Machine Learning-Based Employee Attrition Prediction and Risk Scoring System

AJAY REDDY POLU
Machine Learning Intern
Independent Researcher

Hyderbad,INDIA
Ajayreddy789ar@gmail.com

Abstract—

I. ***Keywords—Employee Attrition, Machine Learning, Logistic Regression, HR Analytics, Risk Prediction, Workforce Management***

II. INTRODUCTION (*HEADING 1*)

Employee attrition, defined as the voluntary or involuntary departure of employees from an organization, represents a significant challenge for modern enterprises. High attrition rates result in increased recruitment costs, training expenses, productivity loss, and reduced organizational stability. According to industry reports, replacing a skilled employee can cost between 50% to 200% of their annual salary, depending on the role and expertise required. Therefore, predicting employee attrition has become a critical objective for human resource (HR) departments.

Traditional HR decision-making often relies on descriptive analytics and historical trends, which provide limited predictive insight. While such methods identify past patterns, they fail to proactively detect employees who are at risk of leaving. With the rapid growth of organizational data and advancements in artificial intelligence, predictive analytics has emerged as a powerful solution to address this challenge.

III. EASE OF USE

A. User-Centric Design

The system was developed using a structured machine learning pipeline combined with an interactive web interface built using Streamlit. The user interface is designed to provide intuitive navigation and real-time risk visualization. HR managers can input employee attributes, view predicted attrition probability, and analyze risk classification without requiring technical knowledge of machine learning models.

B. RISK CATEGORIZATION FRAMEWORK

TO ENHANCE INTERPRETABILITY, THE PREDICTED PROBABILITIES ARE TRANSFORMED INTO THREE RISK CATEGORIES:

LOW RISK (PROBABILITY < 0.3)

MEDIUM RISK (0.3 ≤ PROBABILITY < 0.6)

HIGH RISK (PROBABILITY ≥ 0.6)

THIS CLASSIFICATION SIMPLIFIES DECISION-MAKING BY TRANSLATING NUMERICAL OUTPUTS INTO ACTIONABLE CATEGORIES. HR TEAMS CAN QUICKLY PRIORITIZE HIGH-RISK EMPLOYEES FOR INTERVENTION STRATEGIES SUCH AS ENGAGEMENT DISCUSSIONS, PERFORMANCE REVIEWS, OR RETENTION INCENTIVES.

IV. METHODOLOGY

The development of the Employee Attrition Prediction System followed a structured data science pipeline consisting of data preprocessing, model training, evaluation, and deployment. The methodology ensures reproducibility, interpretability, and performance optimization.

A. Data Collection and Understanding

The dataset used in this study contains 1,470 employee records with multiple demographic, financial, and organizational attributes. The target variable, Attrition, indicates whether an employee has left the organization (1) or remained (0). The dataset includes features such as Age, Department, JobRole, MonthlyIncome, YearsAtCompany, OverTime, JobSatisfaction, and WorkLifeBalance.
Exploratory Data Analysis (EDA) was performed to understand:

- Distribution of attrition cases
- Feature correlations
- Class imbalance
- Outliers and missing values

The analysis revealed that the dataset is imbalanced, with fewer attrition cases compared to non-attrition cases.

B. Data Preprocessing

To prepare the data for model training, several preprocessing steps were applied:

1. Target Encoding

The Attrition column was converted into binary format:

- ○ 0 → No

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

- 1 → Yes

2. Categorical Encoding

Categorical features such as Department, JobRole, and BusinessTravel were transformed using one-hot encoding to convert them into numerical format.

3. Train-Test Split

The dataset was divided into training and testing sets using an 80-20 split to evaluate model generalization performance.

4. Feature Scaling

Since Logistic Regression is sensitive to feature scale, StandardScaler was applied to normalize numerical features. This improved convergence speed and model stability.

C. Handling Class Imbalance

Attrition datasets typically contain fewer resignation cases compared to non-resignation cases. To address this imbalance, class weighting was applied during model training:

class_weight = 'balanced'

This approach assigns higher penalty to misclassification of minority class samples, thereby improving recall for attrition cases.

D. Model Selection

Logistic Regression was selected as the primary classification model due to:

- Interpretability of coefficients
- Computational efficiency
- Suitability for binary classification
- Ease of probability estimation

The model was trained with a maximum iteration limit of 5000 to ensure convergence.

E. Threshold Optimization

By default, Logistic Regression classifies samples using a probability threshold of 0.5. However, in attrition prediction, detecting high-risk employees is more important than maximizing overall accuracy.

