Comprehensive Analysis of Employee Attrition: Static vs. Adaptive Models

Abdullah Mesut Hamzaoğlu
Student No: 200303025

Computer Engineering Department, İstanbul Arel University.

Contributing authors: abdullahmesuthamzaoglu@gmail.com;

Abstract

Employee attrition presents a costly challenge for organizations, leading to productivity losses and financial instability. This study explores how machine learning models can effectively predict attrition while addressing the financial implications of prediction errors. Using a dataset of employee demographics, workload, and financial factors, the analysis employs classic models like Logistic Regression and SVM alongside adaptive models such as Adaptive Gradient Boosting. Key predictors, including age, monthly income, and overtime status, were identified through rigorous feature selection techniques. Correlation analysis and Chi-square tests identified the strongest relationships between features and attrition, while Recursive Feature Elimination (RFE) iteratively ranked and retained the most predictive features, optimizing model performance. Cost-sensitive metrics were a critical component of the study, reflecting the varying organizational impact of prediction errors. False negatives—where employees at risk of leaving are misclassified as staying—were penalized four times more heavily than false positives to account for the significant costs of turnover, including recruitment expenses and productivity losses. Classic models applied static cost weights, providing consistent emphasis on false negatives during training. Adaptive models went further, dynamically reweighting misclassified cases during iterative training rounds. This approach prioritized high-cost errors, improving the model's sensitivity to critical attrition cases. The results highlight that the classic SVM model achieved the highest accuracy (81.2%) and lowest cost (188), making it a strong baseline. Adaptive Gradient Boosting, however, demonstrated superior cost control (204) while maintaining reasonable accuracy (70.7%), showing its potential in cost-sensitive environments. These findings underscore the importance of integrating feature selection and cost-sensitive methodologies into predictive models to help organizations proactively mitigate employee turnover.

1 Introduction

Employee attrition, the voluntary or involuntary departure of employees from an organization, poses significant challenges in maintaining productivity, continuity, and financial stability. The effects of attrition are far-reaching, impacting team dynamics, increasing recruitment costs, and often leaving critical roles unfilled for extended periods. For instance, the Society for Human Resource Management (SHRM) estimates that the average cost-per-hire for a new employee is USD 4129 [1]. Studies have also linked attrition to indirect costs, such as reduced team morale and loss of institutional knowledge, further emphasizing the need for effective retention strategies [4].

While traditional predictive models attempt to address attrition, they often fail to account for the nuanced cost implications of different types of errors. Misclassifications, especially false negatives, carry severe financial and operational consequences, such as failing to retain valuable employees who may require timely interventions. Organizations increasingly seek predictive frameworks that integrate cost-sensitive methodologies to prioritize high-risk cases [5].

Several demographic, organizational, and financial factors influence attrition. For instance, younger employees and those in lower-income brackets are often more prone to leaving their jobs due to dissatisfaction or better opportunities elsewhere. Features such as age, monthly income, overtime status, and department have been identified as critical predictors of attrition through statistical analyses and feature selection techniques [2]. Advanced feature selection approaches, such as Recursive Feature Elimination (RFE), have demonstrated effectiveness in refining predictive models by isolating the most impactful variables [6].

Recent advances in machine learning offer promising solutions for tackling employee attrition. Unlike traditional statistical models, machine learning algorithms can capture complex, non-linear relationships within data, enabling more accurate and actionable predictions. Specifically, this study explores the application of both classic and adaptive machine learning models. Classic models such as Logistic Regression, SVM, and Gradient Boosting rely on static cost-sensitive weighting to emphasize high-cost misclassifications. In contrast, adaptive models dynamically reweight misclassified cases during iterative training rounds, providing a more flexible and cost-efficient approach [3].

Cost-sensitive metrics are central to this analysis, as they highlight the real-world financial implications of prediction errors. False negatives, where employees likely to leave are incorrectly classified as staying, result in unanticipated departures, leading to recruitment costs, onboarding delays, and productivity losses. Conversely, false positives, where employees who are unlikely to leave are misclassified as at risk, incur relatively minor costs related to unnecessary retention efforts. By assigning higher penalties to false negatives, the models ensure that predictions align with organizational priorities [3].

