HR Analytics for Predicting Employee Attrition: A Hybrid Algorithmic Approach

Dr. Suvarchala Rani Korapole[1]
[1]Assistant Professor,
Department of Management Studies,
Bhavan's Vivekananda College of Science, Humanities and Commerce, Sainikpuri, Secunderabad, Telangana, India.
suvarchalarani.korapole@gmail.com

Abstract — In a data-driven business landscape, predictive HR analytics plays a strategic role in managing workforce stability. The employee attrition, a long-standing issue in IT exists because in-house knowledge-based functions and employee mobility are threats to the company's competitiveness. Traditional descriptive techniques give hardly any insights and organizations remain more reactive than proactive. This paper applies a hybrid HR analytics algorithmic framework, that draws from the Kaggle HR attrition data (n = 1,470) and validates the model on IT employee data from India (n = 102), demonstrating that the approach proved methodologically robust and practical feasibility. The predictive algorithms and model construction were carried out using Logistic Regression and Random Forest models, developed and analyzed based on prediction accuracy, interpretability, and generalizability. Based on the Kaggle dataset, Random Forest provided perfect scores to the data, though when it was used for IT workers, it did not generalize, indicating that the model was overfitting. Logistic Regression was less precise initially but generalized better and provided interpretable insight suitable for HR decision-making. Salary, tenure (especially in the first three years), overtime, job satisfaction, and age are stronger predictors of attrition. Younger employees and staff earlier in their careers were also more likely to leave, highlighting the generational and career stage dynamics. According to this study, attrition is influenced by financial indicators, career, lifestyle, and demographic components. Predictive analytics, and its application in HR practice, can provide evidence-based insights for proactive retention strategies in the IT sector

Index Terms — HR Analytics, Employee Attrition, Predictive Modelling, Logistic Regression, Random Forest, Hybrid Model Validation

Introduction

Retention of employees has become a key challenge in the IT industry due to high turnover rates, expansion into global markets, and the strategic emphasis placed on human capital (Bhatnagar, 2020; Sharma & Mishra, 2021). Attrition raises direct costs like recruitment and training, as well as loss of institutional knowledge and disruption of client delivery, which negatively affects the competitive edge of the organization (Hom et al., 2017).

Conventional methods of analysis of attrition rely on descriptive and diagnostic HR metrics that present past and current trends that do not offer predictive insight into the future. Accordingly, organisations are reactive to, rather than proactive on dealing with turnover (Minbaeva, 2018). In answer, HR analytics has become an efficient strategy that allows companies to move from the descriptive to predictive and prescriptive strategies in workforce management (Marler & Boudreau, 2017; Angrave et al., 2016).

Machine learning (ML) methods predict future attrition, e.g., Decision Trees, Random Forest, Logistic Regression, and other similar algorithms are popularly deployed. HR managers have few practical applications for random forest and ensemble methods because they are less interpretable despite their high accuracy (Fernandez & Singh, 2021).However, at the cost of logistic regression offers more interpretability in terms of predictive strength, which helps HR make the results more useful (Kaur & Fenech, 2021).However, in spite of these developments, the Kaggle HR dataset still capture most of the existing literature. But, when benchmarking is used extensively, it fails to encompass sector-specific characteristics regarding IT attrition like early-career turnover, workload intensity, and generational differences (Jain & Sinha, 2022; Reddy & Reddy, 2020).

Additionally, prior studies have tended to favour accuracy over generalizability and interpretability. To overcome these shortcomings, the current study employs a hybrid algorithmic and predictive analytics model using the Kaggle dataset and validated against primary IT staff data. This design increases

generalizability while providing practical utility. Still, there is limited research confirming that global predictive models are reliable in the Indian IT field where the composition of the workforce and the progression of careers are very different. To address these identified research gaps, this study sets out the following objectives: (1) Create and compare predictive models for employee attrition at organizations in IT, (2) Validate generalizability with data from IT employees and (3) Identify key predictors of attrition in the IT sector and offer suggestions relating to retention strategies.
Thus, this study bridges the gap between predictive accuracy and interpretability by developing and validating a hybrid model suitable for IT workforce analytics.

Review of Literature and Research Gaps

Theoretical Perspectives on Attrition.

Theories have been introduced explaining attrition. Human Capital Theory focuses on personnel being priceless, and a high turnover results in diminishing their skill and experience investment in the organizational capital and thus the investment in abilities and background (Becker, 1964; Lepak & Snell, 1999).. Herzberg's Two-Factor Theory differentiates between hygiene factors (e.g., pay and working conditions) and motivators (e.g., recognition, growth), leading toward either job dissatisfaction or turnover (Herzberg, 1959; Bhatnagar, 2020) as motivators and in the case of dissatisfaction on top of it, to leave the organization (Herzberg, 1959; Bhatnagar, 2020). Job Embeddedness Theory improves on retention knowledge by looking at the network aspect, fit, and what employees would leave as a result (Mitchell et al., 2001; Zhang et al., 2012).

Human Resource Analytics and Staff Planning.

HR analytics has revolutionized HRM through shifting descriptive towards predictive and prescriptive measures (Angrave et al., 2016).It is increasingly used to forecast employee attrition, expedite workforce planning, and align with the strategic objectives of the organization (Rasmussen & Ulrich, 2015; Fitz-enz & Mattox, 2014).However, little indication exists of the use of empirical validation in the few organizations, with the bulk of the literature on the topic being conceptual (Minbaeva, 2018; Marler & Boudreau, 2017).

Predictive Modeling in Attrition Studies.

