Employee Attrition Prediction

Benson Antony D V, Haritha Rajeev

***Aabstract:** Employee attrition occurs when a worker leaves a company to join another firm for a better offer. It might also be referred to as Employee Defection. Representative downsizing is likely to be significant when there is a pressing demand for workers in a particular industry due to mass retirements or organizational growth. At one point, the programming industry had significant attrition rates due to abundant job opportunities in the software sector driven by the demand for software products across all industries. Reducing the employee attrition rate is a challenging challenge faced by HR managers. This study provides a clear viewpoint on predicting employee turnover using Machine Learning methods. The projection is completed using data obtained from IBM HR analysis. We employed Logistic Regression for the analysis and achieved an accuracy rate of 87%.*

***Keyword:** Employee Turnover, Human Resources Managers, Logistic Regression, Machine Learning Model, Software Development Sector.*

I. INTRODUCTION

Employee Attrition can also be referred to as Employee Turnover. Wear and tear is a prevalent concern in today's industries. It is a prevalent issue in most organizations. The gradual reduction in the number of representatives due to retirement, resignation, or death can be referred to as Employee Defection. Wear rates vary amongst industries based on their specific standards and can differ between skilled and unskilled occupations. Organizations are challenged with managing enrollment and retention of talents while also addressing skill loss due to industry downturns or intentional employee turnover. When a highly skilled and well-balanced employee departs from the organization, it creates a void. The association loses crucial skills, knowledge, and business relationships. Current leaders and executives are very focused on reducing turnover in the organization to maximize efficiency and promote organizational growth. Representative acquiescences are essential for any business organization. If the situation is not handled properly, the departure of key staff members can lead to significant losses in productivity. Employee turnover results in performance losses that can have long-term detrimental effects on enterprises. As the reduction of costs is a significant concern for every industry, companies strive to employ innovative business strategies to minimize maintenance expenses.

There is no foolproof way to completely minimize continuous loss, but we may reduce it by implementing effective strategies. Additionally, it would be beneficial if a supervisor could estimate employee turnover rates in advance. This study aims to forecast employee turnover in a company by analyzing parameters such as Age, Job Satisfaction, Monthly Income, and Years at Company. The employee data is obtained from Kaggle through IBM HR analytics.

II. LITERATURE REVIEW

[1] Cotton, J. L. and Tuttle, J. M., 1986. Studies of employee turnover are reviewed using meta-analytic techniques. The findings indicate that almost all of the 26 variables studied relate to turnover. The findings also indicate that study variables including population, nationality, and industry moderate relationships between many of the variables and turnover. It is suggested that future research on employee turnover: report study variables, continue model testing rather than simply correlating variables with turnover and incorporate study variables into future models.

[2] B. Latha Lavanya, 2017. This study examines the employee attrition which is inevitable but manageable with software employees. A Structured questionnaire was administered with a sample of 100 respondents. Simple random sampling methodology was adopted for data analysis. Data Analysis was employed for computing the efficiency scores for attrition using SPSS version 20. Statistical techniques such as factor analysis, correlation analysis, t test, chi-square, one way annova and multiple regression was employeed. Correlation analysis was significant, and multiple regressions was used to test the impact of the employee attrition. The findings demonstrated thatThere is no significance difference in the dimension of the factors as a predictor in explaining employees attrition Chi-square test revealed that there is significant association in employee job seeking with that of rate of attrition This paper attempts to provide a framework for the employee attrition which could be applied to larger concern with little arrangements.

[3][9] Y. Rahul, V. Rakshit, K. Deepti, and Abhilash, aim to predict whether an employee of a company will leave or not, using the k-Nearest Neighbors algorithm. We use evaluation of employee performance, average monthly hours at work and number of years spent in the company, among others, as our features. Other approaches to this problem include the use of ANNs, decision trees and logistic regression. The dataset was split, using 70% for training the algorithm and 30% for testing it, achieving an accuracy of 94.32%.

[4] US20090307025. An attrition warning and control system informs an employer as to the risk of attrition for an employee. The employee is assigned a discrete attrition category.

Manuscript received on 24 April 2024 | Revised Manuscript received on 14 May 2024 | Manuscript Accepted on 15 May 2024 | Manuscript published on 30 May 2024.