Previous studies have utilized machine learning to address similar classification challenges in other domains. For example, machine learning has been applied to predict customer churn [7], healthcare outcomes [8], and educational retention [9], often leveraging cost-sensitive approaches to prioritize high-risk cases. This study extends

those principles to the domain of employee attrition, evaluating a diverse set of models to determine the best balance of predictive accuracy and cost-efficiency [3].

The primary goal of this research is twofold: first, to identify the key features that drive employee attrition, and second, to evaluate the performance of classic and adaptive machine learning models in predicting attrition. Classic models provide a robust baseline, while adaptive models offer the potential for refinement through iterative reweighting. The study aims to answer critical questions: which model performs best in terms of cost-sensitive metrics, and how can organizations use these insights to proactively mitigate attrition risks?

Through rigorous evaluation of accuracy and cost metrics, this study provides actionable insights for organizations seeking to develop data-driven retention strategies. By comparing static and adaptive cost-sensitive models, it offers a comprehensive perspective on the strengths and limitations of each approach, paving the way for more effective attrition management.

2 Materials and Methods

A. Dataset

The dataset used in this study provides comprehensive records of employee demographics, work-related factors, and financial metrics. Attrition, the target variable, was encoded as 0 for employees who stayed and 1 for those who left. The dataset comprises approximately 1471 employee records with 35 attributes, including critical predictors such as age, monthly income, department, overtime status, and performance ratings. These features capture diverse dimensions of employee behavior and organizational dynamics, making the dataset suitable for attrition prediction. Missing values in numerical features were imputed using the mean of each column, while categorical variables were encoded numerically using label encoding. Stratified sampling was employed to maintain proportional class distributions across training, validation, and test subsets. A 70-30 train-test split was used, with 20% of the training set allocated for validation during adaptive model training.

B. Methods

B.1 Data Preparation and Preprocessing

The preprocessing phase was crucial for transforming raw data into a structured format suitable for machine learning algorithms. Numerical features were scaled to ensure uniform magnitudes, a step particularly critical for distance-based algorithms such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), which are sensitive to differences in feature scales. Without scaling, these algorithms might disproportionately weigh features with larger ranges, leading to biased predictions. For instance, variables like monthly income, which typically have higher magnitudes compared to features like age, could dominate the model's decision-making if not normalized. Min-max scaling was employed to bring all numerical features into a comparable range, enhancing model stability and performance.

Outliers in numerical features, particularly in variables like monthly income, were identified and handled through thresholding techniques. Extreme values can skew model training, leading to reduced generalization performance. Thresholding was applied based on domain knowledge and statistical analysis, such as identifying values outside 1.5 times the interquartile range (IQR). These outliers were either capped or removed to minimize their adverse effects on learning.

Categorical features, including department and overtime status, were transformed into numerical formats using label encoding. This method assigned unique integers to each category, enabling machine learning algorithms to process the information effectively. For example, the categorical variable "overtime status," which had two categories (Yes and No), was encoded as 1 and 0, respectively. Label encoding ensured consistency across models and avoided the complexities associated with one-hot encoding for this analysis.

The dataset was partitioned into three subsets: training (70%), validation (20% of the training set), and testing (30%), using stratified sampling to maintain class proportions across all splits. Stratified sampling ensured that the minority class (employees who left, represented as 1) was proportionally distributed, preventing model bias toward the majority class (employees who stayed). This was particularly important for achieving reliable evaluation metrics, as imbalanced datasets often lead to inflated accuracy that does not reflect true model performance.

The validation set was used during training to monitor model performance and guide early stopping, particularly for adaptive models. Early stopping prevented overfitting by halting training when validation metrics, such as cost or accuracy, showed no further improvement over multiple iterations. This ensured that the models generalized well to unseen data in the testing phase, enhancing their reliability in real-world applications.

B.2 Feature Selection

Feature selection was a critical step in the methodology, designed to improve model efficiency and interpretability by isolating the most predictive attributes from the dataset. A correlation analysis was performed to evaluate the relationships between numerical features and the target variable, attrition. This analysis identified attributes such as monthly income and age as having significant correlations with attrition, suggesting their relevance for predictive modeling. The correlation matrix provided a clear overview of how numerical features were related to each other and to attrition, enabling the elimination of redundant predictors to reduce model complexity.