We discuss a different paradigm for the above, that of predictive modeling, which is applied to Attrition. Machine learning models have proven to be highly effective for attrition prediction. Decision Trees and Random Forests have been used to capture non-linear relationships well (Nagadevara et al., 2008; Mehta & Gupta, 2022). Logistic Regression continues to be a favored model because of its interpretability and managerial relevance (Fernandez and Singh, 2021; Kaur and Fenech, 2021). Although both neural networks and ensemble models, which enhance accuracy, are viewed as black box methods with limited transparency for HR managers, they are also used to diagnose errors, and not be easily adaptable for real-time anomaly detection (Khan & Maini, 2020; Yadav & Shankar, 2019).

Attrition in the IT Sector.

Attrition is particularly high in IT, where the demand for talent is demanding, and the opportunity to work anywhere in the world attracts talent. A study revealed that the poor work/life balance, long work hours, and salary dissatisfaction are well recognized as major factors of turnover at IT companies (Sharma & Mishra, 2021; Reddy & Reddy, 2020). Especially important is early-tenure dropout, as most employees would most likely quit, in the first three years (Hom et al., 2017; Venkatesh & Choudhury, 2019).In general, younger workers areless likely to remain in the workforce (Patel & Desai, 2021), and generational disparities in mobility preferences and career goals exacerbate this. Previous studies showed that predictive HR analytics could be used to detect attrition trends and core predictors. Nonetheless, most research is constrained by a single dataset, values accuracy over interpretability and little generalization to the IT space. Adopting a mixed-method approach (using both Kaggle and IT employee data), we inform the current attrition studies in ways that are both methodological and practical, providing theory-based and practical retention techniques.
Table 1 consolidates prior literature to position the current study within existing HR analytics frameworks.

Table 1. Summary of Selected Literature on Employee Attrition and HR Analytics

Authors & Year	Context / Method	Key Findings	Limitations	Contribution to the Present Study
Becker (1964); Lepak & Snell (1999)	Theoretical – Human Capital Theory	Employees as strategic assets; attrition erodes firm value.	Conceptual; not empirically validated with predictive analytics.	Provides theoretical foundation linking attrition to loss of human capital in IT sector.
Herzberg (1959); Bhatnagar (2020)	Two-Factor Theory; Indian IT case studies	Attrition driven by dissatisfaction with pay, workload, lack of motivators.	No predictive validation; context specific.	Supports role of motivation and hygiene factors in IT attrition.
Mitchell et al. (2001); Zhang et al. (2012)	Job Embeddedness Theory	Links, fit, and sacrifice predict retention.	Limited testing in IT and predictive modeling.	Provides multidimensional lens for attrition factors beyond pay and tenure.
Angrave et al. (2016); Minbaeva.	HR Analytics frameworks.	HR developments in the direction of predictive and prescriptive analytics.	Mostly conceptual; few empirical validations.	Explains predictive HR analytics as methodological base.
Marler & Boudreau (2017); Fitz-enz & Mattox (2014)	HR analytics applications	Predictive analytics enhances HR decision-making	Limited focus on IT workforce attrition.	Provides rationale for HR predictive analytics in IT sector.
Nagadevara et al. (2008)	Data mining on attrition (Indian firms)	Data mining helps identify turnover risks.	Focus on withdrawal behaviors; interpretability issues.	Shows potential of machine learning for attrition prediction.
Mehta & Gupta, (2022)	proposed ML models for attrition in IT firms	high degree of accuracy provided by Random Forest.	Overfitting risk; limited generalizability.	Highlights risk of overfitting — addressed by hybrid validation here.
Fernandez & Singh (2021)	Logistic Regression & ML comparison	Logistic Regression offers interpretability; ML higher accuracy.	Single dataset; trade-off between accuracy and insights.	Study compares models for both accuracy and interpretability.
Kaur & Fenech (2021)	Predictive analytics in HRM	Predictive analytics helps identify turnover drivers.	Kaggle dataset reliance; limited external validation.	Validates findings with IT dataset to improve generalizability.
Khan & Maini (2020); Yadav & Shankar (2019)	Neural Networks & Ensembles	High predictive accuracy among Attrition studies.	"Black-box" models limit interpretability for HR decision-making.	Strengthens studies, pay more attention to interpretable models, such as Logistic Regression.
Sharma & Mishra (2021); Reddy & Reddy (2020)	IT attrition studies in India	Salary, overtime, work-life balance are key attrition drivers.	Context-specific; descriptive in nature.	Confirms importance of workload and compensation factors in IT dataset validation.
Hom et al. (2017); Venkatesh &	Turnover patterns	Early-career employees most prone to attrition.	General workforce focus; not validated with predictive models.	Study examines tenure as a key predictor in IT attrition.

Authors & Year	Context / Method	Key Findings	Limitations	Contribution to the Present Study
Choudhury (2019)				
Patel & Desai (2021)	Generational analysis of IT attrition	Younger employees (<30) show higher turnover tendencies.	Limited predictive validation; cross-sectional.	Aligns with findings of higher attrition risk among younger IT employees.
Jain & Sinha (2022)	Hybrid validation in HR analytics	Combining datasets strengthens model generalizability.	A low volume of application specific to IT- attrition.	Directly applies to the hybrid approach in this study, Kaggle + IT employee data.

Source: Author

Table 1 illustrates the significant theoretical, conceptual, and methodological advancements of previous studies on employee attrition. Nonetheless, most studies are descriptive, secondary datasets like Kaggle are utilized, or predictive accuracy is prioritized over interpretability and generalizability. In an IT industry characterised by various sources of attrition, such as wages, duration, workload, and generational mobility, such constraints reduce the ability to build theory and apply management practice. This research fills these gaps by utilizing hybrid predictive analytics. It explores and compares predictive models of attrition in the IT industry, validates whether these can be generalised to IT worker data, and presents key predictors to help develop appropriate retention strategies. By such bridging, the study addresses the methodological gap on HR analytics and delivers practical implications for the IT workforce challenges.