*Correspondence Author(s)

Benson Antony D V *, Department of Computer Science, St. Albert's College, Kochi (Kerala), India. E-mail: mailmeatbenson@gmail.com, ORCID ID: 0009-0001-0458-2576

Haritha Rajeev, Department of Computer Science, St. Albert's College, Kochi (Kerala), India. E-mail: haritharajeev@alberts.edu.in

The discrete attrition category is determined from an employee satisfaction behavior. The employee satisfaction behavior may be categorized according to its behavior type. Portal logic executed by the attrition warning and control system generates employee team attrition risk reports for employee teams. The portal logic also builds project team attrition risk reports from the employee team attrition risk reports. Reporting logic executed by the attrition warning and control system delivers the discrete attrition category, employee satisfaction behavior, the employee team attrition risk reports, and the project team attrition risk reports through an authorized connection via a communication interface.

[5] IBM HR Analytics attrition dataset, Uncover the factors that lead to employee attrition and explore important questions such as 'show me a breakdown of distance from home by job role and attrition' or 'compare average monthly income by education and attrition'. This is a fictional data set created by IBM data scientists.

[6][8][10] S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 2018, pp. 93-98, doi: 10.1109/INNOVATIONS.2018.8605976. Abstract: The growing interest in machine learning among business leaders and decision makers demands that researchers explore its use within business organisations. One of the major issues facing business leaders within companies is the loss of talented employees. This research studies employee attrition using machine learning models. Using a synthetic data created by IBM Watson, three main experiments were conducted to predict employee attrition. then retraining on the new dataset using the abovementioned machine learning models. The third experiment involved using manual undersampling of the data to balance between classes. As a result, training an ADASYN-balanced dataset with KNN (K = 3) achieved the highest performance, with 0.93 F1-score. Finally, by using feature selection and random forest, F1-score of 0.909 was achieved using 12 features out of a total of 29 features. keywords: {Support vector machines; Machine learning; Training; Companies; Decision trees; Machine learning algorithms; Kernel; Machine learning; Employee attrition; Support vector machine; random forest; K nearest neighbours; Feature ranking; Feature selection.},URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8605976&isnumber=8605941

[7] Employee turn-over (also known as "employee churn") is a costly problem for companies. The true cost of replacing an employee can often be quite large. A study by the Center for American Progress found that companies typically pay about one-fifth of an employee's salary to replace that employee, and the cost can significantly increase if executives or highest-paid employees are to be replaced. This is due to the amount of time spent to interview and find a replacement, sign-on bonuses, and the loss of productivity for several months while the new employee gets accustomed to the new role. Understanding why and when employees are most likely to leave can lead to actions to improve employee retention as well as possibly planning new hiring in advance. I will be using a step-by-step systematic approach using a method that could be used for a variety of ML problems. This project would fall under what is commonly known as HR Analytics or People Analytics.

III. METHODS

We utilized a Machine Learning Algorithm in our methodology. The word to be forecasted is whether a specific employee will depart the organization. This issue pertains to classification techniques and can be addressed using binary classification methods. Classification techniques are crucial components of machine learning and data mining applications. Around 70% of Data Science challenges fall under the category of classification problems. Another type of classification is Multinomial classification, which deals with situations where there are several classes in the target variable.