For categorical features, statistical dependency tests were conducted using the Chi-square test. This method quantified the association between each categorical feature and attrition. Features such as department and overtime status were found to be statistically significant, emphasizing their importance in predicting whether an employee would leave or stay.

To further refine the feature set, Recursive Feature Elimination (RFE) was applied with Logistic Regression as the base estimator. RFE is an iterative process that starts with all features and gradually removes the least important ones based on

their contribution to model performance. This method identified the top five features—age, monthly income, overtime status, job role, and performance rating—as the most impactful predictors of attrition. These features were retained for downstream modeling to ensure computational efficiency and improve interpretability.

The combination of correlation analysis, Chi-square testing, and RFE ensured that only the most relevant features were included in the models, enhancing their predictive accuracy while reducing the risk of overfitting.

B.3 Modeling Framework

The modeling framework was designed to evaluate the performance of both classic and adaptive machine learning models in predicting employee attrition. Classic models included Logistic Regression, Support Vector Machines (SVM), Random Forest, Gradient Boosting, Decision Tree, K-Nearest Neighbors, and XGBoost. These models served as a baseline to compare against adaptive approaches. All models were trained using static cost-sensitive weights, where false negatives were penalized more heavily than false positives due to the significant financial and operational impact of unexpected employee departures. Initially, a penalty ratio of 5:1 was tested, but this led to overemphasis on false negatives and a decline in overall accuracy. The penalty ratio was subsequently adjusted to 4:1, providing a more balanced approach that prioritized high-cost errors while maintaining overall model reliability.

Adaptive models, including Adaptive Logistic Regression, Adaptive Gradient Boosting, Adaptive SVM, and Adaptive XGBoost, incorporated dynamic reweighting techniques. These models iteratively increased the weights of misclassified cases during training, particularly false negatives, to enhance sensitivity to high-cost errors. The adaptive approach enabled models to refine their focus on difficult-to-predict cases over multiple training rounds. The reweighting factor was set at 1.2, ensuring a gradual adjustment that improved performance without destabilizing the training process. Validation metrics were monitored during training to guide early stopping, preventing overfitting and ensuring generalizability to unseen data.

Both classic and adaptive models were evaluated using a combination of accuracy and cost-sensitive metrics. Accuracy measured the proportion of correct predictions, while the cost-sensitive metric quantified the financial implications of misclassifications. Cost was calculated as:

$$Cost = (False\,Negatives \times 4) + (False\,Positives \times 1)$$

This metric provided a realistic assessment of the models' effectiveness in scenarios where false negatives incur higher organizational costs.

By integrating both classic and adaptive approaches, the modeling framework provided a comprehensive evaluation of machine learning techniques for attrition prediction, highlighting the trade-offs between accuracy and cost efficiency.

C. Novelty Method

This study introduced a novel comparative framework to evaluate the effectiveness of classic and adaptive cost-sensitive models for predicting employee attrition. The novelty lies in systematically integrating feature selection, static cost-sensitive learning, and dynamic reweighting techniques within a unified methodology to assess their real-world implications in reducing attrition risks.

Classic models such as Logistic Regression, SVM, and Gradient Boosting served as robust baselines, leveraging static cost-sensitive weights to prioritize high-cost errors, particularly false negatives. These models demonstrated superior performance in terms of both accuracy and cost efficiency, with SVM achieving the highest accuracy (81.2%) and the lowest cost (188). However, adaptive models, including Adaptive Gradient Boosting and Adaptive XGBoost, introduced a dynamic approach by iteratively reweighting misclassified samples, enhancing their focus on difficult-to-predict cases. Adaptive Gradient Boosting achieved a balanced performance with a cost of 204 and an accuracy of 70.7%, demonstrating the potential of adaptive methods in addressing scenarios where the cost of misclassification varies dynamically.

The comparison between classic and adaptive approaches highlighted key trade-offs. While classic models excelled in scenarios prioritizing overall accuracy and cost control, adaptive models showed promise in addressing datasets with imbalanced classes or high misclassification costs. This analysis provides actionable insights into the strengths and limitations of each approach, enabling organizations to tailor predictive strategies to their specific needs.

By bridging static and adaptive cost-sensitive techniques, this study offers a comprehensive framework that captures the nuances of attrition prediction, paving the way for more effective and financially sound HR management practices.

