High Recall Oriented Employee Attrition Prediction using Stacking Ensemble Model

Garima Singh, Nishkarsh Singhal, Shaivee Sharma

under the scheme of DSKC guided by: Dr. Seema Aggarwal, Dr. Tarun Kumar Gupta, Dr. Tulika

ABSTRACT

Understanding and predicting employee attrition has become essential for effective HR planning and talent retention strategies. In this study, we developed a stacking ensemble machine learning model optimized for high recall on the minority class (Attrition = 'Yes') using the IBM HR Analytics dataset. The dataset showed significant class imbalance (16:84), necessitating the use of SMOTE for synthetic oversampling. After rigorous preprocessing — including feature pruning, label encoding, and standard scaling — we trained base learners (Random Forest, XGBoost, LightGBM) and a meta learner (Extra Trees Classifier) with manually tuned hyperparameters for optimal performance.

To further enhance sensitivity, we applied threshold tuning (optimal at 0.28) to maximize F1-score for the minority class. The final model achieved a recall of 0.83 and an F1-score of 0.57 on the attrition class, significantly outperforming classical baselines.

METHODOLOGY

DATASET AND PREPROCESSING

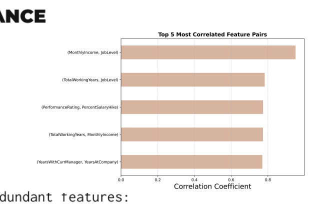
DATASET(S) used

DATASET (employee attrition datasets)	features	Records	Class imbalance
IBM	35	1470	16:84
WATSON HEALTHCARE	35	1676	13:87
STEALTH TECHNOLOGIES	23	14900 test 59600 train	47:52

Feature overview of IBM Dataset

Category	Features
Demographics	Age, Gender, MaritalStatus
Job Role & Department	JobRole, Department, JobLevel, JobInvolvement, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager
Compensation	MonthlyIncome, MonthlyRate, DailyRate, HourlyRate, StockOptionLevel, PercentSalaryHike
Performance & Satisfaction	JobSatisfaction, EnvironmentSatisfaction, PerformanceRating, RelationshipSatisfaction, WorkLifeBalance
Education & Training	Education, EducationField, TrainingTimesLastYear
Experience & Background	TotalWorkingYears, NumCompaniesWorked, DistanceFromHome, OverTime, BusinessTravel

PREPROCESSING AND HANDLING IMBALANCE

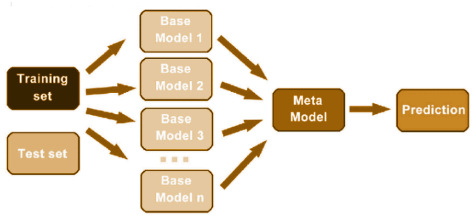


1. **RAW DATA** — Starting point with original dataset features
   - **35** features
   - **1470** datapoints
   - **16:84** class imbalance
2. **Cleaning** — Removed redundant features:
   - **EmployeeCount**
   - **Over18**
   - **StandardHours**

   MonthlyIncome and **JobLevel** showed high correlation (0.95), so **MonthlyIncome** was dropped
3. **Encoding** — Used Label encoding to convert Categorical features into numerical (eg: Gender, JobRole)
4. **SMOTE** — Since the dataset was quiet imbalanced, we applied SMOTE to balance the classes in the train data
5. **Scaling** — Used Standard Scaler to normalize numerical features
6. **Model Input** — Final processed data ready for model training

Model Exploration & Final Choice

1. **Started with Logistic Regression, SVM, and shallow Neural Networks**
   Observed low F1 & recall, especially on minority class and over-fitting
2. **Tree-based models like RF, XGBoost, and LGBM showed stronger performance**
   Inspired by stacking strategies proposed in recent research [7] we explored **Stacking Ensemble**
3. **Final model = RF + XGB + LGBM → Extra Trees**
   Recall boost confirmed



MODEL NAME	Recall(minority class)	F1(minority class)
logistic regression	0.72	0.55
SVM	0.53	0.39
FNN	0.55	0.54
MLP	0.55	0.6
CNN	0.53	0.54
LGBM	0.53	0.51
random forest	0.49	0.51
XGBoost	0.6	0.52
easy ensemble	0.44	0.56
Stacking Ensemble	0.83	0.57

