IPRPD
International Journal of Business & Management Studies
ISSN 2694-1430 (Print), 2694-1449 (Online)
Volume 07; Issue no 01: January, 2026
DOI: 10.56734/ijbms.v7n1a1

EMPLOYEE ATTRITION PREDICTION UNDER DATA CONSTRAINTS: A COMPARATIVE ANALYSIS OF REAL AND SYNTHETIC DATASETS

ESSEN Mmedo[1], FASUNLADE Femi[2], OLADEJO Samson O.[3]

[1]*University of Bradford*

[2] *University of Portsmouth*

[3] *De Montfort University*

Abstract

Employee attrition remains a persistent challenge for organizations, with significant impaction for human capital sustainability and overall strategic organisational performance. However, existing empirical research in this field often relies on limited real-world datasets to predict employee turnover and explicitly assume that insights from real employee datasets are inherently superior despite increasing constraints related to data access, privacy and ethical governance, which raises substantial concerns regarding data accessibility, privacy, and ethical governance. Consequently, there is still limited understanding of whether synthetic data can be a reliable alternative for predictive modeling in human resources.

To address this gap, this study examines whether the employee attrition prediction differ when models are trained on real versus synthetically generated data in the context of employee turnover. The research uses the International Business Machines (IBM) employee dataset, binary logistic regression, random forest, and gradient boosting models are applied to real, bootstrapped, and synthetic datasets to assess predictive reliability. The original sample of 1,470 observations is expanded through bootstrapping and synthetic generation to create comparable datasets of 5,000 observations each. The study systematically compares the performance of binary logistic regression, random forest, and gradient boosting tree models across these different types of data.

The results indicate that synthetic data preserves key attrition-related relationship and yields predictive performance comparable to real data, although minor reductions are observed in identifying rare attrition cases. Factors affecting attrition, such as job satisfaction and age, remain consistent across both real and synthetic datasets. Furthermore, statistical analyses reveal no significant differences in predictive accuracy among the models. The study contributes to employee attrition and HR analytics research by demonstrating that attrition knowledge is not solely dependent on access to real employee data. It further offers practical insights for organisations seeking to leverage privacy-preserving analytics to support workforce planning and retention strategies under data constraints.

Keywords

Synthetic Data, Real Data, Binary Logistics Regression, Machine Learning, Random Forest, Xgboost Regression, Employee Attrition, International Business Machines

Introduction

Employee attrition remains a persistent organisational challenge with well-documented financial and operational implications (Ballinger et al., 2011; Cascio, 2007; Kingsley, 2025; Kuhn & Yu, 2021). Consequently, understanding the factors that drive employee turnover has become a central concern for scholars and managers, particularly in knowledge-intensive industries where skilled labour is a key asset (Pirrolas & Correia, 2022; Timsina, 2024). Prior research describes employee attrition as a multifaceted outcome influenced by demographic, job-related, and organisational factors (Awan et al., 2021; Belete, 2018; Castaldo et al., 2022; Yakut & Kara, 2021). These statistics underscore the necessity for effective retention strategies. In response, modern companies increasingly utilize data-driven human resource (HR) analytics to anticipate turnover risks and design targeted interventions (Skelton et al., 2019). Despite the growing sophistication of HR analytics, a significant methodological limitation exists in current research. Most predictive studies heavily rely on real-world employee datasets, which are often hard to obtain due to strict data protection regulations, confidentiality concerns, and ethical constraints (Vinoodhini, 2022). As a result, researchers frequently depend on a limited number of publicly available benchmarks, notably the IBM HR Analytics dataset, which contains only 1,470 observations. This sample size is often inadequate for effectively training complex machine learning models. Although bootstrapping is commonly used to expand limited datasets, it does not introduce new information and may reinforce existing distributional patterns (Muller et al., 2016). This reliance on limited real-world data creates a significant research gap. Although synthetic data generation offers a promising solution by creating datasets that retain the statistical properties of real data without compromising privacy, its application in employee attrition research remains largely unexplored. Little is known about whether predictive models trained on fully synthetic data can achieve accuracy metrics comparable to those trained on real data (Via et al., 2022; Rousseiet, Pernet, & Wilcox, 2023).

To address this gap, this study rigorously evaluates the utility of synthetic data in predicting employee turnover. By using the IBM dataset as a baseline, this research compares the performance of three distinct classification algorithms: binary logistic regression, random forest, and gradient boosting trees. The study employs both bootstrapping and synthetic data generation to expand the dataset to 5,000 observations, allowing for a robust comparison of model performance across different data conditions. Guided by the necessity to assess the methodological robustness of attrition prediction, this study addresses the following research questions:

- Is there a significant difference in the performance of the logistic regression model between real and synthetic employee datasets?
- Is there a significant difference in the performance of the gradient boosting tree model between real and synthetic employee datasets?
- Is there a significant difference in the performance of the random forest model between real and synthetic employee datasets?

The uniqueness of this study is that beyond its technical relevance, it contributes to employee attrition research by investigating how the nature of data, real versus synthetic shapes empirical insights, predictive reliability, and managerial interpretation. Though existing studies implicitly treat real employee datasets as the golden standard, this assumption remain mostly unexamined. However, our study explicitly compares model performance across data types to reframe employee attrition not only as a behavioural outcome but also as a knowledge problem influenced by data accessibility, ethical constraints and analytical design. In doing so, this study extends understanding of how HR analytics informs decision-making under data limitations, an increasingly salient issue in contemporary human resource management.

Literature Review

This literature review critically examines the drivers of employee attrition, the evolution of HR analytics from statistical to machine-learning methods, and the emerging potential of synthetic data to overcome privacy-related research constraints.

A. Employee Attrition as a Strategic and Organisational Challenge

Employee attrition has significant implications for organisational performance and human capital sustainability, particularly in knowledge-intensive environments. (Alnachef & Alhajjar, 2017; Dess & Shaw, 2001; Hash & Dyer, 2004; Kamoche, 1996; Kiran, Chaubey and Shastri, 2024; Muzammil et al., 2025). Research shows that beyond training and replacement costs, high employee turnover disrupts routines, erodes tacit knowledge, and diminishes productivity and innovation (Awan et al., 2021; Muzammil et al., 2025; Saini, Nimje and Kalra, 2022; Via et al., 2022). From a strategic perspective, attrition reflects a misalignment between employee expectations and organisational practices. Departing employees take explicit skills and social capital, which are hard to replace, raising the importance of understanding and predicting attrition, particularly in competitive markets. Cross-sectional studies have identified various factors influencing employee attrition, including individual attributes (age, education), job-related variables (compensation, workload), and organisational influences (leadership, culture, support) (Awan et al., 2021; Yakut & Kara, 2021; Castaldo et al., 2022). Such findings align with established theories such as job embeddedness and social exchange theory, which suggest that employees are more likely to stay when they perceive fit, support, and reciprocity.

B. Determinants of Employee Attrition: Individual, Job, and Organisational Factors

The decision to leave an organisation typically results from multiple determinants rather than a single event. Empirical studies indicate that job satisfaction and compensation are among the strongest predictors of employee attrition. Factors like lower remuneration, limited promotional opportunities, and inadequate recognition lead employees to seek alternative employment, especially in competitive labour markets (Saini et al., 2022). Additionally, work environment conditions, such as managerial support, interpersonal relationships, schedule flexibility, and workload, can significantly influence retention outcomes. However, the impact of these factors varies across organisational contexts. Employees in larger organisations or knowledge-intensive sectors often prioritise non-monetary factors such as autonomy, development opportunities, and organisational culture, while those in smaller firms or lower-skilled positions tend to emphasise wage stability and job security. Employee motivation serves as a crucial link between organisational practices and turnover outcomes. Motivated employees tend to exhibit higher commitment and lower withdrawal intentions, while low motivation correlates with increased turnover rates (Katsande & Chisoro, 2018). Viewed through this lens, employee attrition is the cumulative result of various interconnected individual, job-related, and organisational conditions. Moreover, employee turnover reflects the effectiveness of management practices and human resource systems. Turnover rates can serve as indicators of organisational alignment with employee expectations, with supportive environments associated with higher retention (Wilkens, 2020).

