International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Employee Attrition Prediction for Human Resources Management Using Career-Dynamics Features and Block-Penalised Logistic Learning: A Protocol-Matched Comparison

Aseel Hadi Mahmood
Department of Quality Assurance Management,
Ministry of Higher Education and Scientific Research, Baghdad, Iraq
aseelhadi@gmail.com

Abstract. Employee attrition is costly, and machine-learning studies on the public IBM HR Analytics dataset report accuracies from about 82% to above 97%. Much of this spread reflects evaluation design: several studies oversample the whole dataset before splitting it, so synthetic copies of test employees enter training. This study proposes a model that adds sixteen career-dynamics indicators, encodes ordinal survey items by category and fits a logistic learner with block-wise penalties, with all preprocessing fitted on training data only. Under ten repetitions of stratified 10-fold cross-validation it reached an AUC of 0.863, accuracy of 89.16% and a Matthews correlation coefficient of 0.549, significantly above eleven baseline classifiers ($p < 0.001$ for AUC), and it ranked first in AUC, F1 and MCC on an untouched test split. Reproducing the oversample-then-split designs of earlier studies, the framework exceeded the accuracy and F1 each reported, reaching 95.46% accuracy and an AUC of 0.989, while moving oversampling inside the training folds cut tree-ensemble accuracy by 5.8 to 8.9 points.

Keywords: human resources management; employee attrition; turnover prediction; data leakage; logistic regression; feature engineering

1. Introduction

Voluntary turnover removes trained staff and forces repeated recruitment, so human resources management (HRM) units increasingly use predictive models to flag employees at risk of leaving. The IBM HR Analytics Employee Attrition dataset is the common benchmark for this task, and studies have compared naive Bayes and support vector machines [1], logistic regression [2], [3] and deep temporal networks [4] on it, reporting accuracies from about 82% to above 97%.

The benchmark is small (1,470 employees) and imbalanced (16.1% leavers), so results depend heavily on evaluation choices. Several studies balance the whole dataset with synthetic minority oversampling before the train-test split [5], [6]; synthetic leavers are then interpolated from records that later sit in the test set, a form of data leakage known to inflate performance [7]. Such figures are not comparable with those from untouched test data, and leakage-free evaluation on this benchmark has only recently been examined [8].

This paper contributes (i) an attrition model combining sixteen career-dynamics indicators, category-level encoding of ordinal survey items and a block-penalised logistic learner whose coefficients managers can read as risk drivers; (ii) a benchmark against eleven classifiers under repeated cross-validation and an untouched hold-out split, and against prior studies under their own protocols; and (iii) a measurement of how far oversampling before the split inflates accuracy.

2. Related Work

Fallucchi et al. [1] found linear SVC most accurate (87.9%) on a 70:30 split, though it recalled only a quarter of leavers. Logistic regression was best in Qutub et al. [2] (88.43%, AUC 0.859) and in Guerranti and Dimitri [3] under 5-fold cross-validation (87.96%, AUC 0.850). Al-Darraji et al. [9] reached 91.16% with a deep network on one 70:30 split and 89.11% under 10-fold cross-validation. Mansor et al. [10] and Pratt et al. [11] reported 87.76% (SVM) and 85.12% (random forest). Mortezapour Shiri et al. [4] obtained 89.65%

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

accuracy, F1 0.616 and AUC 0.776 with a temporal convolutional ensemble, and Căvescu and Popescu [8] reached 87% with XGBoost using in-fold SMOTE.

Studies that balance the data before splitting report 92–94.5% [5], [6], [12], as does the ADASYN variant in [9]. Two studies report 97.5% and 97.7% without such balancing [13], [14], and one an AUC of 0.987 with a sensitivity of 0.43 [15]; these lie about eight points above every other leakage-free result and are not used as references. Table 1 summarises the comparable studies.

TABLE 1. Comparison of the state-of-the-art work.

