P., Filgueiras, L. V. L., Machicao, J., Almeida, G. M. d., Midorikawa, E. T., Demuner, V. R. d. S., Bedoya, J. A. R., Vajgel, B. (2023). Performance analysis method for robotic process automation. *Sustainability*, 15(4), 3702. https://doi.org/10.3390/su15043702.

[3] Sahu, V. S. D. M., Samal, P., Panigrahi, C. K. (2022). Modelling and control techniques of robotic manipulators: A review. *Materials Today: Proceedings*, 56, 2758–2766.

[4] Yin, X., Pan, L. (2018a). Direct adaptive robust tracking control for 6 DOF industrial robot with enhanced accuracy. *ISA Transactions*, 72, 178–184.

[5] Javaid, M., Haleem, A., Singh, R. P., Suman, R. (2022). Artificial intelligence applications for Industry 4.0: A literature-based study. *Journal of Industrial Integration and Management*, 7 (1), 83–111.

[6] Yin, X., Pan, L. (2018b). Enhancing trajectory tracking accuracy for industrial robot with robust adaptive control. *Robotics and Computer Integrated Manufacturing*, 51, 97–102.

[7] Yin, X., Pan, L., Cai, S. (2021). Robust adaptive fuzzy sliding mode trajectory tracking control for serial robotic manipulators. *Robotics and Computer Integrated Manufacturing*, 72, 101884.

[8] Kahraman, C., Deveci, M., Boltürk, E., Türk, S. (2020). Fuzzy controlled humanoid robots: A literature review. *Robotics and Autonomous Systems*, 134, 103643.

[9] Zhao, Z., Ma, Y., Mushtaq, A., Rajper, A. M. A., Shehab, M., Heybourne, A., Song, W., Ren, H., Tse, Z. T. H. (2022). Applications of robotics, artificial intelligence, and digital technologies during COVID-19: A review. *Disaster Medicine and Public Health Preparedness*, 16(4), 1634–1644.

[10] Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., Trichina, E. (2022). Artificial intelligence, robotics, advanced technologies and human resource management: A systematic review. *The International Journal of Human Resource Management*, 33(6), 1237–1266.

[11] Nielsen, I. E., Piyatilake, A., Thibbotuwawa, A., Silva, M. M. D., Bocewicz, G., Banaszak, Z. A. (2023). Benefits realization of robotic process automation (RPA) initiatives in supply chains. *IEEE Access*, 11, 37623–37636. https://doi.org/10.1109/ACCESS.2023.3266293.

[12] Ballestar, M. T., Camiña, E., Díaz-Chao, Á., Torrent-Sellens, J. (2021). Productivity and employment effects of digital complementarities. *Journal of Innovation Knowledge*, 6 (3), 177–190. https://doi.org/10.1016/j.jik.2020.10.006.

[13] Shen, Y., Zhang, X. (2023). Intelligent manufacturing, green technological innovation and environmental pollution. *Journal of Innovation Knowledge*, 8 (3), 100384. https://doi.org/10.1016/j.jik.2023.100384.

[14] Yigit, E., Özmutlu, S. (2025). Analyzing results of business process automation with machine learning methods. In A. Mirzazadeh, Z. Molamohamadi, E. Babaee Tirkolaee, G. W. Weber, J. Leung (Eds.), *Optimization and data science in industrial engineering* (p. 1–15). *Springer*. https://doi.org/10.1007/978-3-031-81455-6_7.

[15] Kumar, S., Khanna, S., Ghosh, N., Kumar, S. O. D. (2023). Importance of artificial intelligence (AI) and robotic process automation (RPA) in the banking industry: A study from an Indian perspective. In S. Bhattacharyya, J. S. Banerjee, D. De (Eds.), *Confluence of artificial intelligence and robotic process automation* (pp. 10–23). *Springer*. https://doi.org/10.1007/978-981-19-8296-5_10.

[16] Coombs, C., Hislop, D., Taneva, S. K., Barnard, S., D'Cruz, P., Noronha, E., Aroles, J., Kelliher, C., Koukoulis, S., Zimerman, I. (2021). What is it about humanity that we can't give away to intelligent machines? A European perspective. *International Journal of Information Management*, 58, 102311. https://doi.org/10.1016/j.ijinfomgt.2021.102311.

[17] Bogue, R. (2023). The role of machine learning in robotics. *Industrial Robot*, 50 (2), 197–202. https://doi.org/10.1108/IR-11-2022-0279.

[18] Huda, M., Rahayu, A., Furqon, C., Sultan, M. A., Susilawati Sugiana, N. S. (2025). Analyzing the impact of robotic process automation (RPA) on productivity and firm performance in the service sector. *International Journal of Advanced Computer Science Applications*, 16(6).

[19] Al Ayyubi, M. C., Shabrina, U. I., Febrianto, R., Naufal, F., Nafis, M. R., Lubis, L. S. (2024). Optimizing core banking operation's ROI with robotic process automation: A case study from a leading Southeast Asian bank. In *2024 IEEE International Conference on Control Automation, Electronics, Robotics, Internet of Things, and Artificial Intelligence (CERIA)* (p. 1–6). *IEEE*. https://doi.org/10.1109/CERIA64726.2024.10915031.

[20] Altarazi, F. (2025). *Integrating process mining, robotic process automation, and data mining: A framework for optimizing inventory management* [Doctoral dissertation, State University of New York at Binghamton].

[21] Heinonen, O. (2025). *Design and implementation of a machine learning based tool and a robotic process automation workflow for SAP S/4HANA item setup* [Master's thesis].

[22] Ahmed, F., Ahmed, M. R., Kabir, M. A., Islam, M. M. (2025). Revolutionizing business analytics: The impact of artificial intelligence and machine learning. *American Journal of Advanced Technology and Engineering Solutions*, 1(1), 147–173. https://doi.org/10.63125/f7yjxw69.

[23] Sharma, C., Shah Bharadwaj, S., Gupta, N., Jain, H. (2023). Robotic process automation adoption: Contextual factors from service sectors in an emerging economy. *Journal of Enterprise Information Management*, 36(1), 252–274.