C. Work Environment, Compensation, and Contextual Drivers of Employee Attrition

Research highlights the crucial impact of the work environment and compensation structures on employee retention and attrition, particularly in high-stress settings such as call centres. Conversely, a supportive work environment enhances job satisfaction and confidence, reducing turnover (Iqbal et al., 2017; Setiyani et al., 2019). Beyond compensation, elements such as equitable job opportunities, skills development, and career advancement opportunities greatly influence organisational attachment. Human resource managers should adopt long-term, employee-centred strategies that foster appreciation and sustainable career paths. The work environment also includes organisational policies and managerial practices (Silva et al., 2019). Absence of coherent retention strategies can lead to dissatisfaction and accelerated turnover. Flexible work schedules have emerged as another key factor in retention. Nasir and Mahmood (2016) demonstrate that such arrangements enhance work-life balance and lower attrition. While compensation is often viewed as the primary retention factor, research indicates that the relationship is complex; factors like company size and job responsibilities also affect attrition rates (Duhautois et al., 2016). Significantly, higher wages tend to reduce the likelihood of turnover (Trembley et al., 2006; Milkovich & Newman, 2004).

D. HR Analytics and Machine Learning Approaches to Employee Attrition

The complex nature of employee attrition has led researchers to increasingly employ data-driven and machine learning methods to model turnover behaviour (Fitz-enz & Mattox, 2014; Fallucchi et al., 2020). Traditional techniques, particularly binary logistic regression, have been favoured for their interpretability and connection to established organisational theories (Hom et al., 2017). While effective in estimating average effects and identifying significant predictors, these methods often rely on linearity and independence assumptions that may not adequately represent the nuanced realities of employment decisions. As workforce data grows larger and more varied, traditional models struggle to capture the non-linear relationships and interactions in employee turnover dynamics (Breiman, 2001; Strohmeier & Piazza, 2015). To address these limitations, recent studies have turned to machine learning algorithms, such as random forests and gradient boosting trees, which excel at modelling non-linear interactions and complex feature spaces (Ali & Bader, 2021; Fallucchi et al., 2020). Research using the IBM employee dataset demonstrates the high classification accuracy of these models in predicting employee attrition based on factors such as income, job role, overtime, and satisfaction (Yang & Islam, 2020). However, many existing studies focus narrowly on maximising predictive accuracy within a single dataset, with less emphasis on the robustness of these models across diverse data conditions or the implications of data quality and availability in HR analytics research.

E. Data, Knowledge Production, and HR Analytics In Employee Attrition Research

Over the past two decades, numerous extant employee attrition studies (such as Alnachef & Alhajjar, 2017; Awan et al., 2021; Castaldo et al., 2022; Dess & Shaw, 2001; Fallucchi et al., 2020; Fitz-enz & Mattox, 2014; Hash & Dyer, 2004; Iqbal et al., 2017; Kamoche, 1996; Kiran, Chaubey & Shastri, 2024; Muzammil et al., 2025; Setiyani et al., 2019; Yakut & Kara, 2021) largely assumes that empirical insights derived from real organisational datasets characterize unbiased and superior reflections of employee behaviour. However, this assumption overlooks the growing constraints surrounding data access, privacy and ethical governance in HR management research. As organisations increasingly restrict access to employee data due to GDPR compliance (Machado et al., 2023), the empirical foundations of attrition research risk becoming narrow and repetitive, potentially limiting theoretical generalization.

Evidence from an HR perspective suggest that, data are not neutral inputs but shape the patterns, relationship and conclusions that emerge from analytical models (xxx, 20xx). Accordingly, understanding whether alternative data sources such as synthetic data can preserve meaningful attrition relationship is crucial for both theory development and practical application. Our study bridges this gap by empirically examining whether attrition predictors and predictive performance remain stable across real and synthetic datasets.

F. Data Constraints, Ethics, and the Emergence of Synthetic Data

A significant challenge in employee attrition research is the lack of access to high-quality employee data. Organisations often hesitate to share such information due to confidentiality, data protection regulations like GDPR, and ethical concerns. This results in "data poverty" in HR analytics, where a scarcity of raw data hinders the use of existing analytical techniques. Many studies rely on limited publicly available datasets, such as the IBM employee dataset, which includes only 1,470 observations. To address these sample-size limitations, researchers often use resampling techniques such as bootstrapping. While this can enhance statistical power, it does not introduce new information and may exacerbate existing biases. Moreover, repeated observations can distort model learning, especially in machine learning contexts where detecting minority-class events, such as attrition, is crucial. Synthetic data generation has emerged as a viable alternative that enables the creation of artificial datasets that are reflective of real data, thereby reducing privacy risks and data acquisition costs. Although synthetic data has been beneficial in other fields for robust modelling and hypothesis testing, its use in employee attrition research is still limited and under-theorised (Via et al., 2022). Specifically, there is insufficient empirical evidence on whether predictive models trained on synthetic data yield results comparable to those trained on real datasets across various modelling techniques. This gap is increasingly important given the rising adoption of machine learning approaches in strategic human resource management.

G. Research Gap

This study presents a conceptual framework that explores how demographic, job-related, and organisational factors influence employee attrition. It operationalises these influences through employee-level attributes such as age, job level, compensation, satisfaction, and work environment. Utilising both traditional statistical methods and machine learning techniques, the study aims to effectively predict attrition outcomes. A significant contribution of this research is the introduction of data type (real versus synthetic) as a factor that affects model performance. It goes beyond merely assessing predictive accuracy to investigate how different data-generating processes can influence the reliability and validity of attrition predictions across various modelling approaches. By comparing logistic regression, random forests, and gradient boosting trees across real and synthetic datasets, the study addresses a critical gap in HR analytics: the viability of synthetic data as a substitute for real employee data. This work enhances the understanding of data-driven decision-making in human resources, offering insights for research design, ethical data use, and managerial practices. In summary, data privacy limits machine learning in attrition research, highlighting a critical literature gap regarding the unproven reliability of synthetic data as a viable alternative for HR predictive modelling. This study seeks to address this gap by systematically comparing the performance of real versus synthetic data across various algorithms.

Methodology

This section outlines the research methodology used to compare the effectiveness of real and synthetic datasets for predicting employee attrition. The primary objective is to determine whether synthetic data, generated to preserve the statistical properties of real-world data, can serve as a reliable substitute for training machine learning models in human resource (HR) analytics. Additionally, this section outlines the research design, data acquisition sources, sampling techniques, and the operationalisation of key variables. Furthermore, it details the analytical procedures, including the specific machine learning algorithms selected, such as Binary Logistic Regression, Random Forest, and Gradient Boosting, and the statistical tests used to validate the hypotheses.

A. Research Design and Rationale

This study adopts a quantitative, non-experimental research design to systematically evaluate whether the performance of employee attrition predictive models differs when trained on real versus synthetic datasets. While traditional HR analytics research has heavily relied on observational data, this study emphasises the comparative utility of synthetic data. Recognising that prediction accuracy, data privacy, and model generalizability are paramount concerns in modern HR analytics, this design aligns with emerging research that compares modelling approaches and data sources within human resource contexts. Recent literature shows that machine learning techniques significantly improve the prediction of employee turnover compared to conventional statistical methods, facilitating better human resource decision-making (Park & Shaw, 2025; Konar et al., 2025). However, the limited availability of open-source HR data due to privacy issues presents a significant challenge. To address this, the study uses the IBM Human Resource Analytics Employee Attrition and Performance dataset as a benchmark and creates a synthetic dataset with identical statistical properties. Building on previous models such as Logistic Regression, Random Forest, and Gradient Boosting (Yang et al., 2020; Ghita & Francesco, 2025), the research systematically compares model performance across both real and synthetic data. Key factors examined include demographics, employment characteristics, and remuneration, with a focus on how these variables impact predictive accuracy in simulated environments.

B. Data Source and Sampling Construction

This study utilises the IBM HR Analytics Employee Attrition and Performance dataset from Kaggle, which is well-regarded for benchmarking attrition models and includes 1,470 employee records with 35 features related to demographics, employment, and satisfaction metrics (Fallucchi et al., 2018). The population encompasses the complete IBM employee base represented in this dataset. Since secondary data is being used, traditional inclusion and exclusion criteria are less applicable; the dataset has already been validated (PM & Balaji, 2018). To investigate the effectiveness of synthetic data in attrition prediction, the study created two expanded datasets to maintain sample size parity:

- Bootstrapped Real Dataset: This involves resampling the original dataset to produce 5,000 observations without altering the underlying data distribution.
- Synthetic Dataset: A similarly-sized dataset of 5,000 observations was generated using principled techniques, ensuring the preservation of the original data's distributional properties without replicating identifiable records. This approach aligns with ongoing research on privacy-preserving analytics in HR (Baydili & Tasci, 2025) and enables a controlled comparison of model performance based solely on data type (real vs. synthetic).