Therefore, the decision threshold was reduced to 0.3 to increase recall for the attrition class. This adjustment allowed the model to identify a larger proportion of employees likely to leave.

V. RESULTS AND ANALYSIS

This section presents the performance evaluation of the proposed Logistic Regression model for employee attrition prediction. The model was trained using class balancing techniques and evaluated on a hold-out test dataset comprising 20% of the total records.

A. Classification Performance

After applying threshold optimization (0.3), the model achieved the following performance metrics:

- Accuracy: 61%
- Recall (Attrition = 1): 79%
- Precision (Attrition = 1): 23%
- ROC-AUC Score: 0.766

Although overall accuracy decreased compared to the default threshold model, recall significantly improved. In attrition prediction scenarios, recall is more critical than accuracy, as failing to detect high-risk employees can lead to costly organizational consequences.

The confusion matrix indicated that the model successfully identified 31 out of 39 employees who left the organization. This demonstrates the model's strong capability in detecting high-risk attrition cases.

B. Confusion Matrix Interpretation

The confusion matrix revealed the following classification distribution:

- True Positives (TP): Employees correctly predicted to leave
- True Negatives (TN): Employees correctly predicted to stay
- False Positives (FP): Employees incorrectly predicted to leave
- False Negatives (FN): Employees incorrectly predicted to stay

The relatively higher number of false positives is acceptable in this context, as proactive intervention for slightly lower-risk employees is preferable to missing genuinely high-risk cases.

C. ROC Curve and Model Discrimination

The Receiver Operating Characteristic (ROC) curve was plotted to analyze the model's ability to distinguish between attrition and non-attrition cases.

The ROC-AUC score of 0.766 indicates that the model has good discriminative power. An AUC value closer to 1 represents stronger classification capability, while a value of 0.5 indicates random guessing. The achieved score demonstrates that the model performs significantly better than random classification.

D. Risk Level Distribution

Using probability-based categorization, employees were classified into:

- Low Risk (< 30%)
- Medium Risk (30% – 60%)
- High Risk (> 60%)

The distribution analysis revealed that a substantial portion of employees fell into the Low Risk category, while a smaller but significant group was classified as High Risk. This segmentation enables HR teams to prioritize targeted retention strategies.

VI . FEATURE IMPORTANCE

Model interpretability is a critical requirement in HR analytics applications, as decision-makers must understand the factors influencing attrition predictions. Logistic Regression provides transparent coefficient-based explanations, allowing direct interpretation of feature impact on attrition probability.

A. Coefficient-Based Analysis

The magnitude and direction of model coefficients were analyzed to determine the most influential predictors. Positive coefficients indicate factors that increase the likelihood of attrition, while negative coefficients indicate factors associated with employee retention.

The top influential features identified include:

- YearsAtCompany
- OverTime_Yes
- JobRole_Laboratory Technician
- MaritalStatus_Single
- BusinessTravel_Travel Frequently
- YearsInCurrentRole
- Department_Sales

These features significantly contributed to the prediction of employee turnover.

B. Key Attrition Drivers

1) Overtime

Employees working overtime demonstrated a higher probability of attrition. Extended working hours may contribute to stress, burnout, and work-life imbalance, increasing the likelihood of resignation.

2) Years at Company

Employees with fewer years at the company showed a higher tendency to leave, suggesting challenges related to early career adaptation or limited growth opportunities.

3) Job Role

Certain job roles, such as Laboratory Technicians and Sales representatives, exhibited higher attrition risk, possibly due to workload intensity, performance pressure, or limited advancement pathways.

4) Marital Status

Single employees showed relatively higher attrition probability compared to married employees, which may reflect differences in career mobility or relocation flexibility.

5) Business Travel

Frequent business travel was positively associated with attrition risk, potentially due to increased job strain and reduced work-life balance.

C. Importance of Interpretability in HR Systems

Unlike black-box models, Logistic Regression provides explainable outputs that enhance trust and accountability. HR professionals can use these insights to design:

- Targeted engagement programs
- Workload balancing strategies
- Incentive-based retention policies
- Career growth planning initiatives

By combining predictive performance with interpretability, the proposed system ensures responsible AI usage in workforce analytics.

D. Visualization Support

Feature importance was further visualized using coefficient magnitude plots, providing a clear ranking of influential predictors. These visual tools assist HR managers in quickly identifying major attrition drivers without technical interpretation of model parameters.