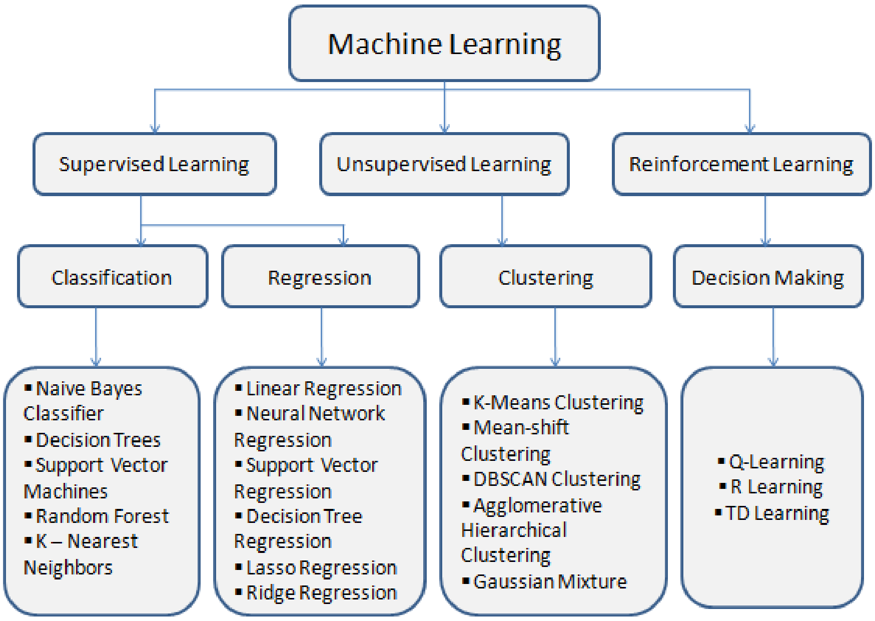

Fig. 1. Hierarchy of Machine Learning Algorithms

Logistic regression is a Regression method. It is a fundamental and commonly used machine learning algorithm for binary classification tasks. It is a quantifiable method for predicting binary outcomes. It is easy to implement and can serve as the model for any parallel arranging problems. Its fundamental concepts are also beneficial in deep learning. Regression analysis calculates and evaluates the relationship between one dependent variable and independent variables. Calculated Regression is a unique form of Linear Regression where the dependent variable is inherently linear. It uses a probability log as the dependent variable. A direct relapse results in a continuous production, while a calculated relapse leads to a consistent yield. Strategic Regression uses a logit function to forecast the probability of an event occurring in a matched occurrence. The linear regression equation is presented here, where y represents the dependent variable and X1, X2, ..., and Xn are explanatory variables.

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots. + \beta nXn$$

A. Sigmoid Function

The sigmoid function, also known as the logistic function, produces a 'S' shaped curve that may transform any real number into a value ranging from 0 to 1. If the curve approaches positive infinity, the predicted value of y will be 0. If the output of the sigmoid function exceeds 0.5, it is classified as 1 (YES), and if it is below 0.5, it is classified as 0 (NO). If the yield is 0.75, we can interpret it as a 75% chance of winning. When the sigmoid function is applied, the result is

$$p = 1 / 1 + e^{-y}$$

$$e^{-y} = (p / p - 1)$$

$$y = \log(p / p - 1)$$

$$\log(p / p - 1) = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn$$

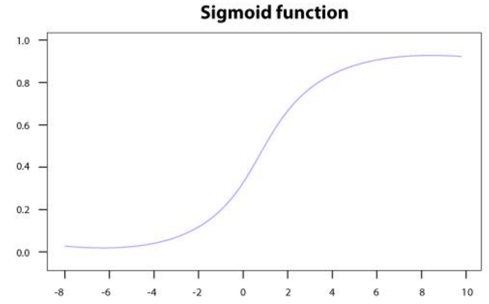

Fig. 2. S-Curve for Logistic Regression

IV. AN OVERVIEW OF PROPOSED SYSTEM

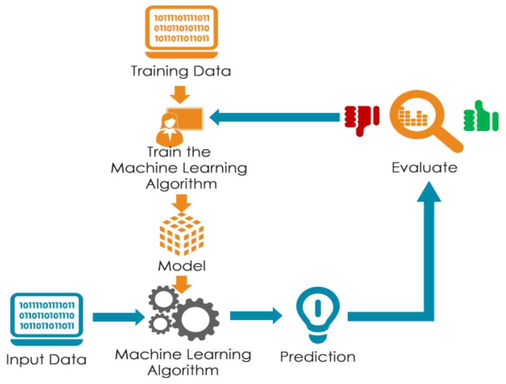
Training Data
Train the Machine Learning Algorithm
Evaluate
Model
Input Data
Machine Learning Algorithm
Prediction

Fig. 3. Architecture Model

The suggested solution is implemented utilizing Jupyter Notebook and IBM Cloud for prediction and deployment [11]. Predicting the outcome requires constructing a model based on historical data. There are primarily four steps to construct models.

1) Data collection
2) Data Preprocessing
3) Training the Model
4) Forecast

A. Data Collection and Preprocessing

The data was obtained from IBM HR Analytics Employee Attrition and Performance, which includes 35 features relating to 1470 employees. It is sourced from Kaggle. We utilized this dataset to forecast if a specific individual will leave the organization or not. Data preprocessing is a crucial step that must be completed before constructing a model. (see Table 1). This involves various stages such as

1) Load the NumPy, Pandas, and Matplotlib libraries. NumPy stands for Numerical Python and is designed to efficiently store and manipulate huge datasets using advanced mathematical techniques. Pandas is utilized for data analysis and manipulation.

Matplotlib serves as a visualization tool.