C. Data Variables and Operationalisation

The operationalisation of variables focuses on investigating the patterns that differentiate employees who leave the organisation from those who stay. The dataset includes 35 variables, which encompass numerical, categorical, and ordinal data types. The dependent variable is “Employee Attrition”, a binary indicator coded as 1 for employees who have left the organisation and 0 for those who have remained. This binary classification aligns with previous attrition modelling approaches found in the literature. Additionally, the independent variables are categorised into three primary domains, reflecting findings from prior studies that have identified these predictors as significant in predicting turnover (Awan et al., 2021; Cheng & Zhao, 2024; Saini et al., 2022; Liu & Batt, 2010):

- Demographic Factors: Age, Gender, Education Level, Marital Status.
- Job Characteristics: Job Level, Job Role, Overtime, Distance from Home.
- Attitudinal/Organisational Variables: Job Satisfaction, Environment Satisfaction, Work-Life Balance, Monthly Income.

D. Data Preprocessing, Data Analysis and Quality Assurance

Extensive preprocessing ensured data quality, confirming no missing values among the 1,470 entries. The dataset comprised 26 integer variables and 9 factor variables. To prepare for machine learning analysis, variables with no predictive value were removed: Employee Count, Over18, Standard Hours, and Employee Number. The target variable "Attrition" was relabeled from "Yes"/"No" to “1/0”, and variable names were standardised to improve code clarity (e.g., removing prefixes such as "Travel_"). After these cleaning steps, the dataset was ready for bootstrapping and synthetic generation. Furthermore, this study employs inferential statistics and supervised classification algorithms to analyse employee attrition data, focusing on three benchmark models: Binary Logistic Regression, Random Forest Classification, and Gradient Boosting Trees (Alpaydin, 2020).

- Binary Logistic Regression: Used as a baseline for its interpretable coefficients, allowing for the analysis of odds ratios for specific attrition drivers.
- Random Forest Classification: A tree-based ensemble method noted for capturing non-linear variable interactions and reducing overfitting through bootstrap aggregation.
- Gradient Boosting Trees: An iterative ensemble technique that enhances prediction accuracy by correcting errors from previous models, effective for identifying complex patterns.

These models were chosen for their proven effectiveness in the literature on employee attrition and for the rise of ensemble methods in HR predictions (Akintunde, 2024). Standard preprocessing techniques were applied to the real and synthetic datasets, including One-Hot Encoding for categorical variables and removing features with negligible variance to improve model performance. Given the class imbalance often present in attrition datasets, specific evaluation metrics were chosen to address this issue, following best practices in HR analytics (Căvescu & Popescu, 2025).

E. Ethical Considerations, Evaluation and Comparison

The study adheres to ethical standards by utilising anonymised secondary data. The inclusion of synthetic data enhances privacy protection, aligning with emerging standards in data governance and privacy-preserving research methodologies (Baydili & Tasci, 2025). Overall, this research contributes to the understanding of predictive modelling in HR analytics and offers a careful examination of data ethics and treatment methodologies. Moreover, each model was trained on both synthetic and real bootstrapped datasets, and its predictions were assessed against actual attrition rates. A multi-metric evaluation approach was used, reflecting recent trends in HR analytics, emphasising holistic performance assessment in Explainable AI (Konar et al., 2025). To rigorously test differences in model performance, a Chi-square test of independence was conducted for each model pair, addressing previous research gaps that focused solely on within-dataset comparisons. Based on the comparative study design, four hypotheses were formulated to investigate whether synthetically generated datasets produce performance metrics equivalent to those of real datasets:

- H1: There is a significant difference in attrition rates between synthetic and real IBM employee datasets.
- H2: The real dataset yields higher accuracy than the synthetic dataset in predicting attrition rates using logistic regression.
- H3: The real dataset yields higher accuracy than the synthetic dataset in predicting attrition rates using the gradient boosting trees model.
- H4: The real dataset yields higher accuracy than the synthetic dataset in predicting attrition rates using the random forest model.

In summary, this methodology evaluates real versus synthetic data interchangeability in attrition modelling using a quantitative design. Expanding the IBM dataset into comparable 5,000-observation samples via bootstrapping and synthetic generation, the study employs Logistic Regression, Random

Forest, and Gradient Boosting to benchmark performance, establishing a rigorous foundation for empirical analysis. Therefore, the findings of this study discussed below have implications beyond model performance comparison. Our findings suggest that employee attrition knowledge is not exclusively dependent on access to real organisational data but can be evocatively generated under alternative data conditions. This unique insight challenges dominant assumptions in attrition research and underscores the role of analytical design in shaping HR knowledge and decision-making.

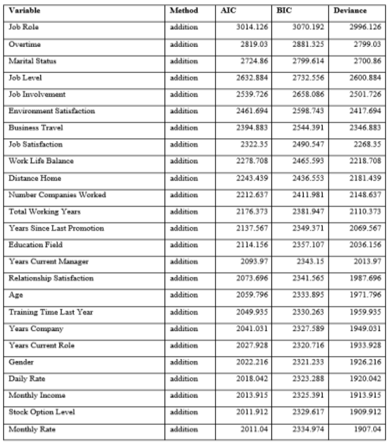
Results & Discussion

This study sets out to examine whether predictive models of employee attrition yield systematically different results when trained on real versus synthetically generated datasets. The findings provide several important insights that extend existing employee attrition and HR analytics research. Machine learning modelling techniques (Tranmer & Elliot 2008; Bhardwaj & Pal, 2012; Ali & Bader, 2021) were used to investigate factors that may lead to attrition of employees in a company. Three modelling techniques were used. The machine learning algorithms included binary logistic regression, random forest classification, and gradient boosting trees classification. The algorithms were trained on subsets of both synthetic and original data for easier comparison.

A. Binary Logistic Regression Model

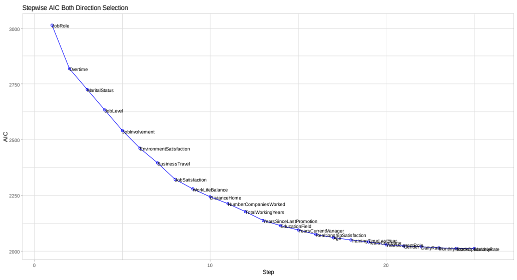
Table 1 was obtained for stepwise logistic regression model for original training subset. Model selection method used was stepwise logistic regression and using Akaike Information Criterion (AIC) as selection method. As factors leading to employee attrition were added, the AIC kept reducing until it could not reduce further. Variables such as job role, overtime, marital status, job level, job involvement, environment satisfaction, business travel, job satisfaction, work-life balance, and distance from home are the main factors employees use to decide whether to stay or quit their jobs.