Theoretical Framework

The phenomenon of employee attrition has been studied for a long time: in organizational behavior, human resource management, and labor economics theory, and from various theoretical perspectives. Three interrelated frameworks underpin this research which assist in understanding why people leave companies and how predictive analytics offers pragmatic information on why employees leave organizations. The theory is underpinned by the conceptual models and visualization schemes included in the below drawings.

Human Capital Theory

The Human Capital Theory (Becker, 1964) argues that the employees are valued assets who invest in their skills and knowledge; organizations can be in optimal situation by retaining the employees. Attrition is a decline in the sums invested in knowledge, skills, and experience. In IT there is a lot of intellectual and technical capital that is critical to competitiveness; attrition is about the strategic erosion of human capital.

Herzberg's Two-Factor Theory

Herzberg's theory of motivation-hygiene distinguishes between motivators (intrinsic factors such as growth, recognition, and meaningful work) and hygiene factors (extrinsic elements such as pay, job security, and working conditions) in the workplace. Dissatisfaction with hygiene factors encourages turnover, while the lack of motivators curtails retention; while pay makes sense under a competitive economy, it can stifle retention.

Job Embeddedness Theory

Mitchell et al. (2001) also explain employee retention theory as Job Embeddedness Theory of retention, where connections to employees (including working relationships), belongingness to the organization (which is how one fits in one's work), and how well, for example, an employee sacrifices (losing something if the employee leaves the company) are considered key reasons why a firm keeps employees. In IT, attrition has been observed to be quite high due to low perceived sacrifice in the sense that other companies provide positions with similar jobs and better pay and flexibility; employees feel the opportunity is theirs to be taken upon.

Socio-Technical Perspective

From a socio-technical perspective, predictive HR analytics can be seen as a method of decision-support in conjunction with technical models based on social factors. But machine learning models are only effective on terms that may only be number-correct; their true value comes when the outputs are in line

with theories of employee behavior. Following are the ideas described in the conceptual and the theoretical linkages between these ideas.
The conceptual framework (Fig. 1.) integrates theoretical and analytical perspectives to explain attrition predictors in the IT sector.

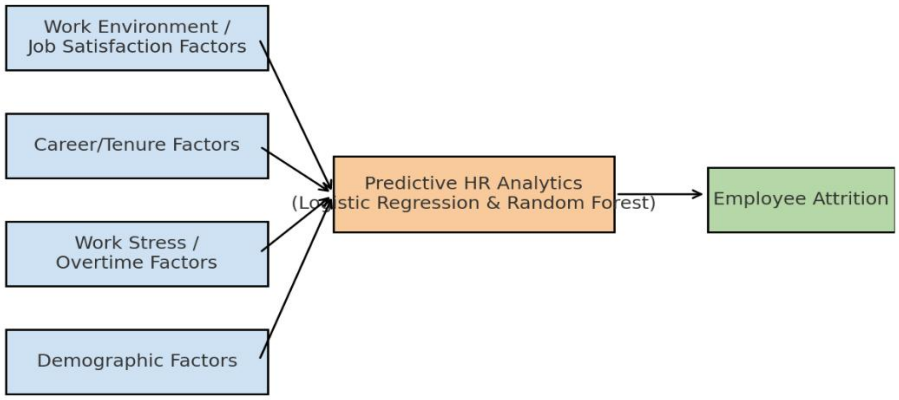


Fig 1. Conceptual Framework of Attrition Prediction in IT Sector
Source: Author

As seen in Fig 1, attrition in the IT industry has been seen to be a complex combination of various features combined with predictive analytics. This figure highlights the fact that attrition is a multidimensional phenomenon that needs data-driven analysis and theoretical foundations.
Theoretical linkages between existing HR theories and predictive HR analytics are demonstrated in Fig 2.

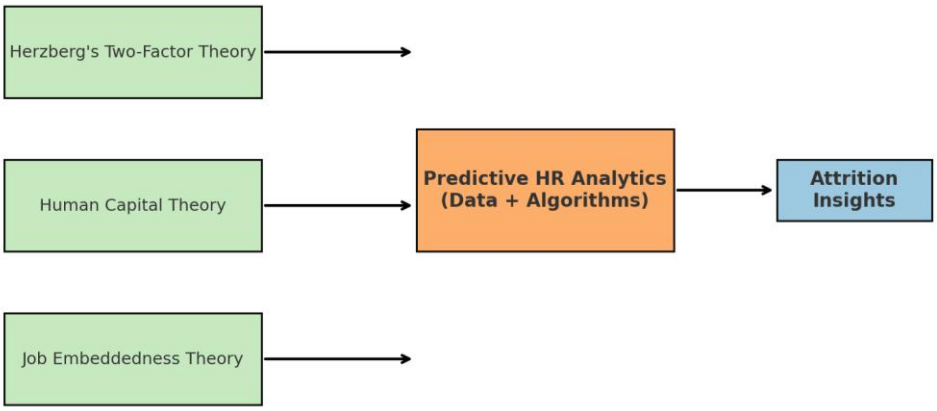


Fig 2. Theoretical Linkages to Predictive HR analytics
Source: Author

Predictive models as shown in Fig 2 are not an additional technical tool but a product of several decades of HR theory. The picture suggests that employee motivation, investment in human capital, and embeddedness are quantifiable and testable through predictive analytics.
Fig 3 incorporates the socio-technical perspective to elucidate how social and technical systems interact in HR analytics.