Study	Model	Evaluation	Acc. (%)	F1	AUC
Leakage-free evaluation (original class distribution)					
[1], 2020	Linear SVC	70:30 split	87.90	0.358	–
[2], 2021	Logistic regression	Hold-out	88.43	0.570	0.859
[9], 2021	Deep neural network	10-fold CV	89.11	–	–
[10], 2021	SVM	80:20 split	87.76	–	–
[11], 2021	Random forest	80:20 split	85.12	–	0.808
[3], 2023	Logistic regression	5-fold CV	87.96	0.313	0.850
[4], 2025	Bi-TCN ensemble	5-fold CV	89.65	0.616	0.776
[8], 2026	XGBoost	In-fold SMOTE	87.00	0.530	0.797
Proposed	**Block-penalised LR**	**10 × 10-fold CV**	**89.16**	**0.585**	**0.863**
Whole dataset balanced before splitting					
[5], 2022	Extra trees	SMOTE, 85:15	93.00	0.930	–
[9], 2021	Deep neural network	ADASYN, 70:30	94.16	0.940	–
[6], 2023	Nested ensemble	SMOTE, 10-fold	94.53	0.945	0.985
[12], 2025	GAN + transformer	GAN, split	92.00	0.916	0.963
Proposed	**Framework (extra trees)**	**SMOTE, 10-fold**	**95.46**	**0.955**	**0.989**

3. Materials and Methods

3.1. Dataset

The IBM HR Analytics Employee Attrition & Performance dataset is a fictional benchmark created by IBM data scientists, with 1,470 employee records and 35 attributes on demographics, role, level, pay, tenure, overtime, travel and five survey scores; 237 employees (16.1%) left. It has no missing values or duplicates. The employee number and three constant columns were removed, and three pay-rate codes were excluded from the linear model after showing no signal in development.

3.2. Career-dynamics features

Sixteen indicators are derived row by row. Twelve are continuous: a satisfaction index (sum of five survey scores), income per job level, log income, income-to-age ratio, years per previous employer, the shares of the career spent at the company and of company tenure spent in the current role and with the current manager, promotion lag, age at hire, career start age and log commuting distance. Four flags mark no stock options, early career, overtime at job level 1, and single employees without stock options. Each uses only the employee's own record, so it leaks nothing.

3.3. Level encoding and block-penalised learner

Nominal attributes are one-hot encoded. The ten ordinal items (for example satisfaction, job level and stock option level) enter both as standardised scores and as one-hot levels, so non-monotone effects can be represented; attrition is 24.4% at stock option level 0, 7.6% at level 2 and 17.6% at level 3. The logistic model uses an L2 penalty ($C = 0.3$), and each standardised block is multiplied by a weight (numeric 0.5, nominal 2.0, ordinal levels 1.0, engineered 0.5), which divides its effective penalty by the squared weight. Weights and C were chosen by 10 × 5-fold cross-validation on the training split; the decision threshold is 0.5 (Fig. 1).

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

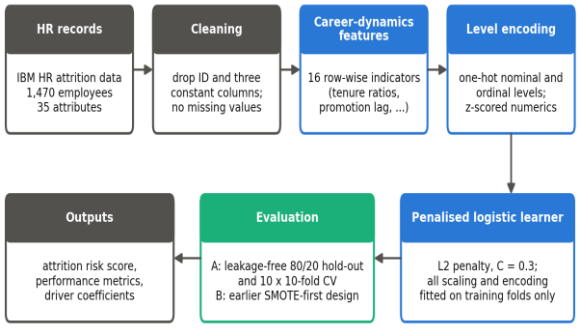


Figure 1. Workflow of the framework and the evaluation protocols.

3.4. Evaluation protocols

Protocol A is leakage-free. A stratified 80/20 split, with its seed fixed before modelling, gave a training part for all development and a test part scored once. Performance was also estimated by ten repetitions of stratified 10-fold cross-validation, refitting all preprocessing inside each training fold. Eleven baselines (logistic regression, RBF-SVM, k-NN, naive Bayes, decision tree, random forest, extra trees, XGBoost, LightGBM, CatBoost and a multilayer perceptron) used the original attributes, and paired differences over the 100 folds were tested with the corrected resampled t-test [16]. Protocol B reproduces earlier designs: the whole dataset is balanced with SMOTE [17] or ADASYN [18] and split 85:15 [5], 10-fold [6], 70:30 [9] or 80:20 [12], with the learner (logistic or extra trees) chosen by inner 5-fold cross-validation. Class metrics refer to leavers in Protocol A and are class-averaged in Protocol B, as in those studies.