Table 1 Original Dataset Stepwise Summary for Binary Logistic Regression Model



Variable	Method	AIC	BIC	Deviance
Job Role	addition	3014.126	3070.192	2996.126
Overtime	addition	2819.03	2881.325	2799.03
Marital Status	addition	2724.86	2799.614	2700.86
Job Level	addition	2632.884	2732.556	2600.884
Job Involvement	addition	2539.726	2658.086	2501.726
Environment Satisfaction	addition	2461.694	2598.743	2417.694
Business Travel	addition	2394.883	2544.391	2346.883
Job Satisfaction	addition	2322.35	2490.547	2268.35
Work Life Balance	addition	2278.708	2465.593	2218.708
Distance Home	addition	2243.439	2436.553	2181.439
Number Companies Worked	addition	2212.637	2411.981	2148.637
Total Working Years	addition	2176.373	2381.947	2110.373
Years Since Last Promotion	addition	2137.567	2349.371	2069.567
Education Field	addition	2114.156	2357.107	2036.156
Years Current Manager	addition	2093.97	2343.15	2013.97
Relationship Satisfaction	addition	2073.696	2341.565	1987.696
Age	addition	2059.796	2333.895	1971.796
Training Time Last Year	addition	2049.935	2330.263	1959.935
Years Company	addition	2041.031	2327.589	1949.031
Years Current Role	addition	2027.928	2320.716	1933.928
Gender	addition	2022.216	2321.233	1926.216
Daily Rate	addition	2018.042	2323.288	1920.042
Monthly Income	addition	2013.915	2325.391	1913.915
Stock Option Level	addition	2011.912	2329.617	1909.912
Monthly Rate	addition	2011.04	2334.974	1907.04

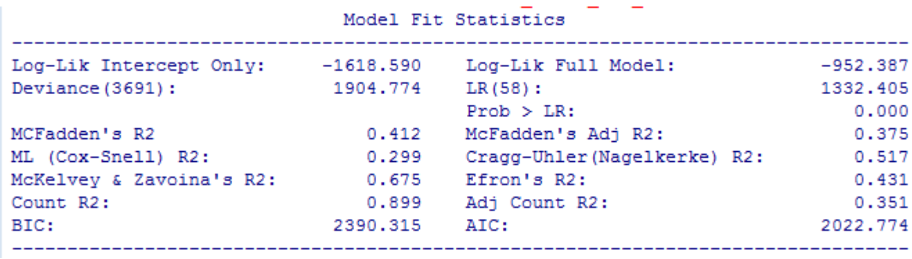
Figure 1 Stepwise Regression on Original Training Subset



The model, as observed in Table 1 and Figure 1, which summarises stepwise processes conducted, reveals that the highest AIC was 3014.13. At the same time, after adding all significant factors leading to attrition, the AIC reduced to 2011.04.

Additionally, the model results for model fit statistics show that at least one of the factors used in modelling predicts attrition rates among employees. Hence, the likelihood ratio (LR) is: LR (58) = 1332.41, p<.0001. The model also shows that about 89.9% of the total variations of attrition rates among employees would be explained in the model.

Table 2 Model Fit Statistics for Original Data.



```
                               Model Fit Statistics
--------------------------------------------------------------------------------
Log-Lik Intercept Only:      -1618.590     Log-Lik Full Model:             -952.387
Deviance(3691):               1904.774     LR(58):                         1332.405
                                           Prob > LR:                         0.000
MCFadden's R2                    0.412     McFadden's Adj R2:                 0.375
ML (Cox-Snell) R2:               0.299     Cragg-Uhler(Nagelkerke) R2:        0.517
McKelvey & Zavoina's R2:         0.675     Efron's R2:                        0.431
Count R2:                        0.899     Adj Count R2:                      0.351
BIC:                          2390.315     AIC:                            2022.774
--------------------------------------------------------------------------------
```

From the association of predicted probabilities in Table 2, it can be concluded that the model can predict about 90.24% of attrition rate variations, leaving out only 9.76% that were not correctly predicted. This is a high accuracy of prediction hence the best model.

Figure 2 Stepwise Regression on Synthetic Training Subset.

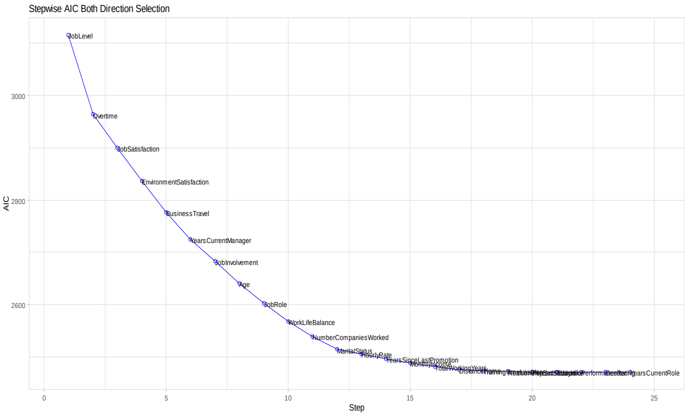
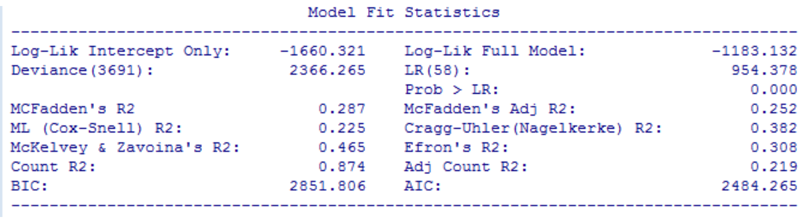


Figure 2 shows all the factors added to the regression model, starting with job level and completing with the current role.

Table 3 Model Fit Statistics for Synthetic Data.



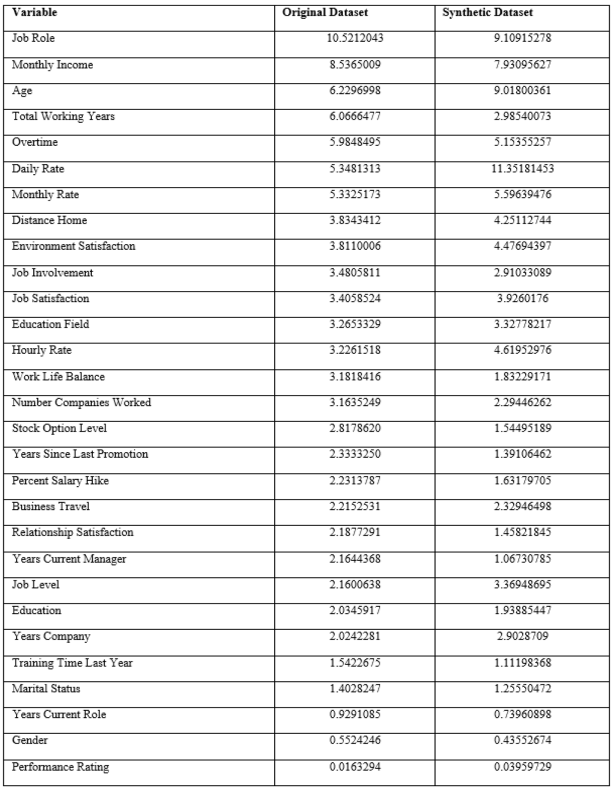
```
                           Model Fit Statistics
--------------------------------------------------------------------------------
Log-Lik Intercept Only:      -1660.321     Log-Lik Full Model:            -1183.132
Deviance(3691):               2366.265     LR(58):                          954.378
                                           Prob > LR:                         0.000
MCFadden's R2                    0.287     McFadden's Adj R2:                 0.252
ML (Cox-Snell) R2:               0.225     Cragg-Uhler(Nagelkerke) R2:        0.382
McKelvey & Zavoina's R2:         0.465     Efron's R2:                        0.308
Count R2:                        0.874     Adj Count R2:                      0.219
BIC:                          2851.806     AIC:                            2484.265
--------------------------------------------------------------------------------
```

As seen in Table 3, results through the model fit statistics summary above show that at least one of the factors was significant in predicting attrition rates among employees, LR (58) = 954.38, p<.0001. It was noted that the synthetic logistic regression model was better than the original dataset since it had more factors explaining employee attrition rates than their counterparts.

B. Gradient Boosting Trees Classification

Another model used to investigate differences in the distributions of attrition rate predictions by the two datasets was through use of gradient-boosting trees. The models show a significant difference in that for original dataset; job role plays a more significant role on employee attrition rates than any other factor followed by their monthly income. On the other hand, daily rates contribute immensely to attrition rates among simulated datasets.

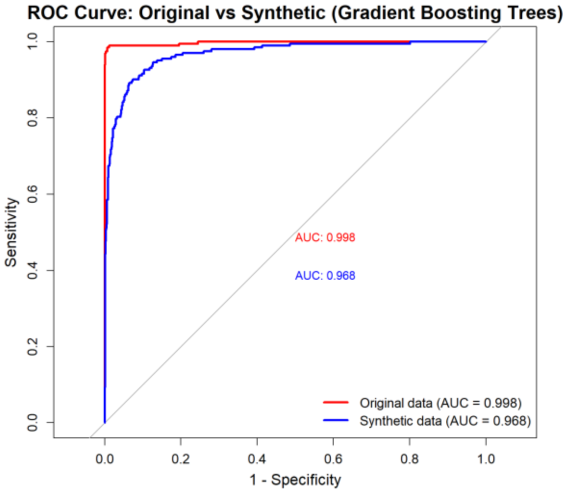
Table 4 Xgboost Regression Variable Importance Results.