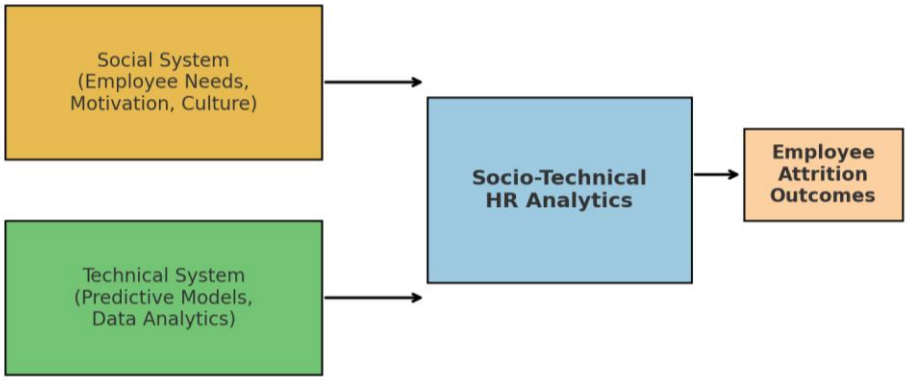


Fig 3 Socio-Technical Perspective on HR Analytics
Source: Author

Predictive HR analytics, exemplified in Fig 3, must be perceived not simply as a technology, but as a socio-technical system in which patterns derived from data are interpreted in ways that relate not only to human behaviour, but to practices at the organizational level.

4. Methodology

Using a hybrid predictive analytics approach the study examines and predicts employee attrition in the IT industry. Using valid data from a publicly-available Kaggle dataset as a secondary source and 102 individuals of primary IT employees in India, the methodology combines methodological rigor with practical relevance.

4.1 Research Design

The study followed a multi-step plan. First, data were obtained from the Kaggle HR Attrition dataset (for model preparation) and a smaller IT employees' dataset (for validation). The data were cleaned, encoded, and missing values were processed. The logistic regression and random forests-based predictive algorithms are interpretable and ensemble-based algorithms, which were adopted. These models were tested with IT employee data to establish the generalizability of the models, and interpreted by methods of the performance measures, confusion matrices and importance scores.

4.2 Data Sources

Kaggle HR Attrition Dataset

The Kaggle dataset is a popular HR analytics research database with demographic, job and organizational information. Central variables include age, job satisfaction, overtime work, monthly earnings, years worked, number of companies and attrition status.

Dataset on IT Employee Validation

To test model generalizability, the researcher collected a smaller dataset from IT professionals using HR records and structured surveys from colleagues in the researcher's professional network. The sample was collected using convenience sampling from professional networks and HR contacts within IT organizations. This dataset had approximately 102 employees on different levels. To validate the model, we aligned variables with those in the Kaggle dataset.

4.3 Variables

Employee Attrition (Yes/No) was the dependent variable. Independent variables included:

- Financial parameters: Monthly Income, Stock Options.
- Career/Tenure variables: Years at Company, Total Working Years, Job Role.
- Workplace/Lifestyle factors: Overtime, Distance from Home, Job Satisfaction.
- Key demographic attributes: Age, Gender, Number of Companies Worked.

4.4 Model Development

Two models were developed:

- Logistic Regression, selected as interpretable and relevant for practically making HR decisions.
- Random Forest, selected as a more sophisticated ensemble approach to see the improvement in performance.

Both algorithms were trained on Kaggle dataset using Python's scikit-learn library, hyperparameter tuning using cross-validation. The models were developed using Python 3.10 and scikit-learn version 1.3

maintained with 70:30 training, test ratio for validation. These models were tested for generalizability on IT employee data, and interpretation was conducted with performance measurements, confusion matrices, and feature importance analysis.

4.5 Analytical Tools

Validation Strategy

To evaluate the adaptability and generalizability of the algorithms, they were validated using the IT employee dataset. Accuracy, Precision, Recall, F1-score, and ROC-AUC were used to evaluate performance. We also built confusion matrices to measure patterns of misclassification, especially when it comes to false negatives — particularly expensive in the HR world.

Exploratory Analysis

In addition to predictive modeling, exploratory visualizations were employed to derive interpretations of the attrition patterns. Attrition was assessed by job satisfaction, overtime, tenure, and age group. The importance of Random Forest features was assessed to determine some important predictors of the attrition. All data analysis was performed in Python 3.10 and scikit-learn version 1.3 for modelling, pandas for manipulating the data and matplotlib/seaborn for visualization. All data were de-identified and solely made available for academic research using this principle of participant confidentiality and ethical conduct.

5. Discussions and Findings

5.1 Model Performance and Validation:

Predictive models were built utilizing the Kaggle dataset for the baseline analysis. The comparative statistics of performance are shown in Table 2. The model development and validation findings are presented sequentially below.

Table 2. Model performance on Kaggle dataset

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.660	0.148	0.667	0.242	0.687
Random Forest	1.000	1.000	1.000	1.000	1.000

Source: Author

As shown in Table 2, Random Forest achieved perfect scores (Accuracy, Precision, Recall, F1-score, ROC-AUC = 1.000). These perfect results can look amazing on paper but they signify overfitting, as the model memorised the training data rather than recognizing any kind of trend that generalizes. Logistic Regression, on the other hand, brought out average but balanced results (Accuracy = 0.660; ROC-AUC = 0.687) more useful for HR-related solutions.

The ROC curves of both models are shown in Fig. 4, which presents their predictive features.