3.5. Implementation

The code was written in Python 3.11 (scikit-learn, imbalanced-learn, XGBoost, LightGBM and CatBoost) and run on a desktop computer; it is available from the author on reasonable request.

4. Results

4.1. Leakage-free performance

The proposed model led all twelve learners in accuracy, F1, AUC and MCC (Table 2). Its AUC of 0.863 (SD 0.043) was 0.027 above plain logistic regression and at least 0.037 above the boosted trees, all differences being significant ($p < 0.001$); accuracy reached 89.16% and MCC 0.549. Tree ensembles kept high specificity but recalled only 16–33% of leavers. On the hold-out split, where every model scored a lower AUC than in cross-validation, the proposed model again had the highest AUC (0.852), F1 (0.494) and MCC (0.436) (Figs. 2 and 3).

TABLE 2. Mean 10 × 10-fold cross-validation results and hold-out AUC.

Model	Acc. (%)	F1	AUC	MCC	Hold-out AUC
Proposed	**89.16**	**0.585**	**0.863**	**0.549**	**0.852**
Logistic regression	88.08	0.526	0.836	0.491	0.811
RBF-SVM	87.53	0.473	0.825	0.451	0.811
XGBoost	87.55	0.449	0.826	0.443	0.813
LightGBM	87.44	0.455	0.825	0.441	0.795
CatBoost	87.23	0.440	0.820	0.429	0.803
Random forest	85.94	0.264	0.811	0.318	0.794
Extra trees	86.05	0.302	0.814	0.334	0.807
MLP	85.50	0.493	0.789	0.418	0.776
k-NN	84.26	0.056	0.713	0.105	0.672
Naive Bayes	65.95	0.409	0.750	0.279	0.687
Decision tree	83.16	0.339	0.669	0.267	0.662

In an ablation on the same folds, AUC rose from 0.836 to 0.858, 0.861 and 0.863, and accuracy from 88.08% to 89.16%, as level encoding, career-dynamics features and block penalties were added in turn.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

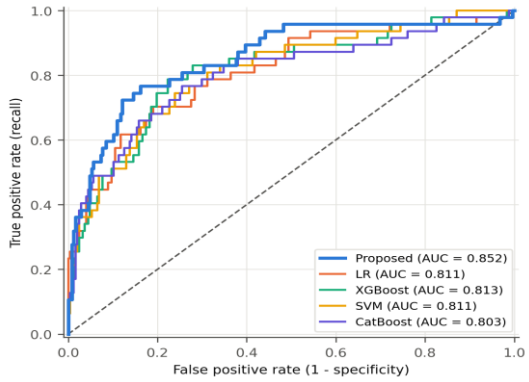


Figure 2. ROC curves on the hold-out test set.

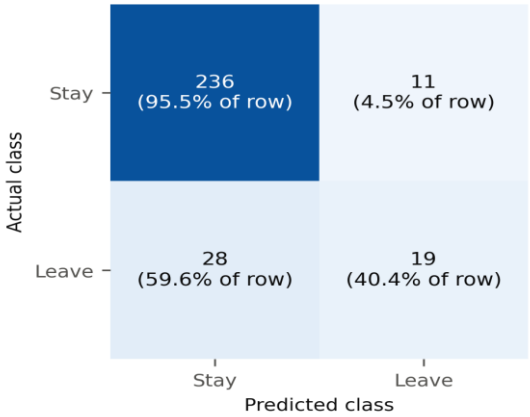


Figure 3. Hold-out confusion matrix of the proposed model.

4.2. Attrition drivers

Overtime raised the log-odds of leaving most (+1.43), followed by frequent travel (+0.91) and technician and sales roles, while research directors, research scientists and non-travellers were least likely to leave (Fig. 4).