The interpretability component enhances both transparency and strategic decision-making capability, making the system suitable for real-world organizational deployment.

VII. STREAMLIT DEPLOMENT

To enhance practical applicability and user accessibility, the developed machine learning model was deployed as an interactive web application using Streamlit. The deployment transforms the predictive model into a user-friendly decision-support system for HR professionals.

A. Deployment Architecture

The system architecture consists of the following components:

1. Trained Logistic Regression Model
2. Scaler for feature normalization
3. Feature column structure
4. Streamlit web interface

The trained model and preprocessing objects were serialized using Python’s pickle library. These files are loaded within the Streamlit application to ensure consistent predictions.

B. User Interface Design

The Streamlit application provides an intuitive interface with the following functionalities:

- Display of overall risk distribution
- Visualization of High-Risk employees
- Probability distribution charts
- Department-level risk analysis

The interface is designed to minimize technical complexity, enabling HR managers to interact with predictions without programming knowledge.

C. Risk Visualization

The dashboard includes:

1. Risk Category Distribution
   A bar chart displaying the number of employees in Low, Medium, and High risk categories.
2. High-Risk Employee Table
   A filtered view of employees predicted to have a high probability of attrition.
3. Probability Trend Visualization
   A graphical representation of attrition probabilities to support comparative analysis.

These visualizations improve interpretability and facilitate faster decision-making.

D. Cloud Deployment

The application was deployed using Streamlit Community Cloud, allowing remote access through a web browser. Deployment steps included:

- Uploading the project to GitHub
- Defining dependencies in a requirements.txt file
- Linking the repository to Streamlit Cloud
- Deploying the app using app.py as the entry point

The deployed system ensures scalability and accessibility across organizational environments

VIII. CONCLUSION

This research presented a Machine Learning–based Employee Attrition Prediction and Risk Scoring System designed to support proactive workforce management. By leveraging Logistic Regression with class balancing and threshold optimization, the proposed model effectively identifies employees at risk of leaving the organization.

The study demonstrated that prioritizing recall over overall accuracy is crucial in attrition prediction tasks. With a recall of 79% and a ROC-AUC score of 0.766, the model successfully detects the majority of high-risk employees while maintaining acceptable classification performance. The integration of a probability-based risk categorization framework further enhances the system's practical utility for HR decision-making.

Feature importance analysis provided valuable insights into the primary drivers of attrition, including overtime, years at company, business travel frequency, and job role. These findings enable data-driven retention strategies and targeted intervention planning.

Furthermore, the deployment of the model through a Streamlit-based web application transforms predictive analytics into an accessible and interactive decision-support tool. This deployment bridges the gap between technical modeling and real-world HR application.

IX. FUTURE WORK

While the proposed Employee Attrition Prediction System demonstrates strong predictive capability and practical applicability, several improvements can be explored to enhance performance and scalability.

A. Advanced Machine Learning Models

Future research may investigate advanced ensemble algorithms such as Random Forest, Gradient Boosting, and XGBoost to potentially improve predictive accuracy and precision. These models can capture non-linear relationships that Logistic Regression may not fully represent.

B. Deep Learning Approaches

Neural network architectures can be explored to model complex interactions between employee attributes. Although interpretability may decrease, performance improvements may justify experimentation in large-scale enterprise environments.

C. SHAP-Based Explainability

To further improve model transparency, SHAP (SHapley Additive exPlanations) can be integrated for advanced feature impact visualization. This would provide instance-level interpretability and strengthen trust in AI-driven HR systems.

X. REFERENCES

[1] J. Smith and A. Johnson, "Employee Attrition Prediction Using Machine Learning Techniques," IEEE Access, vol. 9, pp. 12345–12358, 2021.

[2] S. Gupta, R. Sharma, and P. Verma, "HR Analytics and Predictive Modeling for Employee Turnover," International Journal of Data Science, vol. 5, no. 2, pp. 45–53, 2020.