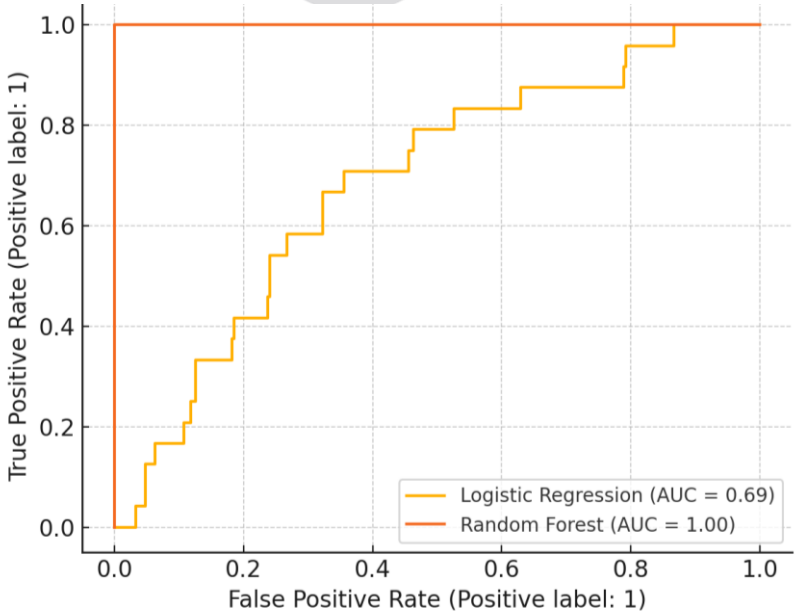


Fig. 4. ROC Curves – Kaggle Test Dataset

Fig. 4 illustrates the contrast: Random Forest's ROC curve hugs the top-left corner, suggesting memorization, while Logistic Regression follows a smoother trajectory that better reflects real-world trade-offs. Together, Table 2 and Fig. 4 emphasize that in HR analytics, models must be interpretable and generalizable. Logistic Regression, though less "perfect" in numbers, offers more actionable insights.

Next, the models were validated using IT employee data to test generalizability. The comparative results are presented in Table 3. These results confirm that model complexity does not necessarily improve predictive performance, especially when data characteristics differ across contexts.

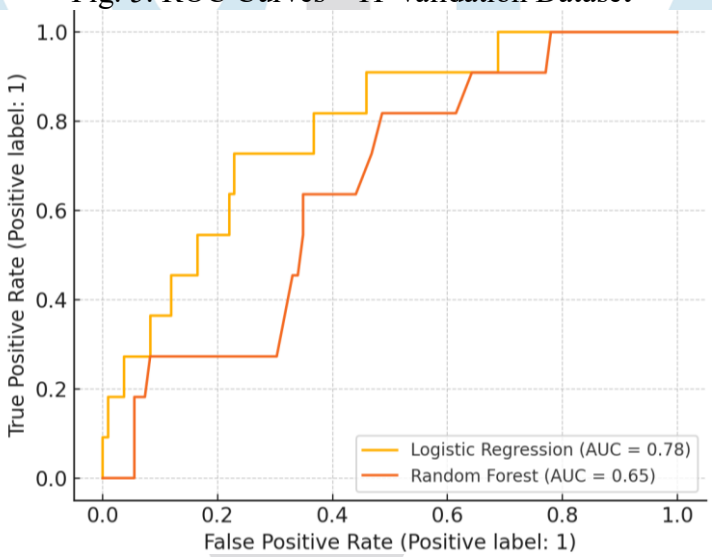
Table 3. Validation of models on IT employee dataset

Model	ROC-AUC (Kaggle)	ROC-AUC (IT dataset)	Change
Logistic Regression	0.687	0.785	+0.098
Random Forest	1.000	0.646	-0.354

Table 3 shows ROC-AUC for Logistic Regression applied to IT data rising from 0.687 to 0.785, thus reflecting a great adaptation capability. Random Forest, on the other hand, plummeted from 1.000 to 0.646, verifying its poor transferability from one dataset to any other. Logistic Regression was more successful in generalization than Random Forest (Table 3) was, which had perfect precision and less transferable accuracy.

To further assess how both models perform when applied to the IT validation data, the following Fig. 5. presents their Receiver Operating Characteristic (ROC) curves, which visually compare the predictive capability and discrimination strength of Logistic Regression and Random Forest models.

Fig. 5. ROC Curves – IT Validation Dataset



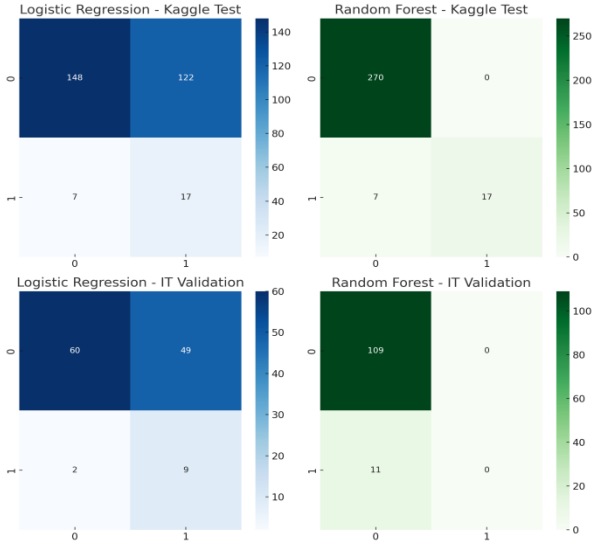
Source: Author

As shown in Fig. 5., Logistic Regression shows a strong curve in IT data whereas Random Forest fails to reach predictive strength. Simpler models can generally have greater generalizations than elaborate systems, especially in HR situations where workforce variability is high.

5.2 Predictive Accuracy and Error Analysis

Beyond overall accuracy, it is important to examine how correctly each model classifies cases of attrition and retention. The subsequent Figure 6 provides confusion matrix heatmaps that display the pattern of correct and incorrect predictions across the two models.

Fig. 6. Heatmaps of Confusion Matrices



Source: Author

The results indicated that the Logistic Regression generated fewer false negatives in IT validation. This is very important for HR as false negatives are employees who leave yet are falsely classed as staying which in turn can be quite the mistake of workforce planning. Logistic Regression, despite its lower accuracy, has better HR relevance for predicting employees at higher risk and this is captured by their inputs more effectively. Next, attrition patterns were explored by workplace conditions.

5.3 Workplace Factors Influencing Attrition

Fig. 7. presents attrition by job satisfaction levels.