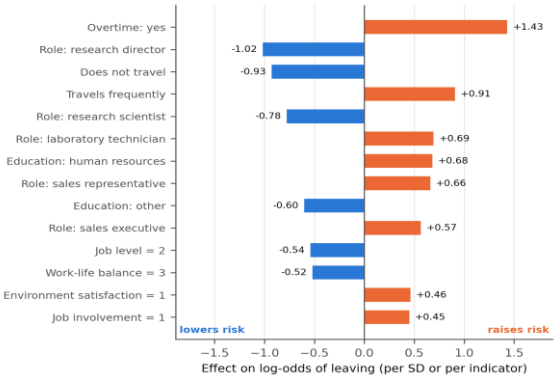


Figure 4. Largest effects in the proposed model.

4.3. Oversample-then-split designs

Inner cross-validation chose extra trees in all four reproduced designs, giving 93.24% (SMOTE, 85:15), 95.46% with AUC 0.989 (SMOTE, 10-fold), 94.71% (ADASYN, 70:30) and 93.93% with AUC 0.984 (SMOTE, 80:20). These are not deployable accuracies: moving SMOTE inside the training folds cut accuracy by 6.6 points for random forest, 8.9 for extra trees and 5.8 for CatBoost (Fig. 5).

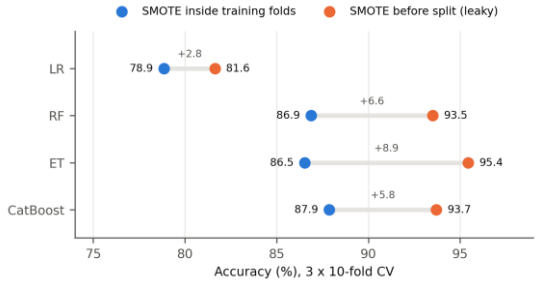


Figure 5. Accuracy with SMOTE before the split and inside training folds.

5. Discussion

Among leakage-free studies (Table 1) the proposed model has the highest AUC (0.863 against 0.859 [2]) and the highest 10-fold cross-validated accuracy (89.16% against 89.11% [9]), and its precision (0.766) exceeds that of [2] and [4]. Two single-run accuracies are higher: 91.16% on one 70:30 split [9] and 89.65% from one 5-fold run [4]. With a few hundred test employees, accuracy shifts by several points between splits: our hold-out accuracy fell 2.4 points below the cross-validated mean, every model lost AUC on that split, and the same authors report 89.11% under 10-fold cross-validation [9]. The 89.65% model also ranks employees less well, with an AUC 0.087 below ours [4]. Under the oversample-then-split designs the framework exceeds all four studies in accuracy and F1, and in AUC where reported (Table 3).

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

TABLE 3. Gains of the proposed model over the best previously reported results.

Protocol	Metric	Best previous	Proposed	Gain
A	AUC	0.859 [2]	0.863	+0.004
A	AUC vs. most accurate model	0.776 [4]	0.863	+0.087
A	Accuracy, 10-fold CV (%)	89.11 [9]	89.16	+0.05
B, SMOTE 85:15	Accuracy (%)	93.00 [5]	93.24	+0.24
B, SMOTE 85:15	F1	0.930 [5]	0.932	+0.002
B, SMOTE 10-fold	Accuracy (%)	94.53 [6]	95.46	+0.93
B, SMOTE 10-fold	F1	0.945 [6]	0.955	+0.010
B, SMOTE 10-fold	AUC	0.985 [6]	0.989	+0.004
B, ADASYN 70:30	Accuracy (%)	94.16 [9]	94.71	+0.55
B, ADASYN 70:30	F1	0.940 [9]	0.947	+0.007
B, balanced split	Accuracy (%)	92.00 [12]	93.93	+1.93
B, balanced split	F1	0.916 [12]	0.939	+0.023
B, balanced split	AUC	0.963 [12]	0.984	+0.021

For HR practice, the model ranks employees by risk and names the factors behind each score, so actions such as reviewing overtime and travel load can be targeted. The data are fictional and cross-sectional, and recall at the default threshold (0.48) is moderate; validation on longitudinal company records is the next step.