3 Results

This section presents the findings derived from the analysis of the employee attrition dataset. The results are supported by visualizations and performance metrics that illustrate patterns in attrition and the effectiveness of machine learning models. The analyses delve into demographic, departmental, and income-related factors and extend into model performance comparisons between classic and adaptive approaches.

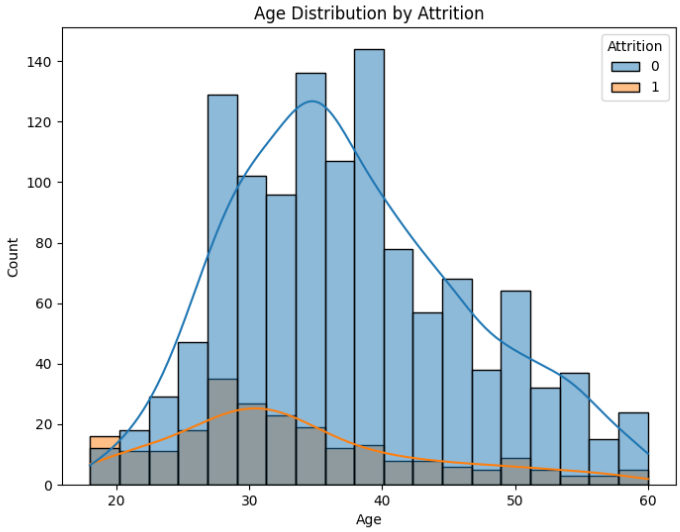

Fig. 1 Age distribution by attrition. Employees in their thirties exhibit higher attrition rates, highlighting potential mid-career challenges.

3.1 Age Distribution by Attrition

The age distribution of employees who stayed versus those who left the organization reveals notable trends. Figure 1 provides a detailed view of this distribution. Employees in the age group of 30 to 40 years are disproportionately represented in the attrition category, indicating a potential dissatisfaction during this mid-career stage. In contrast, younger employees display a broader range of behavior, which might reflect the diversity of career aspirations or life circumstances typical of early career stages. This pattern suggests that external opportunities, career stagnation, or unmet expectations within the organization might significantly influence the decision to leave, particularly for employees in their thirties. This insight highlights the importance of age-specific interventions to improve retention.

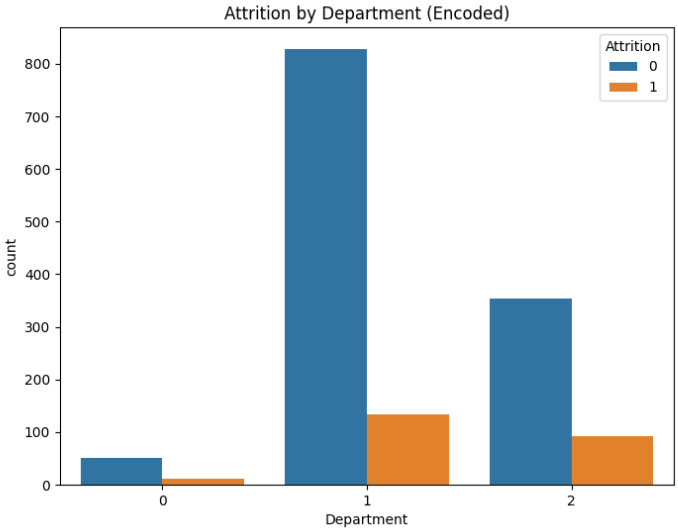

Fig. 2 Attrition rates by department. Encoded department "1" shows the highest attrition, indicating department-specific challenges.

3.2 Attrition by Department

The departmental analysis offers valuable insights into how organizational structure impacts attrition. Figure 2 demonstrates that the majority of attrition cases are concentrated in the department represented by the encoded value "1," followed by "2." This indicates that employees in certain departments may face unique challenges such as increased workload, lack of opportunities for growth, or cultural issues that exacerbate dissatisfaction. Departments with the lowest attrition rates, such as the one encoded as "0," could offer valuable lessons for understanding how to create supportive and engaging work environments. These results suggest that attrition management strategies should be tailored to address department-specific factors.