Model architecture

We meticulously tested all parameters of each model manually to achieve optimal results

Input Dataset
Preprocessed HR data with balanced classes

Base Learners

Random Forest [n_estimators=120,max_depth=5,min_samples_split=10,min_samples_leaf=5,max_features='sqrt',bootstrap=True,classweight='balanced_subsample']

XGBoost [use_label_encoder=False,eval_metric='logloss',n_estimators=100,learning_rate=0.03,max_depth=3,subsample=0.6,colsample_bytree=0.6,scale_pos_weight=2,reg_lambda=10,reg_alpha=2,gamma=10]

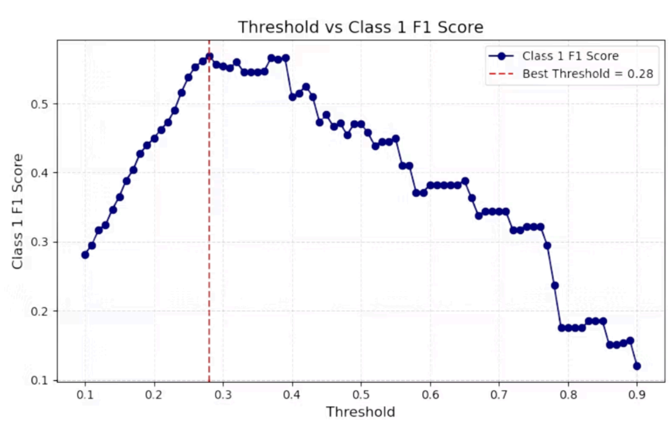
LightGBM [n_estimators=100,learning_rate=0.05,max_depth=5,subsample=0.7,colsample_bytree=0.7,class_weight='balanced',reg_lambda=5,reg_alpha=1]

Meta Learner
ExtraTrees Classifier [n_estimators=100,max_depth=6,min_samples_split=10,min_samples_leaf=5,max_features='sqrt',class_weight='balanced',n_jobs=-1]
Combines predictions from base models

Final Prediction
Optimized for high recall on attrition class

Threshold Tuning & F1 Optimization

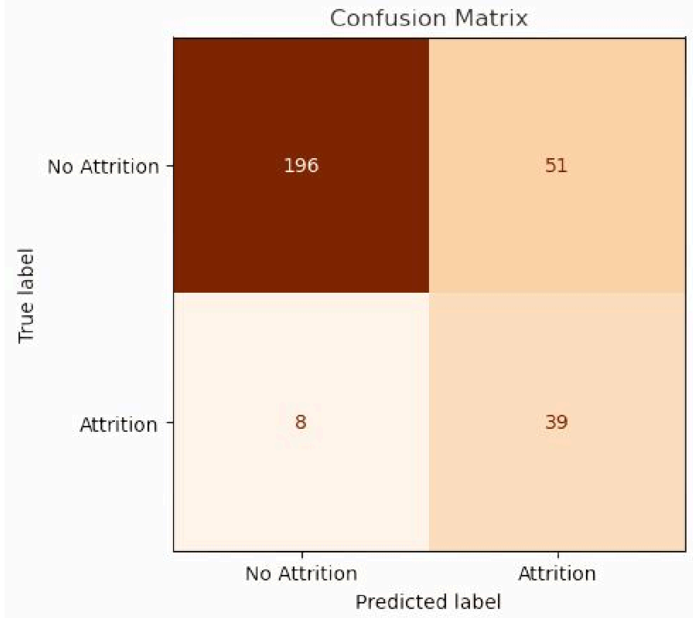
1. After training, we performed threshold tuning to optimize performance
2. Used model's predicted probabilities
3. Iterated threshold from **0.1** to **0.9** (**step = 0.01**)
4. Evaluated each on **class 1 F1-score**
5. Final selected threshold = 0.28
6. Helped improve **recall** and **F1-score** without retraining the model



Tuning the threshold allowed us to better balance false positives and false negatives, especially important for the minority class

RESULTS & CONCLUSION

Metric	Score
Accuracy	0.80
Precision	0.88
Recall	0.81
F1 score	0.83
Support	294
Recall (minority class)	**0.83**



High Recall of positive attrition cases (0.83) ensures most attrition cases are detected, aligning with our real-world objective of early identification & intervention.