6. Conclusion

Career-dynamics features, level encoding and block-wise penalties made a logistic model the strongest of twelve learners on the IBM attrition benchmark, with the best AUC among leakage-free studies. Reproducing oversample-then-split designs showed that the framework also exceeds published figures on their own terms, and that those figures overstate deployable accuracy by up to nine points.

Data Availability

The dataset is available at https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset (mirror: https://github.com/IBM/employee-attrition-aif360).

References

[1] F. Fallucchi et al., "Predicting employee attrition using machine learning techniques," Computers, vol. 9, no. 4, p. 86, 2020, doi:10.3390/computers9040086.

[2] A. Qutub et al., "Prediction of employee attrition using machine learning and ensemble methods," Int. J. Mach. Learn. Comput., vol. 11, no. 2, pp. 110-114, 2021, doi:10.18178/ijmlc.2021.11.2.1022.

[3] F. Guerranti and G. M. Dimitri, "A comparison of machine learning approaches for predicting employee attrition," Appl. Sci., vol. 13, no. 1, p. 267, 2023, doi:10.3390/app13010267.

[4] F. Mortezapour Shiri, S. Yamaguchi, and M. A. B. Ahmadon, "A deep learning model based on bidirectional temporal convolutional network (Bi-TCN) for predicting employee attrition," Appl. Sci., vol. 15, no. 6, p. 2984, 2025, doi:10.3390/app15062984.

[5] A. Raza et al., "Predicting employee attrition using machine learning approaches," Appl. Sci., vol. 12, no. 13, p. 6424, 2022, doi:10.3390/app12136424.

[6] M. S. Alshiddy and B. N. Aljaber, "Employee attrition prediction using nested ensemble learning techniques," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 7, pp. 932-938, 2023, doi:10.14569/IJACSA.2023.01407101.

[7] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, p. 100804, 2023, doi:10.1016/j.patter.2023.100804.

[8] A. M. Căvescu and A. N. Popescu, "Leakage-free evaluation for employee attrition prediction on tabular data," Information, vol. 17, no. 3, p. 308, 2026, doi:10.3390/info17030308.

[9] S. Al-Darraji et al., "Employee attrition prediction using deep neural networks," Computers, vol. 10, no. 11, p. 141, 2021, doi:10.3390/computers10110141.

[10] N. Mansor, N. S. Sani, and M. Aliff, "Machine learning for predicting employee attrition," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 11, pp. 435-445, 2021, doi:10.14569/IJACSA.2021.0121149.

[11] M. Pratt, M. Boudhane, and S. Cakula, "Employee attrition estimation using random forest algorithm," Baltic J. Mod. Comput., vol. 9, no. 1, pp. 49-66, 2021, doi:10.22364/bjmc.2021.9.1.04.

[12] İ. Tanyıldızı Baydili and B. Tasci, "Predicting employee attrition: XAI-powered models for managerial decision-making," Systems, vol. 13, no. 7, p. 583, 2025, doi:10.3390/systems13070583.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

[13] M. Nandal et al., "Employee attrition: Analysis of data driven models," EAI Endorsed Trans. Internet Things, vol. 10, 2024, doi:10.4108/eetiot.4762.
[14] G. Nassreddine et al., "Employee attrition prediction: An explanatory and statistically robust ensemble learning model," Computers, vol. 15, no. 3, p. 185, 2026, doi:10.3390/computers15030185.
[15] F. Alsubaie and M. Aldoukhi, "Using machine learning algorithms with improved accuracy to analyze and predict employee attrition," Decis. Sci. Lett., vol. 13, no. 1, pp. 1-18, 2024, doi:10.5267/j.dsl.2023.12.006.
[16] C. Nadeau and Y. Bengio, "Inference for the generalization error," Mach. Learn., vol. 52, no. 3, pp. 239-281, 2003, doi:10.1023/A:1024068626366.
[17] N. V. Chawla et al., "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321-357, 2002, doi:10.1613/jair.953.
[18] H. He et al., "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in Proc. IEEE IJCNN, 2008, pp. 1322-1328, doi:10.1109/IJCNN.2008.4633969.