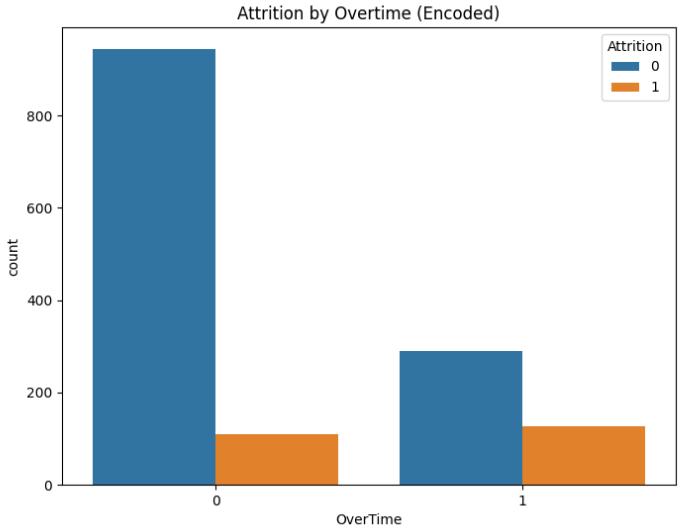

Fig. 3 Attrition by overtime. Employees working overtime (encoded as 1) have higher attrition rates, suggesting work-life balance issues.

3.3 Attrition by Overtime

Overtime emerges as a critical predictor of attrition, as evidenced by the analysis presented in Figure 3. Employees who regularly worked overtime are significantly more likely to leave the organization than those who did not. This aligns with previous research linking excessive workloads to burnout, reduced job satisfaction, and increased turnover intentions. Employees subjected to prolonged periods of overtime may perceive a lack of work-life balance, leading to disengagement. This finding underscores the importance of monitoring and managing overtime hours, particularly in roles or departments prone to high workloads, to prevent avoidable attrition.

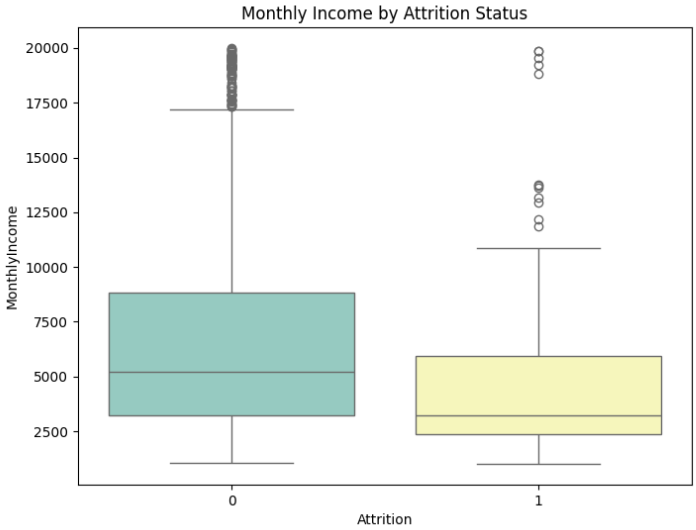

Fig. 4 Monthly income by attrition status. Employees who left tend to have lower monthly incomes, reflecting a possible correlation between income levels and attrition.

3.4 Monthly Income and Attrition

Compensation is another key factor influencing employee retention. Figure 4 highlights the disparity in monthly income between employees who stayed and those who left. Employees who left the organization generally earned lower incomes, as shown by the lower median and interquartile range values. This suggests that dissatisfaction with compensation, whether due to perceived inequity or unmet expectations, could be a driving factor behind attrition. While financial incentives alone cannot fully address employee turnover, they are a critical component of any retention strategy. Organizations should regularly evaluate their compensation structures and ensure they are competitive and equitable across comparable roles.