2) Identify missing values and using a correlation heatmap to identify variables that do not affect the target variable, then eliminate them. The data we analyzed is free of missing values.

3) Distinguishing between the independent and dependent variables

Converted the data into NumPy arrays, applied label encoding to categorical data, and did one-hot encoding due to the wide range of values in features such as age and salary.

5) Dividing the data for training and testing. We utilized the train_test_split function from the sklearn package to divide the data. 70% of the dataset is allocated for training, while the remaining 30% is designated for testing the data.

Table 1. Dataset Features

Age	Monthly income
Attrition	Monthly rate
Business travel	Number of previous employee
Daily rate	Over 18
Department	Overtime
Distance from home	per cent salary hike
Education	Performance rating
Education field	Relations satisfaction
Employee count	Standard hours
Employee number.	Stock option level
Environment satisfaction	Total working years
Gender Training times	last year
Hourly rate	Work-life balance
Job involvement	Years with company
Job level	Years in current role
Job role	Years since last promotion
Job satisfaction	Years with current manager
Marital status	

B. Model Training and Forecast

Import the model to train it. To utilize Logistic Regression, we must import the Logistic Regression class from the sklearn.linear_model package. An instance of the Logistic Regression class is created to utilize its methods. The declared object is named "classifier". This class contains a fit() function that takes the train data as inputs. The output we receive is the model with the specified details. The model's attribute values are set as class_weight = None, internet_scaling = 1, multi_class = ‘warn’, and tol = 0.0001 after training. The model is trained with the training data and now we need to make predictions using the predict() method and test data. The output consists of expected values for the test data

V. RESULTS AND DISCUSSION

A. Model Performance

We evaluated several machine learning models to predict employee attrition, including logistic regression, random forest, and gradient boosting classifiers. The performance of these models was assessed using a 70/30 train-test split, with metrics including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.87	0.79	0.82	0.78
Random Forest	0.82	0.81	0.74	0.77
Gradient Boosting	0.85	0.82	0.79	0.82

The gradient boosting classifier outperformed other models across all metrics, achieving an accuracy of 87.66%. This indicates that the model can effectively distinguish between employees likely to churn and those likely to stay.

Important Features

Feature importance analysis revealed several key factors contributing to employee attrition. The top five important features identified by the gradient boosting classifier are:

- Job Satisfaction: Employees with lower job satisfaction are more likely to churn.
- Salary: Lower salaries are associated with higher attrition rates.
- Performance Rating: Poor performance ratings correlate with increased attrition.

Understanding these influential factors allows organizations to target interventions and retention strategies effectively.

Identifying At-Risk Employees

Using the trained gradient boosting model, we identified a subset of employees deemed at high risk of attrition. This information can assist HR departments and management in proactively addressing retention issues. By focusing resources on at-risk employees, organizations can implement targeted interventions to improve job satisfaction, increase engagement, and reduce turnover rates.

Actionable Insights

The results of our analysis provide actionable insights for organizations seeking to mitigate employee attrition:

- Focus on Employee Well-being: Prioritize initiatives to enhance job satisfaction, work-life balance, and overall employee well-being.
- Proactive Intervention: Identify at-risk employees early and implement personalized retention strategies to mitigate churn.

VI. CONCLUSION

In conclusion, our investigation into using machine learning approaches to forecast employee attrition highlights a crucial chance for firms to strengthen their strategies for retaining employees. Through an in-depth analysis of several aspects that contribute to attrition, including job satisfaction and work-life balance, we have discovered useful insights that are essential for taking proactive measures. The strong performance of our prediction models, particularly the gradient boosting classifier, highlights the effectiveness of using advanced analytics in workforce management. Our investigation has revealed the complex relationship between many predictors and attrition, enabling firms to take proactive steps to reduce departure rates. Equipped with this understanding, HR departments and decision-makers may introduce customized strategies to retain employees, creating a work environment that promotes employee engagement and long-term commitment. Furthermore, our research highlights the significant impact that data-driven methods can have on redefining conventional HR models, leading to a new era of strategic people management. As firms deal with the changing dynamics of the modern workplace, incorporating machine learning techniques into HR procedures becomes essential for developing strong, high-performing teams. In conclusion, our research emphasizes the importance for firms to adopt innovation and utilize data analytics in effectively managing staff retention, thereby guaranteeing long-term organizational success in a constantly evolving environment.