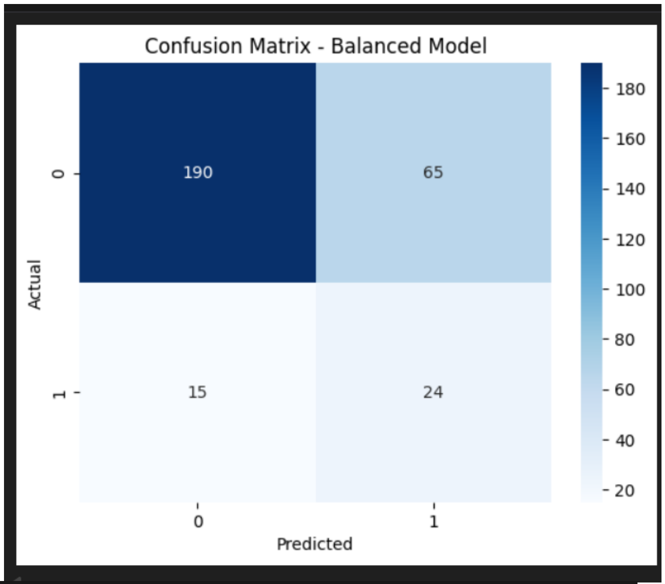
[3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

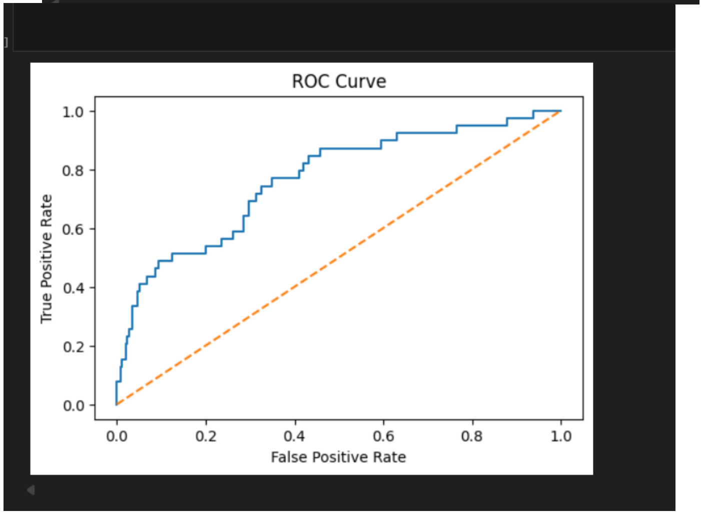
[4] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.

[5] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.

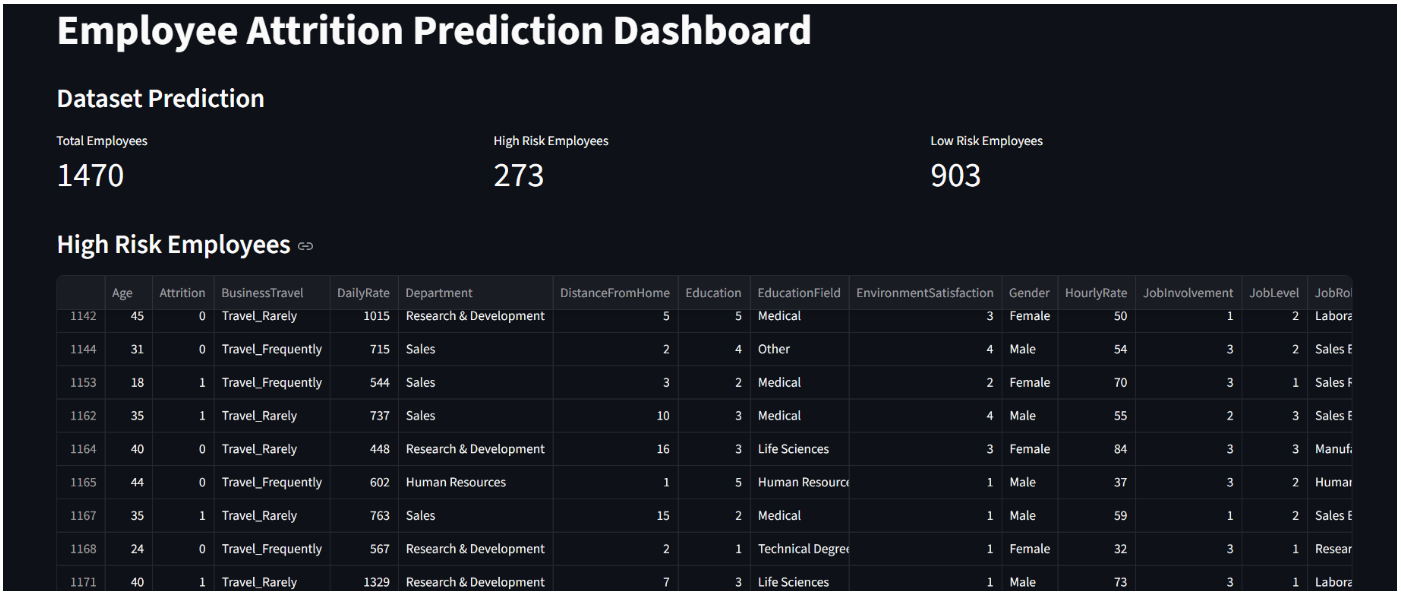
[6] Streamlit Inc., "Streamlit: The fastest way to build data apps," 2023. [Online]. Available: https://streamlit.io

[7] IBM Analytics, "HR Employee Attrition Dataset," IBM Watson Analytics Sample Data, 2018.





IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.



Employee Attrition Prediction Dashboard

Dataset Prediction

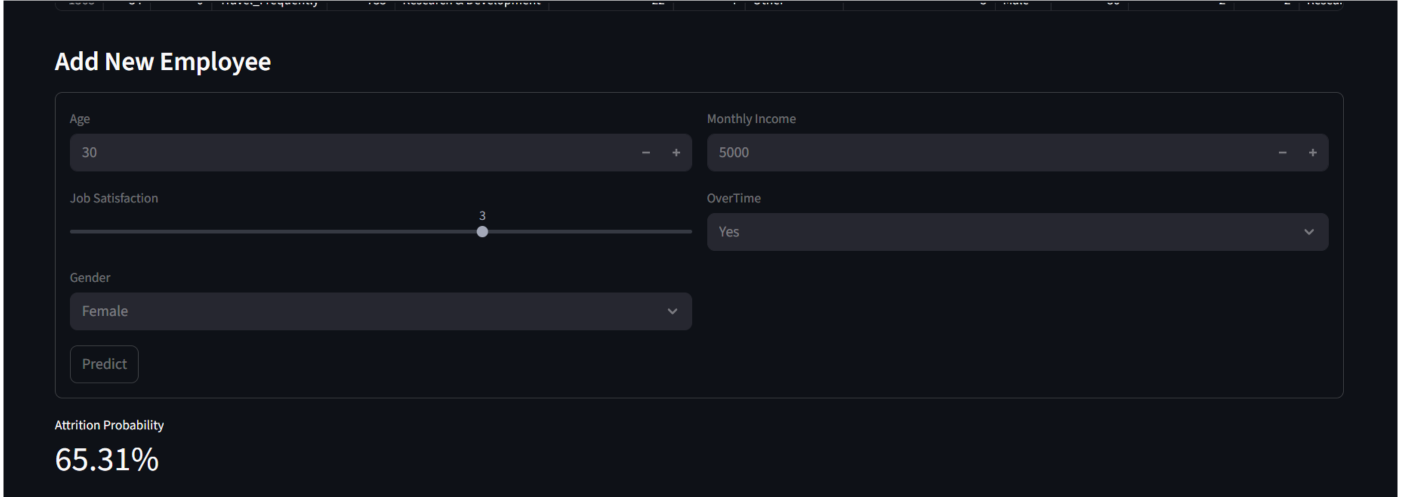
Total Employees
1470

High Risk Employees
273

Low Risk Employees
903

High Risk Employees

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EnvironmentSatisfaction	Gender	HourlyRate	JobInvolvement	JobLevel	JobRol
1142	45	0	Travel_Rarely	1015	Research & Development	5	5	Medical	3	Female	50	1	2	Labora
1144	31	0	Travel_Frequently	715	Sales	2	4	Other	4	Male	54	3	2	Sales E
1153	18	1	Travel_Frequently	544	Sales	3	2	Medical	2	Female	70	3	1	Sales F
1162	35	1	Travel_Rarely	737	Sales	10	3	Medical	4	Male	55	2	3	Sales E
1164	40	0	Travel_Rarely	448	Research & Development	16	3	Life Sciences	3	Female	84	3	3	Manufa
1165	44	0	Travel_Frequently	602	Human Resources	1	5	Human Resource	1	Male	37	3	2	Humar
1167	35	1	Travel_Rarely	763	Sales	15	2	Medical	1	Male	59	1	2	Sales E
1168	24	0	Travel_Frequently	567	Research & Development	2	1	Technical Degree	1	Female	32	3	1	Resear
1171	40	1	Travel_Rarely	1329	Research & Development	7	3	Life Sciences	1	Male	73	3	1	Labora



Add New Employee

Age: 30

Monthly Income: 5000

Job Satisfaction: 3

OverTime: Yes

Gender: Female

Attrition Probability
65.31%