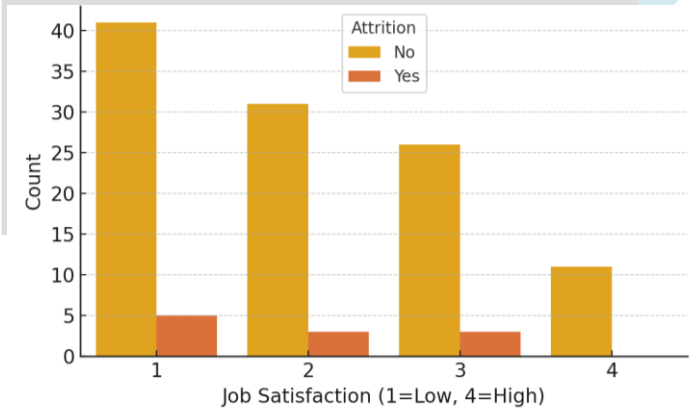


Fig. 7. Attrition by Job Satisfaction (IT Dataset)
Source: Author

Low job satisfaction employees (levels 1–2) exhibited in Fig. 7. significantly greater attrition rate evidencing dissatisfaction as the primary source of turnover.

To investigate workload's impact on attrition, we measured the relationship between overtime and attrition in Figure 8. Considering its pivotal role on workload in IT, overtime was considered as a possible driver of attrition.

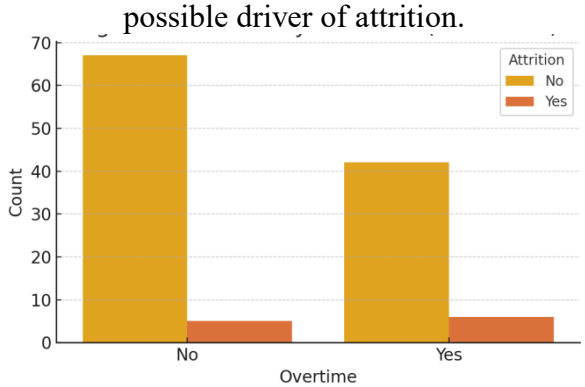


Fig. 8. Attrition by Overtime (IT Dataset)

Source: Author

Fig. 8. illustrates that employees who report overtime had a higher attrition indicating that it is important in information technology firms work on workloads as well as work-life balance problems. This demonstrates that attrition is not just about compensation but that engagement and work life balance are equally important.

5.4 Key Predictors and Demographic Trends

Next, Key primary indicators of attrition were detected. The findings are presented in Table 4 and illustrated in Fig. 9.

Table 4. Key Predictors of Attrition in IT Sector

Rank	Predictor	Implication
1	Monthly Income	Competitive salaries reduce attrition risk
2	Total Working Years	Experience influences stability; mid-career exit risk
3	Years at Company	Early-tenure employees face higher attrition
4	Distance from Home	Commute reduces satisfaction in IT hubs
5	Number of Companies Worked	Frequent movers are more likely to leave
6	Age	Younger employees switch jobs more often

Source: Author

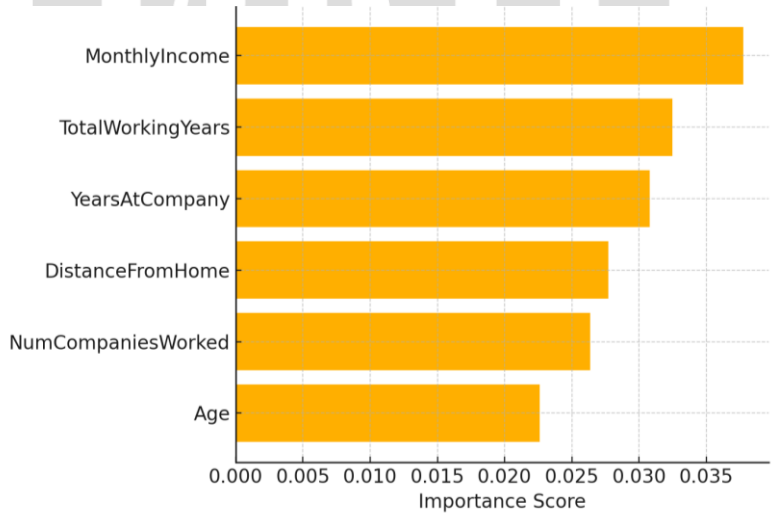
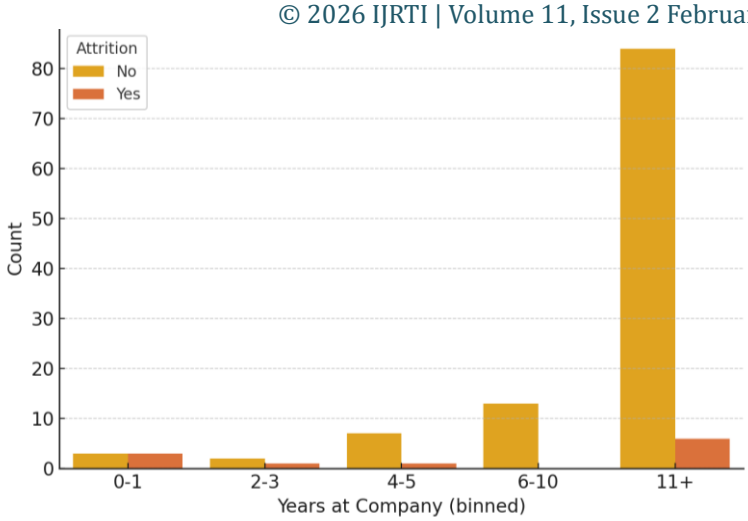


Fig. 9. Feature Importance – Top Predictors of Attrition
Source: Author

The results from Table 4 and Fig. 9. indicated that financial indicator (Monthly Income) had the highest predictive role, then tenure-related variable (Total Working Years, Years at Company), Workplace/lifestyle factors (Distance from Home, Overtime), and demographics (Age, Number of Companies Worked). IT attrition is attributed to the interplay of financial, career, workplace/lifestyle, and demographic factors.