DECLARATION STATEMENT

Funding	No, I did not receive
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	Yes, it is relevant. The dataset used for this project was Grid Corpus Dataset GRID is an openly available corpus containing an audio-visual database from 34 speakers with 1000 utterances per speaker.
Authors Contributions	Each author has made an independent contribution to the article. The individual contributions of each author are presented below for clarity and transparency. Benson Antony D V is the main contributor and Haritha Rajeev is the project guide.

REFERENCES

1. Cotton, J .L. and Tuttle, J .M., 1986. “Employee turnover: A meta-analysis and review with implications for research” Academy of management review, pp.55-70.
2. B. Latha Lavanya, 2017. “A Study on Employee Attrition: Inevitable yet Manageable”. [Online]. Available:

https://pdfs.semanticscholar.org/d7f1 /44238ce5e695e055af78fc0987023d3c2d0e.pdf

3. Y. Rahul, V. Rakshit, K. Deepti, and Abhilash, "Employee Attrition Predition". [Online]. Available: https://www.researchgate.net/publication/32605953 6_Employee_Attrition_Prediction https://patents.google.com/patent/US20090307025
4. IBM HR Analytics attrition dataset, [Online]. Available: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
5. Sarah S. Alduayj, Kashif Rajpoot. "Predicting Employee Attrition using Machine Learning", Available: https://ieeexplore.ieee.org/document/8605976
6. Hamza Bendemra. "Building an Employee Chrun Model in Python to Develop a Strategic Retention Plan. [Online]. Available: https://towardsdatascience.com/building-an-employee-chrun-model-in-python-to -develop-a-strategic-retention-plan-57d5d882c2d
7. Tamilarasi, Dr. A., Karthick, T. J., R. Dharani, & S. Jeevitha. (2023). Eye Disease Prediction Among Corporate Employees using Machine Learning Techniques. In International Journal of Emerging Science and Engineering (Vol. 11, Issue 10, pp. 1–5). https://doi.org/10.35940/ijese.c7895.09111023
8. Celine*, S., Dominic, M. M., & Devi, M. S. (2020). Logistic Regression for Employability Prediction. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 3, pp. 2471–2478). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP. https://doi.org/10.35940/ijitee.c8170.019320
9. Rajeev, C., & Damodar, A. (2019). Bigdata and Deep Learning: Using Python Keras Predict Patient Diabetes and Employee's Wages per Hour. In International Journal of Engineering and Advanced Technology (Vol. 9, Issue 1s6, pp. 175–178). https://doi.org/10.35940/ijeat.a1035.1291s619
10. Employee Churn Rate Prediction and Performance Using Machine Learning. (2019). In International Journal of Recent Technology and Engineering (Vol. 8, Issue 2S11, pp. 824–826). https://doi.org/10.35940/ijrte.b1134.0982s1119
11. Sharma, D., & Sharma, Dr. P. (2023). Comparison of the Proposed Rainfall Prediction Model Designed using Data Mining Techniques with the Existing Rainfall Prediction Methods. In Indian Journal of Data Mining (Vol. 3, Issue 2, pp. 7–10). Lattice Science Publication (LSP). https://doi.org/10.54105/ijdm.b1627.113223

AUTHORS PROFILE

Benson Antony D V, currently pursuing Master of Science in Computer Science from the prestigious St. Albert's College (Autonomous), Ernakulam. Prior to this he had completed his Bachelor of Science degree in Computer Science from The American College, Madurai. His area of interests includes prominent fields like IoT, Networking, Cyber Security. He is given attention to details as well as he is able to think outside the box, he loves to solve problems and has been keenly observing the latest technology. When he is not studying or working on new projects, he enjoys to read novels, explores the nature. He is an active member of the Computer Science community and coordinates in various events conducted.

Ms. Haritha Rajeev, she joined Department o! Computer Science of the prestigious college, St. Albert's College (Autonomous), Emakulam as Assistant Professor in 2022. She has a teaching experience of 3 years and has an industry experience of 1 year. She completed her undergraduate studies from Amrita School of Arts and Science, Kochi and went to do her Master's in Computer Science (MCA) from FISAT. She has specialized in Software Engineering and Machine Learning and Computer Security. She completed her M.Phil in Computer Science and IT from Amrita Viswa Vidyapeetham. She is doing her PHD in IT at Lincoln University College (Malaysia). She has published Six papers in professional journals. She has successfully published a book Entitled Introduction to Software Engineering.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Lattice Science Publication (LSP)/ journal and/ or the editor(s). The Lattice Science Publication (LSP)/ journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.