Reference Study for SOTA: "Developing a hybrid machine learning model for employee turnover prediction: Integrating LightGBM and genetic algorithms" Hojat Talebi et al., Journal of Open Innovation: Technology, Market, and Complexity, Volume 11 (2025), Article ID: 100557 DOI: https://doi.org/10.1016/j.joitmc.2025.100557

Comparison with State-of-the-Art Model

It Uses GA for feature selection + LightGBM for classification

Performance Comparison
Proposed model outperforms SOTA across **all key** metrics

Ensemble Advantage
Stacking ensemble captures diverse learning patterns from RF, XGBoost & LGB

Threshold Tuning
Additional threshold tuning improves recall (crucial for attrition prediction)

Metric	SOTA Model	Proposed Model
F1-Score	0.73	**0.83**
Precision	0.75	**0.88**
Recall	0.72	**0.81**
Accuracy	0.78	**0.80**

Cross-Dataset Validation

The model was tested on two additional datasets with different class balances and feature dimensions, to assess its generalization and recall consistency.

Dataset	Records	Features	Class Balance	Accuracy	Weighted Recall	Weighted F1-score
IBM	294	35	16:84	0.80	**0.81**	**0.83**
Watson	336	35	13:87	0.93	**0.93**	**0.93**
StealthTech	14,900	23	47:52	0.74	**0.74**	**0.74**

- High weighted metrics across all datasets show strong generalization
- Watson (Healthcare) dataset had **best overall metrics**, even with class imbalance
- Model adapts well to both **imbalanced** and **balanced** scenarios
- Reflects strong potential for real-world implementation across domains

Business Impact

- **High Recall = Early Identification of At-Risk Employees** → Even if few false positives exist, better to **consult and retain proactively**
- **Helps in flagging "Asset-Class" Employees** → Companies can't afford to lose top performers
- **Supports Strategic HR Decisions** → Personalized retention plans, counseling, incentive tweaking
- **Saves Cost of Rehiring & Retraining** → Retaining employees is cheaper than replacing them

Conclusion

High-Performance Predictor
Stacked ensemble model + **threshold tuning** creates a powerful attrition prediction system

Recall-Focused Strategy
Recall-focused approach aligns with modern HR priorities by catching all potential exits

Cross-Dataset Performance
Generalizes well across diverse datasets including corporate, healthcare, and synthetic environments

"Missing an attrition case can be costlier than wrongly flagging one"

so high recall gives the company that protective edge.

REFERENCES

1. Optimising HRM Practices in Call Centres
https://doi.org/10.1108/IJOA-12-2024-5117
Date: 13th May, 2025

2. Predicting Employee Attrition using ML Approaches
https://doi.org/10.61591/jslhu.20.717
Date: 15th March, 2025

3. ML Applications in HRM: Turnover & Performance
https://doi.org/10.53032/tvcr/2025.v7n2.37
Date: 30th April, 2025

4. Predicting Employee Attrition Using ML
https://doi.org/10.3390/app12136424
Date: 24th June, 2022

5. ML for Predicting Employee Attrition
http://dx.doi.org/10.14569/IJACSA.2021.0121149
Date: 30th Nov, 2021

6. Employee Attrition: Analysis of Data-Driven Models
https://doi.org/10.4108/eetiot.4762
Date: 3rd Jan, 2024

7. Predictive Model Using Stacking Ensemble
https://doi.org/10.1016/j.eswa.2022.119364
Date: 7th Dec, 2022

8. Hybrid Model for Turnover Prediction (LightGBM + GA)
https://doi.org/10.1016/j.joitmc.2025.100557
Date: 31st May, 2025

9. Comparative Analysis using IBM HR Dataset
https://doi.org/10.1016/j.procs.2025.04.659
Date: 10th May, 2025