Variable	Original Dataset	Synthetic Dataset
Job Role	10.5212043	9.10915278
Monthly Income	8.5365009	7.93095627
Age	6.2296998	9.01800361
Total Working Years	6.0666477	2.98540073
Overtime	5.9848495	5.15355257
Daily Rate	5.3481313	11.35181453
Monthly Rate	5.3325173	5.59639476
Distance Home	3.8343412	4.25112744
Environment Satisfaction	3.8110006	4.47694397
Job Involvement	3.4805811	2.91033089
Job Satisfaction	3.4058524	3.9260176
Education Field	3.2653329	3.32778217
Hourly Rate	3.2261518	4.61952976
Work Life Balance	3.1818416	1.83229171
Number Companies Worked	3.1635249	2.29446262
Stock Option Level	2.8178620	1.54495189
Years Since Last Promotion	2.3333250	1.39106462
Percent Salary Hike	2.2313787	1.63179705
Business Travel	2.2152531	2.32946498
Relationship Satisfaction	2.1877291	1.45821845
Years Current Manager	2.1644368	1.06730785
Job Level	2.1600638	3.36948695
Education	2.0345917	1.93885447
Years Company	2.0242281	2.9028709
Training Time Last Year	1.5422675	1.11198368
Marital Status	1.4028247	1.25550472
Years Current Role	0.9291085	0.73960898
Gender	0.5524246	0.43552674
Performance Rating	0.0163294	0.03959729

Table 4 shows that there exist disparities in the two datasets despite creating a synthetic dataset from the original employee dataset. Additionally, other differences in terms of variable importance are summarised in Table 4.

Figure 3 ROC curve for Gradient Boosting Trees.



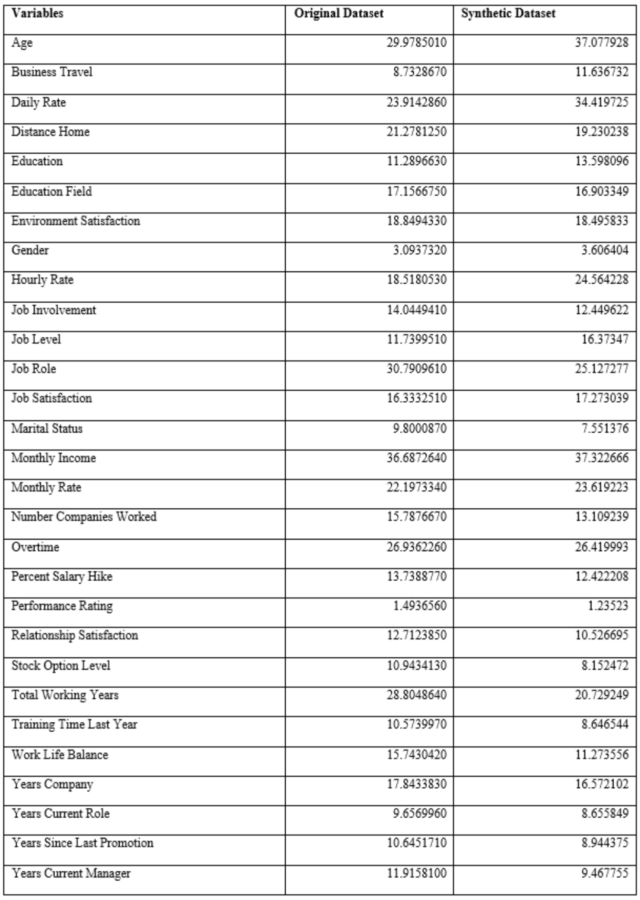
As shown in Figure 3, the ROC curve was generated to compare the predictive performance of the Gradient Boosting Trees model for both original and synthetic dataset. The ROC curve shows the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate) across different classification thresholds.

Both models showed excellent discrimination ability, with the original dataset exhibiting a slightly higher AUC = 0.99 compared to the synthetic data model (AUC = 0.97). The high AUC value indicates that both models can accurately distinguish between employees who were attrited and those who did not. The marginal difference in AUC (0.03) suggests that the synthetic dataset preserved most of the underlying structural and statistical relationships present in the original data, thereby supporting the reliability of the synthetic data for modelling purposes.

C. Random Forest Classification

Random forest classification models were also built to investigate significant differences in the two datasets, such as original and simulated datasets.

Table 5 Random Forest Variable Importance Results.



Variables	Original Dataset	Synthetic Dataset
Age	29.9785010	37.077928
Business Travel	8.7328670	11.636732
Daily Rate	23.9142860	34.419725
Distance Home	21.2781250	19.230238
Education	11.2896630	13.598096
Education Field	17.1566750	16.903349
Environment Satisfaction	18.8494330	18.495833
Gender	3.0937320	3.606404
Hourly Rate	18.5180530	24.564228
Job Involvement	14.0449410	12.449622
Job Level	11.7399510	16.37347
Job Role	30.7909610	25.127277
Job Satisfaction	16.3332510	17.273039
Marital Status	9.8000870	7.551376
Monthly Income	36.6872640	37.322666
Monthly Rate	22.1973340	23.619223
Number Companies Worked	15.7876670	13.109239
Overtime	26.9362260	26.419993
Percent Salary Hike	13.7388770	12.422208
Performance Rating	1.4936560	1.23523
Relationship Satisfaction	12.7123850	10.526695
Stock Option Level	10.9434130	8.152472
Total Working Years	28.8048640	20.729249
Training Time Last Year	10.5739970	8.646544
Work Life Balance	15.7430420	11.273556
Years Company	17.8433830	16.572102
Years Current Role	9.6569960	8.655849
Years Since Last Promotion	10.6451710	8.944375
Years Current Manager	11.9158100	9.467755

The results revealed that monthly income contributes more to attrition rates than any other attribute in the original dataset, unlike the simulated, which shows that age is the main factor influencing attrition rates. The variable importance results are summarised in Table 5.

Figure 4 ROC curve for Random Forest Classification.

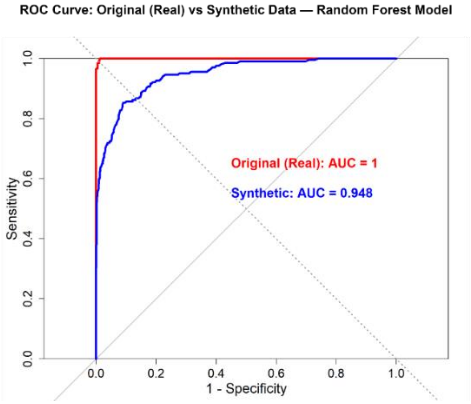


Figure 4 shows the performance of the Random Forest models evaluated using ROC analysis for both the original and synthetic datasets. The original dataset achieved an AUC of 1, indicating near perfect classification performance while the synthetic dataset achieved an AUC of 0.95 which still reflects a high degree of predictive accuracy. Overall, these results suggest that the synthetic dataset provides a reliable approximation of the real data for modelling employee attrition.

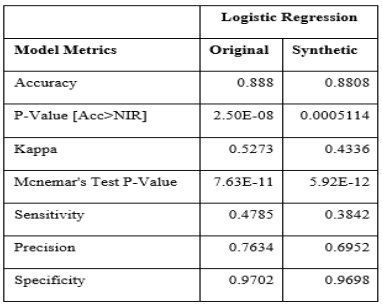
Model Evaluation

This study aimed to use machine learning classification algorithms to investigate whether there is any statistically significant difference in the prediction of attrition rates among employees from the original dataset and synthetic dataset obtained from the real dataset used. In an attempt to investigate significant differences, four hypotheses were used. To unmask differences in predicted attrition rates from real and synthetic datasets; machine learning classification algorithms such as binary logistic regression, gradient boosting trees, and random forest analyses algorithms were used to train models and then evaluated using test subsets of real and synthetic datasets. The attrition rates obtained from each model presented only show each prediction's accuracy but do not compare real dataset predicted values to those of synthetic dataset predictions. To intuitively assess differences, two-way chi-square analyses were used for each pair of predicted attrition rates in each of the three models. The hypotheses were therefore evaluated using chi-square analyses. The following provides a framework on how each research question formulated from the four hypotheses was tested. As model evaluation was the central area of concern in this research project, the model metrics utilised to assess the performance of each model for both original and synthetic datasets are accuracy, precision, Kappa statistics and sensitivity.