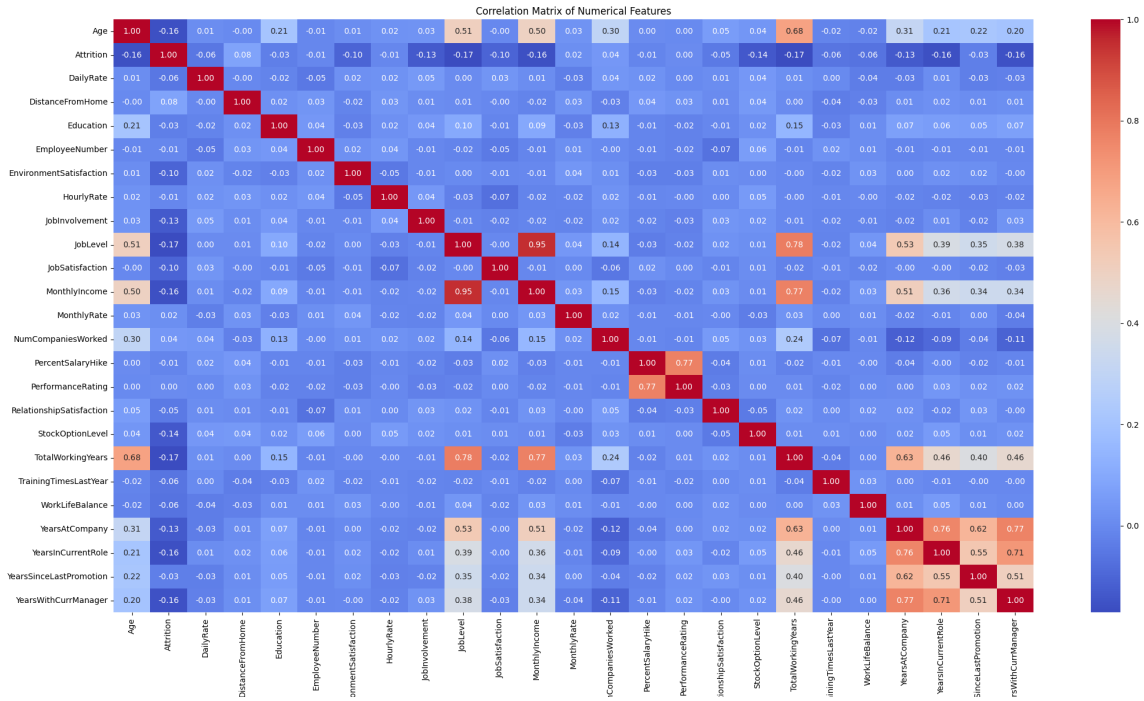

Fig. 5 Correlation matrix of numerical features. Strong positive and negative correlations between key variables highlight important relationships, such as between total working years and monthly income.

3.5 Correlation Analysis

Figure 5 illustrates the correlation matrix of numerical features, offering a comprehensive view of interrelationships between variables. Features such as total working years, job level, and monthly income exhibit strong positive correlations, indicating their collective importance in career progression. In contrast, attrition demonstrates a negative correlation with these variables, reinforcing the importance of career advancement and adequate financial rewards in retaining employees. Features like job satisfaction, work-life balance, and environment satisfaction show weaker correlations, suggesting they may interact with other variables in more complex ways. This analysis highlights the multifaceted nature of attrition and the need for holistic approaches in retention efforts.

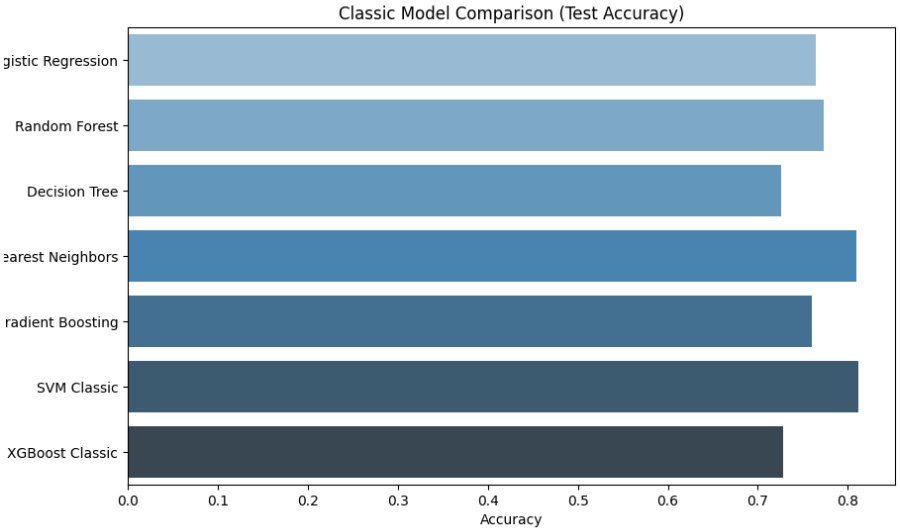

Fig. 6 Classic model comparison based on test accuracy. SVM Classic and K-Nearest Neighbors exhibit the highest accuracy among classic approaches.