To understand how tenure impacts turnover, Fig. 10. shows the distribution of attrition by years at the company, highlighting early-stage departures in IT organizations.

Fig. 10. Attrition by Years at Company (IT Dataset)

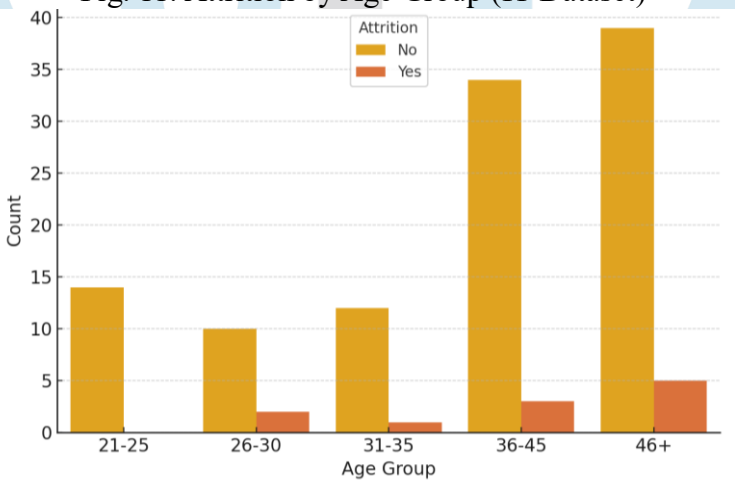


Source: Author

As indicated in Fig. 10. Attrition accelerates within the first 3 years, signalling that a tougher approach in terms of orientation at the outset and early involvement is required. This highlights the need for organizations to strengthen onboarding, engagement, and early career development initiatives to reduce early-tenure exits.

Beyond tenure, age also plays a crucial role in understanding attrition trends. Fig. 11. illustrates how turnover varies across different age groups within the IT workforce.

Fig. 11. Attrition by Age Group (IT Dataset)



Source: Author

Fig. 11. reveals that younger employees (below 30 years) exhibit significantly higher attrition rates, indicating generational shifts in career expectations and a preference for mobility and rapid growth opportunities.

5.5 Model Comparison and Generalizability

Finally, to evaluate model generalizability, Fig. 12. compares the ROC-AUC scores of Logistic Regression and Random Forest models across the Kaggle and IT datasets.

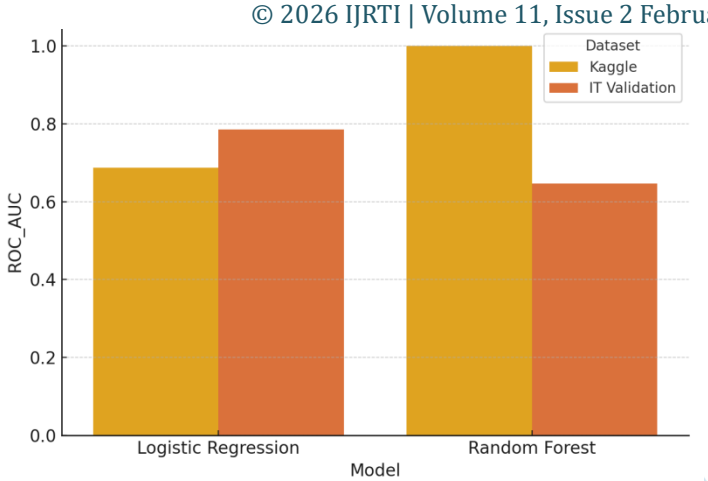


Fig. 12. Comparative ROC-AUCs – Kaggle vs IT Dataset

As demonstrated in Fig. 12., Logistic Regression shows better adaptability and generalization across datasets, while Random Forest performance declines, confirming that interpretable models are more reliable for HR predictive analytics in real-world contexts. The comparison confirms that Logistic Regression, though less complex, is more robust and interpretable across dataset making it a suitable model for HR predictive analytics.

6. Policy & Managerial Implications

The results from the hybrid predictive analytics approach offer theoretical as well as practical insights on employee attrition in the IT sector. It is shown in the analysis that complex models such as Random Forest are likely to have higher training data precision, while models which can be interpreted, such as Logistic Regression offer more generalizable and actionable knowledge that can help HR with decision making. In fact, tenure, job satisfaction, overtime, and age proved to be the most potent predictors of attrition, illustrating the multidimensional nature of workforce stability. As a result of these empirical findings, the following section will convert their results into policy and management suggestions for advancing employee retention and optimizing HR strategies. Drawing from the empirical evidence presented, the following managerial and policy implications are proposed.

The results offer actionable insights to IT organizations:

• Compensation approach: Set an ongoing salary benchmark and incentivise in a transparent manner.

• Career Growth: Offer structured career paths and leadership programs for early- and mid-career employees.

• Work-Life Balance: Address overtime with flexible scheduling, hybrid work, and wellness initiatives.

• Retention of younger/early staff: Improve onboarding, mentorship, and provide accelerated growth opportunities.

• Embedding analytics on-site with interpretation: using interpretable models to install predictive dashboards that can identify workers who are at risk in real time.

By integrating these evidence-based insights into HR dashboards and decision systems, IT organizations can proactively mitigate attrition risk and strengthen talent sustainability.

7. Conclusion

Such predictive HR analytics can be validated with actual data, and it brings meaningful insights into employee attrition. Random Forest performed perfectly well on the Kaggle dataset but failed to generalize to IT employees, indicating overfitting risk. While Logistic Regression was simpler, it was more robust and interpretable, illustrating that transparency far outweighs complexity when attempting to make HR decisions. Ultimately, HR decision systems that are built with algorithmic techniques can shift retention management from reactive strategies to predictive ones. The findings support the notion that attrition in the IT industry is multidimensional: competitive compensation is important, but also tenure, overtime, job satisfaction, and age are associated with a risk of turnover. The first three years have a significant turnover peak, and strong onboarding and career engagement are critical to retain a high rate of new graduates, with younger recruits at the greatest risk because of mobility and career ambition. For IT companies, predictive analytics should be integrated into HR dashboards to identify at-risk employees. These must include pay

benchmarking, career advancement, mentor guidance, flexible work schedules, and wellbeing initiatives. Ultimately, attrition is not solely an HR challenge but a strategic one, and predictive analytics can convert retention into a source of competitive advantage. This study reinforces the value of transparent and interpretable analytics in strategic HR decision-making.