A. Evaluation of Logistics Regression Between Original Data and Synthetic Data

The research question aimed at investigating whether or not the logistic regression modelling technique would significantly differ in predicting employee attrition rates between real dataset and simulated dataset. The null hypothesis stated that there is no significant association in employee attrition rates from logistic regression models built from real and simulated datasets. On the contrary, the alternative hypothesis claimed that there is a significant association in the employee attrition rates of logistic regression models built from real and synthetic datasets.

Table 6 Evaluation of Logistics Regression.

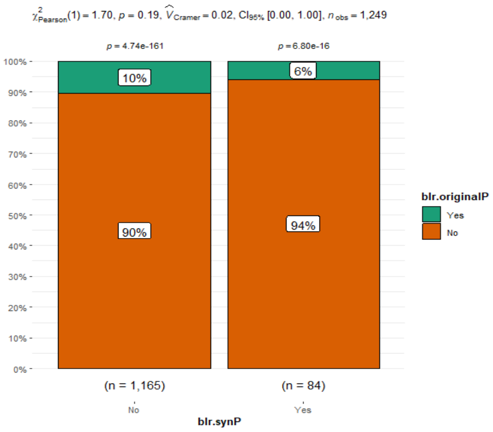


	Logistic Regression	
Model Metrics	Original	Synthetic
Accuracy	0.888	0.8808
P-Value [Acc>NIR]	2.50E-08	0.0005114
Kappa	0.5273	0.4336
Mcnemar's Test P-Value	7.63E-11	5.92E-12
Sensitivity	0.4785	0.3842
Precision	0.7634	0.6952
Specificity	0.9702	0.9698

As seen in Table 6, the two-way chi-square results were non-significant, Chi-square (1) = 1.70, p=.19, V=.002. It is concluded that at 5% level of significance, there is no association of employee attrition rates between real dataset and simulated dataset. Research question 1 was therefore not rejected and concluded that there is no notable significant association in performances between real and simulated datasets from logistic regression model.

In relation to the model metrics, logistic regression showed very similar levels of accuracy between the real and synthetic datasets (0.89 and 0.88 respectively). This supports the non-significant chi-square result, suggesting consistent performance across data types. The Kappa statistics slightly decreased from 0.53 to 0.43, indicating a minor reduction in classification agreement. Sensitivity reduced from 0.48 to 0.38, implying the synthetic data model was less effective in identifying employees who actually left, while specificity remained very high (0.97) for both datasets, showing that non-attrition employees were correctly identified. Precision also decreased slightly (0.76 to 0.69), meaning the model generated more false positives for attrition under synthetic data. Overall, logistic regression performance was stable across datasets, though the real dataset offered slightly more balanced predictive reliability.

Figure 5 Testing Difference in Logistic Regression Prediction between Original and Synthetic Datasets.

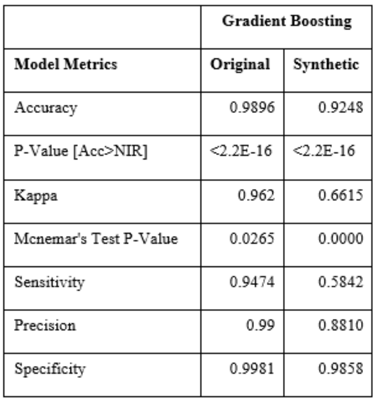


As seen in Figure 5, the distribution of predicted attrition (“Yes”) and non-attrition (“No”) cases is very similar across both data types. The chi-square test shows no statistically significant difference between the two predictions ($x^2(1) = 1.70$, p = .19).

B. Evaluation of Gradient Boosting Between Original Data and Synthetic Data

Chi-square test of independence was used to investigate the hypothesis with the null hypothesis stating that there is no association in gradient boosting trees classification model performances obtained from real and synthetic datasets. On the contrary, the alternative hypothesis stated that a statistically significant association exists in gradient boosting trees classification model performances predicted from real and artificial datasets. The results were inestimable since only employees who left the company were predicted to leave out all employees who chose to stay.

Table 7 Evaluation of Gradient Boosting.



	Gradient Boosting	
Model Metrics	**Original**	**Synthetic**
Accuracy	0.9896	0.9248
P-Value [Acc>NIR]	<2.2E-16	<2.2E-16
Kappa	0.962	0.6615
Mcnemar's Test P-Value	0.0265	0.0000
Sensitivity	0.9474	0.5842
Precision	0.99	0.8810
Specificity	0.9981	0.9858

The model evaluation metrics shown in Table 7, however, show that the gradient boosting tree model performed best overall among all algorithms tested. Accuracy for the real dataset was exceptionally high at 0.99 compared to 0.93 for the synthetic dataset. Kappa decreased from 0.96 to 0.66, suggesting moderate reliability when applied to synthetic data. Sensitivity declined from 0.95 to 0.58, while specificity remained very high (0.99 for real and 0.98 for synthetic). Precision also remained strong, decreasing only slightly from 0.99 to 0.88. These results suggest that while gradient boosting retained high predictive performance and stability, synthetic data led to a noticeable reduction in correctly identifying true attrition cases, likely due to minor variations in class representation.

Figure 6 Testing Difference in Xgboost Trees Prediction between Original and Synthetic Datasets.

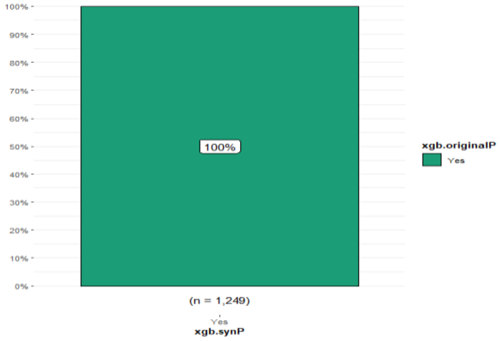
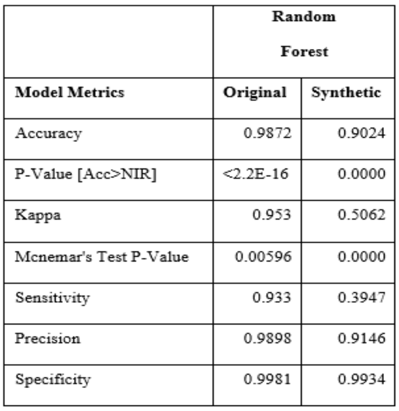


Figure 6 shows that XGboost model predicts all observations as attrition ("Yes") for both the original and synthetic datasets. Due to the lack of variability in predicted outcomes, a chi-square test of independence could not be estimated. This limits statistical comparison of XGBoost predictions between the two data types.

C. Evaluation of Random Forest Between Original Data and Synthetic Data

Chi-square test of independence was used to investigate the hypothesis with the null hypothesis stating that there is no association in random forest classification model performances obtained from real and synthetic datasets. On the contrary, the alternative hypothesis stated that a statistically significant association exists in random forest classification model performances predicted from real and artificial datasets.

Table 8 Evaluation of Random Forest.



	Random Forest	
Model Metrics	**Original**	**Synthetic**
Accuracy	0.9872	0.9024
P-Value [Acc>NIR]	<2.2E-16	0.0000
Kappa	0.953	0.5062
Mcnemar's Test P-Value	0.00596	0.0000
Sensitivity	0.933	0.3947
Precision	0.9898	0.9146
Specificity	0.9981	0.9934

As seen in Table 8, the results were non-significant, Chi-square (1) = .34, p=.56. This signifies a non-existent association between performances of random forest classification models from synthetic and real datasets at 5% level of significance. The random forest model maintained a high level of accuracy across both datasets (0.99 for real and 0.90 for synthetic), consistent with the chi-square result of no significant difference. The Kappa statistics reduced from 0.95 to 0.51, indicating a moderate drop in classification consistency. Sensitivity decreased from 0.93 to 0.39, suggesting that attrition detection was less effective with synthetic data, while specificity remained near perfect (0.99 versus 0.99). Precision declined slightly from 0.99 to 0.92. These differences reveal that although the random forest model remained highly accurate, its ability to detect minority-class attrition events was influenced by the representational quality of the synthetic dataset.