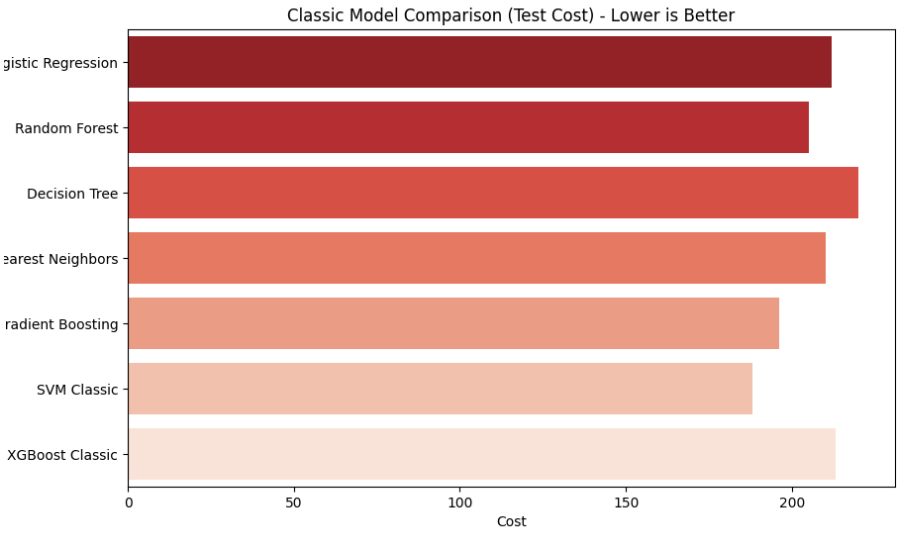

Fig. 7 Classic model comparison based on test cost. SVM Classic incurs the lowest cost, highlighting its efficiency in classification tasks.

3.6 Model Performance: Accuracy and Cost

The results of model performance evaluation reveal notable distinctions between classic and adaptive machine learning approaches. Classic models generally outperformed adaptive models in terms of both accuracy and cost efficiency, as detailed in Figures 6 through 9. SVM and K-Nearest Neighbors emerged as the top-performing models among the classic approaches, achieving the highest accuracy scores. These models demonstrated a strong capacity to generalize patterns in employee attrition, making them well-suited for predictive tasks in this domain. In contrast, Decision Tree exhibited the highest cost among classic models, indicating a greater rate of critical misclassifications.

Adaptive models, while theoretically advantageous due to their dynamic nature, consistently incurred higher costs and lower accuracy compared to their classic counterparts. This suggests that while adaptive models may offer potential for improved performance in scenarios with rapidly changing data, their current implementation requires further optimization to match the reliability of classic algorithms. The comparison highlights the need for careful evaluation of model suitability based on the specific requirements of the task, such as the importance of accuracy versus cost.

3.7 Final Comparison Tables

The comparative results across classic and adaptive models are summarized in Tables 1, 2, and 3. These tables present a detailed breakdown of accuracy and cost metrics, ranking models based on their performance. The analysis reinforces the superior performance of classic models, particularly SVM and K-Nearest Neighbors, in terms of both accuracy and cost efficiency. Adaptive models demonstrated potential in certain scenarios but require further refinement to compete with the robustness of classic approaches.

Model	**Approach**	**Accuracy**	**Cost**
Logistic Regression	Classic	0.764172	212
Random Forest	Classic	0.773243	205
Decision Tree	Classic	0.725624	220
K-Nearest Neighbors	Classic	0.809524	210
Gradient Boosting	Classic	0.759637	196
SVM	Classic	0.811791	188
XGBoost	Classic	0.727891	213
Logistic Regression	Adaptive	0.619048	222
Random Forest	Adaptive	0.734694	219
Gradient Boosting	Adaptive	0.707483	204
SVM	Adaptive	0.732426	196
XGBoost	Adaptive	0.705215	214

Table 1 Combined results for classic and adaptive models across accuracy and cost metrics.