8. Limitations and Scope for Future Research

The research provided strong evidence to the extent the predictive relationships found in the overall IBM HR Analytics dataset can be generalized to Indian IT employee data; however, the results are indicative, rather than representative. The validation dataset (n = 102) offers contextual evidence of model robustness in the Indian IT domain, but studies with larger and more heterogeneous data would be required to generalize the results across the Indian IT sector. Further research could investigate constructs such as organisational culture, support from managers, hybrid work habits, and career advancement variables to improve predictive power and contextual detail. Expanding validation on various IT subsectors and company sizes can expand and amplify the external validity of the model. Future research can also test longitudinal datasets to track employee movement over time, providing deeper insights into retention dynamics and temporal attrition patterns. Future studies extending this hybrid validation across sectors and over time can further advance predictive HR analytics as an applied discipline.

References

[1] D. Angrave, A. Charlwood, I. Kirkpatrick, M. Lawrence, and M. Stuart, "HR and analytics: Why HR is set to fail the big data challenge," *Human Resource Management Journal*, vol. 26, no. 1, pp. 1–11, 2016.
[2] G. S. Becker, *Human Capital*. Chicago, IL, USA: University of Chicago Press, 1964.
[3] J. Bhatnagar, "Talent management and employee retention in Indian IT sector," *Human Resource Development International*, vol. 23, no. 4, pp. 342–360, 2020.
[4] R. Fernandez and A. Singh, "Balancing accuracy and interpretability in HR predictive models," *International Journal of Human Resource Analytics*, vol. 8, no. 2, pp. 110–127, 2021.
[5] J. Fitz-enz and J. Mattox, *Predictive Analytics for Human Resources*. Hoboken, NJ, USA: Wiley, 2014.
[6] F. Herzberg, *The Motivation to Work*. New York, NY, USA: Wiley, 1959.
[7] P. W. Hom, T. W. Lee, J. D. Shaw, J. P. Hausknecht, and D. G. Allen, "One hundred years of employee turnover theory and research," *Journal of Applied Psychology*, vol. 102, no. 3, pp. 530–545, 2017.
[8] R. Jain and P. Sinha, "Validating predictive models in HR analytics: Hybrid approaches," *International Journal of Data Science in HR*, vol. 4, no. 2, pp. 89–104, 2022.
[9] P. Kaur and R. Fenech, "Predictive analytics in HRM: Applications and implications," *Journal of Business Research*, vol. 124, pp. 536–545, 2021.
[10] S. Khan and R. Maini, "Neural networks in HR attrition prediction: A case study," *Procedia Computer Science*, vol. 173, pp. 421–429, 2020.
[11] D. P. Lepak and S. A. Snell, "The human resource architecture: Toward a theory of human capital allocation and development," *Academy of Management Review*, vol. 24, no. 1, pp. 31–48, 1999.
[12] J. H. Marler and J. W. Boudreau, "An evidence-based review of HR analytics," *International Journal of Human Resource Management*, vol. 28, no. 1, pp. 3–26, 2017.
[13] R. Mehta and S. Gupta, "Employee attrition prediction using machine learning: Evidence from IT firms," *Journal of Human Resource Management*, vol. 10, no. 1, pp. 45–57, 2022.
[14] D. Minbaeva, "Building credible human capital analytics for organizational competitiveness," *Human Resource Management Journal*, vol. 28, no. 3, pp. 353–369, 2018.
[15] T. R. Mitchell, B. C. Holtom, T. W. Lee, C. J. Sablynski, and M. Erez, "Why people stay: Using job embeddedness to predict voluntary turnover," *Academy of Management Journal*, vol. 44, no. 6, pp. 1102–1121, 2001.
[16] V. Nagadevara, V. Srinivasan, and R. Valk, "Establishing a link between employee turnover and withdrawal behaviours: Application of data mining techniques," *Research and Practice in Human Resource Management*, vol. 16, no. 2, pp. 81–97, 2008.
[17] M. Patel and A. Desai, "Generational trends in IT employee attrition," *South Asian Journal of Human Resource Management*, vol. 8, no. 2, pp. 134–150, 2021.
[18] T. Rasmussen and D. Ulrich, "Learning from practice: How HR analytics avoids being a management fad," *Organizational Dynamics*, vol. 44, no. 3, pp. 236–242, 2015.

[19] K. Reddy and S. Reddy, "Drivers of attrition in Indian IT companies: An empirical study," *Global Business Review*, vol. 21, no. 5, pp. 1120–1137, 2020.
[20] N. Sharma and S. Mishra, "Workforce mobility and attrition challenges in IT-enabled services," *Asia Pacific Journal of Human Resources*, vol. 59, no. 2, pp. 225–242, 2021.
[21] V. Venkatesh and S. Choudhury, "Early-career turnover in IT firms: A longitudinal study," *Information Systems Research*, vol. 30, no. 1, pp. 210–227, 2019.
[22] P. Yadav and R. Shankar, "Ensemble models for employee attrition prediction," *Expert Systems with Applications*, vol. 125, pp. 272–285, 2019.
[23] M. Zhang, D. D. Fried, and R. W. Griffeth, "A review of job embeddedness: Conceptual, measurement issues, and directions for future research," *Human Resource Management Review*, vol. 22, no. 3, pp. 220–231, 2012.