Figure 7 Testing Difference in Random Forest Prediction between Original and Synthetic Datasets.

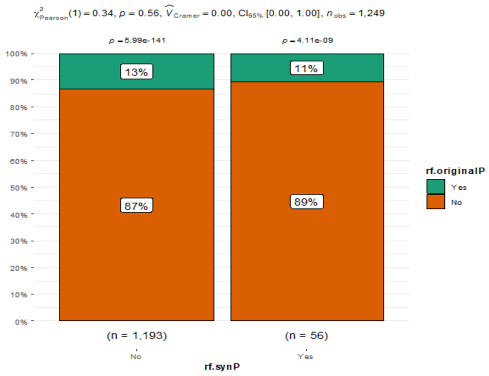


Figure 7 compares random forest attrition predictions between the original and synthetic datasets. The distribution of predicted attrition ("Yes") and non-attrition ("No") cases is highly similar across both data types. The chi-square test shows no statistically significant difference between the predictions (x2(1) = 0.34, p = .56).

Furthermore, the results indicate that synthetic datasets preserve the core predictive structure of the real employee data. Key determinants of employee attrition, including job level, age, job satisfaction, environmental satisfaction, income, and commuting distance, emerged consistently across both data types. This convergence suggests that synthetic data, when properly generated, can retain meaningful relationships between employee characteristics and attrition outcomes rather than merely replicating marginal distributions. It results further reveal that the comparative evaluation of model performance reveals no statistically significant differences in overall predictive accuracy between real and synthetic datasets for logistic regression and random forest models. While gradient boosting trees exhibited marginal reductions in minority-class sensitivity when applied to synthetic data, overall discrimination power remained high. These findings suggest that synthetic data can support robust predictive modelling, albeit with some caution regarding the detection of rare attrition events—a limitation that aligns with broader concerns in machine learning research on class imbalance and data representation. Lastly, the results demonstrate that model choice matters less than data representativeness when predicting employee attrition. While ensemble methods outperformed logistic regression in terms of raw predictive accuracy, the relative stability of performance across real and synthetic datasets underscores that the integrity of the underlying data-generating process is a critical determinant of model reliability.

Conclusion

Using machine learning classification modelling techniques, this research project investigated whether a relationship exists between predicted employee attrition from real datasets and synthetic datasets. The modelling techniques used included binary logistic regression models, gradient boosting classification, and random forest classification. Since getting employee data is usually challenging and most companies are unwilling to give out their employee data, the project aimed at assessing the predictability of employee attrition from real and synthetic datasets to investigate whether or not synthetic datasets could be used in building models. Like any other research, this research had few limitations as it was non-experimental research. The main research limitation was scarcity of employee dataset which saw us using IBM employee dataset that has limited data points as only 1470 employees' data are in the data sample. The research then increased this sample by bootstrapping to 5000 observations. The other concern regarding bootstrapped data is that it tends to have repeated measurements which may not reflect the true distribution of a variable or datapoints. For instance, out of 1470 observations, an increased data to 5000 tripled observations hence the likelihood of counting the same characteristic more than three times, which increases data constant ability and may cause misleading results. Another notable challenge was regarding other variables in the dataset since they could easily model some machine learning algorithms but return errors when using other models. Examples of such variables included performance ratings showing some data partitioning resulted to only two levels such as excellent and outstanding captured in the data subsets leading to errors in some models. Other variables, such as daily rates, were constant throughout the data and hence could not be included in the model.

A. Key contributions and Limitations

Following the discussion of the results, the paper explicitly outlines its theoretical, methodological and practical contributions to employee attrition and HR analytics research.

Existing attrition research implicitly assumes that findings derived from real employee datasets are inherently more valid than those based on alternative data sources. By demonstrating that synthetically generated data can yield comparable predictive outcomes, this study challenges this assumption and extends employee attrition research into the methodological domain. The results suggest that attrition-related relationships are structurally robust rather than artefacts of a specific dataset, indicating that the theoretical mechanisms underlying employee withdrawal decisions may generalise beyond individual-level observations. In doing so, the study also advances HR analytics theory beyond a narrow focus on predictive accuracy. Rather than emphasising how accurately attrition can be predicted, the findings shift attention toward the data conditions under which such predictions remain reliable. By positioning data type as a moderating methodological variable, this research contributes to a growing body of literature that conceptualises HR analytics as a strategic capability embedded within organisational decision-making processes, rather than as a purely technical exercise. Furthermore, by empirically evaluating synthetic data as a substitute for real employee data, the study contributes to emerging debates on ethical analytics and data governance. The findings indicate that organisations can pursue data-driven HR decision-making while mitigating privacy risks and regulatory constraints, thereby advancing theory at the intersection of human resource management, information systems, and organisational governance. From a practical perspective, the results suggest that organisations facing data access restrictions can confidently leverage synthetic data for attrition modelling, workforce planning, and talent management. However, predictive outputs should be used as decision-support tools rather than deterministic forecasts, particularly given reduced sensitivity for minority attrition cases. Finally, this study is subject to limitations, including reliance on a single dataset and distribution-based synthetic data generation.

B. Directions for Future Research

Future analyses could benefit from using non-bootstrapped datasets to enhance the validity of empirical findings. Incorporating basic employee factors will allow for a comprehensive assessment of outcomes. Additionally, employing various comparison methods, such as predicting employee attrition probabilities and evaluating differences in average attrition rates with paired t-tests, should be explored. The paired t-test is suitable here as the distribution of attrition probabilities ranges from 0 to 1, making it easy to validate its assumptions prior to testing. Human resource managers often struggle to assess employees based on their characteristics. To mitigate the resources spent on replacing outgoing employees, a systematic approach to evaluating prospective hires is essential. Future research should shift focus from viewing employee attrition as a dependent variable to considering job satisfaction measured on the Likert scale and attrition as the main independent variable. This approach will allow for a better analysis of how job satisfaction influences attrition, providing insights into employee preferences and motivations. research could employ multi-organisational data, causal generative techniques such as GANs, and probability-based predictions to further strengthen theoretical inference and practical relevance.

REFERENCES

Ajunwa, I. (2020). The Paradox of Automation as Anti-Bias Intervention. Cardozo Law Review, 41(5), 1671–1726.

Akintunde Adetoye Fadare (2024). Prediction of HR Employee Attrition with Machine Learning:

Ali, W. and Bader, A., 2021. Prediction of Employee Turn Over Using Random Forest Classifier with Intensive Optimized Pca Algorithm. *Wireless Personal Communications*, *119*(4), pp.3365-3382.

Allen, M., Titsworth, B., & Thompson, P. (2021). A Systematic Review of Employee Attrition Prediction using Machine Learning. Journal of Human Resources Management Research, 26(1), 1-17.

Alnachef, T.H. and Alhajjar, A.A., 2017. Effect of human capital on organizational performance: A literature review. *International Journal of Science and Research*, *6*(8), pp.1154-1158.

Alpaydin, E., 2020. *Introduction to machine learning*. MIT press.

Austin-Egole, I.S., Iheriohanma, E.B. and Nwokorie, C., 2020. Flexible working arrangements and organizational performance: An overview. *IOSR Journal of Humanities and Social Science (IOSR-JHSS)*, *25*(5), pp.50-59.

Awan, K., Ahmad, N., Naveed, R. T., Scholz, M., Adnan, M., & Han, H. (2021). The impact of work–family enrichment on subjective career success through job engagement: A case of banking sector. Sustainability, 13(16), 8872. https://doi.org/10.3390/su13168872

Bagging and Random Forest Application. INTERNATIONAL JOURNAL OF RESEARCH AND SCIENTIFIC INNOVATION. Vol. XI Issue VIII. Pg. 410 DOI: 10.51244. Available at www.rsisinternational.org

Ballinger, G., Craig, E., Cross, R. and Gray, P., 2011. A stitch in time saves nine: Leveraging networks to reduce the costs of turnover. California Management Review, 53(4), pp.111-133.

Baydili, İ. T., & Tasci, B. (2025). Predicting Employee Attrition: XAI-Powered Models for Managerial Decision-Making. *Systems*, *13*(7), 583. https://doi.org/10.3390/systems13070583

Bhardwaj, B.K. and Pal, S., 2012. Data Mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*.

Cascio, W.F., 2007. The Costs-and Benefits-of Human Resources. *International Review of Industrial and Organizational Psychology 2007*, pp.71-110.

Castaldo, S., Ciacci, A., & Penco, L. (2022). Perceived corporate social responsibility and job satisfaction in the retail industry: A systematic literature review and research agenda. International Series in Advanced Management Studies, 33-55. https://doi.org/10.1007/978-3-031-12027-5_3

Căvescu, A. M., & Popescu, N. (2025). Predictive Analytics in Human Resources Management: Evaluating AIHR’s Role in Talent Retention. Applied Math, 5(3), 99. https://doi.org/10.3390/appliedmath5030099

Cheng Yuxiang, & Zhao Xiaomiao. (2024). Enhancing Employee Retention Strategies Through Advanced Predictive Analytics. *JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING (JRTCSE)*, *12*(1), 1-5. https://jrtcse.com/index.php/home/article/view/JRTCSE.2024.1.1

Dess, G.G. and Shaw, J.D., 2001. Voluntary turnover, social capital, and organizational performance. *Academy of management review*, *26*(3), pp.446-456.

Duhautois, R., Gilles, F. and Petit, H., 2016. Decomposing the relationship between wage and churning. *International Journal of Manpower*.

Fallucchi, F., Coladangelo, M., Giuliano, R. and William De Luca, E., 2020. Predicting employee attrition using machine learning techniques. *Computers*, *9*(4), p.86.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

Ghita Regasse, Francesco Venier, (2025). Implementing machine learning for predictive analytics: An empirical study of employee turnover, Volume 2, Issue 4, https://doi.org/10.1016/j.nexres.2025.100873.

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. Journal of Management, 26(3), 463-488.

Hatch, N.W. and Dyer, J.H., 2004. Human capital and learning as a source of sustainable competitive advantage. *Strategic management journal*, *25*(12), pp.1155-1178.

Hom, P. W., Mitchell, T. R., Lee, T. W., & Griffeth, R. W. (2012). Reviewing the attrition literature: The role of data, theory, and research design. Journal of Management, 38(3), 1032-1053.

Hytter, A., 2007. Retention strategies in France and Sweden. *Irish Journal of Management*, *28*(1).

Jordon, J., Yoon, J., & van der Schaar, M. (2018). PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. International Conference on Learning Representations (ICLR) Workshop.

Kaggle 2025. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Kamoche, K., 1996. Strategic human resource management within a resource‐capability view of the firm. *Journal of Management studies*, *33*(2), pp.213-233.

Katsande, I.T.E. and Chisoro, L., An Investigation into the Impact of Management Style on Employee Motivation at a South African Consulting Firm. International Journal of Human Resources Management (IJHRM), 7(6), pp.1-12.

Keramati, A. and Ardabili, S.M., 2011. Churn analysis for an Iranian mobile operator. *Telecommunications Policy*, *35*(4), pp.344-356.

Kerr, A., 2018. Job flows, worker flows and churning in South Africa. *South African Journal of Economics*, *86*, pp.141-166.

Kingsley, J., 2025. The hidden cost of attrition: predictive modeling as a retention strategy.

Kiran, P.R., Chaubey, A. and Shastri, R.K., 2024. Role of HR analytics and attrition on organisational performance: a literature review leveraging the SCM-TBFO framework. *Benchmarking: An International Journal*, *31*(9), pp.3102-3129.

Kuhn, P. and Yu, L., 2021. How costly is turnover? Evidence from retail. *Journal of Labor Economics*, *39*(2), pp.461-496.

Lavoie-tremblay, M.é.l.a.n.i.e., O'brien-pallas, l.i.n.d.a., Viens, C., Brabant, Macey, W. H., & Schneider, B. (2008). The meaning of employee engagement. Industrial and Organizational Psychology, 1(1), 3-30.

Machado, P., Vilela, J., Peixoto, M., & Silva, C. (2023, May). A systematic study on the impact of GDPR compliance on organizations. In *Proceedings of the XIX Brazilian Symposium on Information Systems* (pp. 435-442).

Milkovich, George T., Jerry M. Newman, and Carolyn Milkovich. *Compensation*. Nova Iorque: McGraw-Hill/Irwin, 2014.

Modau, F.D., Dhanpat, N., Lugisani, P., Mabojane, R. and Phiri, M., 2018. Exploring employee retention and intention to leave within a call centre. *SA Journal of Human Resource Management*, *16*(1), pp.1-13.

Muller, M., Shami, N.S., Guha, S., Masli, M., Geyer, W. and Wild, A., 2016,

Muzammil, A., Mir, M.M., Tunio, M.K. and Jariko, M.A., 2025. Talent development as a tool to achieve competitive advantage through organizational culture: A moderation and mediated model to prove the relation. *Ijss*, *4*(3), pp.36-52.

Nasir, S.Z. and Mahmood, N., 2016. HRM practices for employee retention: an analysis of Pakistani companies. *European Journal of Business and Management*, *8*(30), pp.96-104.

Pirrolas, O.A.C. and Correia, P.M.A.R., 2021. The Theoretical-Conceptual Model of Churning in Human Resources: The Importance of Its Operationalization. *Sustainability*, *13*(9), p.4770.

Pirrolas, O.A.C. and Correia, P.M.A.R., 2022. Literature review on human resource churning—theoretical framework, costs and proposed solutions. *Social Sciences*, *11*(10), p.489.

PM, U. and Balaji, N.V., 2019. Analysing Employee attrition using machine learning. *Karpagam Journal of Computer Science*, *13*, pp.277-282.

Ponnuru, S.R., Merugumala, G.K., Padigala, S., Vanga, R. and Kantapalli, B., 2020. Employee attrition prediction using logistic regression. *International Journal for Research in Applied Science and Engineering Technology*, *8*(5), pp.2871-2875.

Rousselet, G., Pernet, C.R. and Wilcox, R.R., 2023. An introduction to the bootstrap: a versatile method to make inferences by using data-driven simulations. *Meta-Psychology*, *7*.

Saini, K., Nimje, M. and Kalra, S., 2022. *Analysis of employee attrition and strategies for employee retention in the it sector* (Doctoral dissertation).

Setiyani, A., Djumarno, D., Riyanto, S. and Nawangsari, L., 2019. The effect of work environment on flexible working hours, employee engagement and employee motivation. *International review of management and marketing*, *9*(3), pp.112-116.

Shankar, R.S., Rajanikanth, J., Sivaramaraju, V.V. and Murthy, K.V.S.S.R., 2018, July. Prediction of employee attrition using datamining. In *2018 ieee international conference on system, computation, automation and networking (icscan)* (pp. 1-8). IEEE.

Shockley, K.M. and Allen, T.D., 2012. Motives for flexible work arrangement use. *Community, Work & Family*, *15*(2), pp.217-231

Silva, M.R.A., de Amorim Carvalho, J.C. and Dias, A.L., 2019. Determinants of employee retention: a study of reality in Brazil. In *Strategy and Superior Performance of Micro and Small Businesses in Volatile Economies* (pp. 44-56). IGI Global.

Skelton, A. R., Nattress, D., & Dwyer, R. J. (2019). Predicting manufacturing employee turnover intentions. Journal of Economics, Finance and Administrative Science, 25(49), 101-117. https://doi.org/10.1108/jefas-07-2018-0069

Talha, M., 2013. A study of flexible working hours and motivation. *Asian Social Science*.

Timsina, S., 2024. Employee Turnover and Engagement Programs for Retention.

Tranmer, M. and Elliot, M., 2008. Binary logistic regression. *Cathie Marsh for census and survey research, paper*, *20*.

Using Sampling Techniques. In *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1-6). IEEE.

Via, M., Chen, G., Canonaco, F., Daellenbach, K.R., Chazeau, B., Chebaicheb, H.,

Vinoodhini, D., 2022, April. Effective Classification of IBM Hr Analytics Employee Attrition

Wilkens, M., 2020. Employee churn in after-school care: Manager influences on retention and turnover. *Journal of Youth Development*, *15*(1), pp.94-121.

Yang, S. and Islam, M.T., 2020. IBM Employee Attrition Analysis. *arXiv preprint arXiv:2012.01286*.