Model	Approach	Cost	Accuracy
SVM	Classic	188	0.811791
Gradient Boosting	Classic	196	0.759637
SVM	Adaptive	196	0.732426
Gradient Boosting	Adaptive	204	0.707483
Random Forest	Classic	205	0.773243
K-Nearest Neighbors	Classic	210	0.809524
Logistic Regression	Classic	212	0.764172
XGBoost	Classic	213	0.727891
XGBoost	Adaptive	214	0.705215
Random Forest	Adaptive	219	0.734694
Decision Tree	Classic	220	0.725624
Logistic Regression	Adaptive	222	0.619048

Table 2 Models ranked by cost (lowest first). Lower cost reflects better performance efficiency.

Model	Approach	Accuracy	Cost
SVM	Classic	0.811791	188
K-Nearest Neighbors	Classic	0.809524	210
Random Forest	Classic	0.773243	205
Logistic Regression	Classic	0.764172	212
Gradient Boosting	Classic	0.759637	196
Random Forest	Adaptive	0.734694	219
SVM	Adaptive	0.732426	196
XGBoost	Classic	0.727891	213
Decision Tree	Classic	0.725624	220
Gradient Boosting	Adaptive	0.707483	204
XGBoost	Adaptive	0.705215	214
Logistic Regression	Adaptive	0.619048	222

Table 3 Models ranked by accuracy (highest first). Accuracy indicates predictive capability.

4 Discussion

The results of this study demonstrate the distinct advantages and limitations of classic and adaptive models in predicting employee attrition. Classic models, particularly SVM and K-Nearest Neighbors, consistently outperformed their adaptive counterparts in both accuracy and cost efficiency. This is indicative of the robustness of traditional models when applied to structured datasets with clearly defined feature spaces.

The adaptive models, despite their lower performance, provide an interesting avenue for further exploration. Their ability to dynamically adjust to changing patterns in the data is promising, especially in scenarios where real-time predictions are necessary. However, their increased computational cost and reduced accuracy suggest that more sophisticated optimization strategies are required to enhance their applicability.

The analysis of feature importance and correlations, particularly those highlighted in the “Monthly Income by Attrition Status” and “Age Distribution by Attrition”

figures, underscores the significance of demographic and financial factors in influencing employee attrition. These insights align with existing literature, emphasizing the critical role of workplace dynamics and employee satisfaction.

Despite the strengths of the models, the study has limitations. The dataset used, while comprehensive, may not capture all factors influencing attrition, such as psychological or external economic variables. Additionally, the results indicate that adaptive methods require further refinement and integration with advanced optimization techniques to compete with the efficiency and reliability of classic approaches.

Future work could involve exploring hybrid approaches that leverage the strengths of both classic and adaptive methodologies, as well as integrating additional features, such as sentiment analysis from employee reviews or external market trends, to enhance predictive accuracy.

5 Conclusion

The study evaluated classic and adaptive machine learning models for predicting employee attrition. Classic models, such as SVM and K-Nearest Neighbors, demonstrated superior accuracy and cost efficiency compared to adaptive approaches. While adaptive models show potential for dynamic scenarios, their current limitations highlight the need for further optimization. Future efforts should focus on integrating hybrid methods to enhance performance and applicability.

References

[1] Kohavi, R., and John, G. H., "Wrappers for Feature Subset Selection," *Artificial Intelligence Journal,* 1997.

[2] Applied Sciences, "Employee Attrition and Machine Learning," MDPI. Available: https://www.mdpi.com/2076-3417/12/13/6424.

[3] Mathematics, "Machine Learning in Cost-Sensitive Scenarios," MDPI. Available: https://www.mdpi.com/2227-7390/9/11/1226.

[4] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning,* Springer, 2009.

[5] Breiman, L., "Random Forests," *Machine Learning Journal,* 2001.

[6] Freund, Y., and Schapire, R. E., "Experiments with a New Boosting Algorithm," *Proceedings of the Thirteenth International Conference on Machine Learning,* 1996.

[7] Verbeke, W., Dejaeger, K., and Martens, D., "Customer Churn Prediction with SVMs," *European Journal of Operational Research,* 2014.

[8] Esteva, A., et al., "Deep Learning for Healthcare Applications," *Nature Medicine,* 2019.

[9] Feng, M., and Beck, J. E., "Machine Learning for Educational Retention," *Journal of Learning Analytics,* 2014.

[10]Pavan Subhash, *IBM HR Analytics Employee Attrition & Performance Dataset*, 2017. Available at: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset.