Predicting Employee Attrition Using Machine Learning Approaches

GUNASREE M.[1], JEEVA AYER J[2], IMMANUEL C.[3], HARISH BALA D[4], DR. S THAIYALNAYAKI[5]
[1, 2, 3, 4] *Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India*
[5] *Assistant Professor, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India*

Abstract*— Organizations today have a plethora of technological options at their disposal to bolster decision-making processes, with artificial intelligence (AI) emerging as a frontrunner in innovation. AI is being leveraged across various domains to aid organizations in shaping business strategies, streamlining organizational operations, and optimizing people management practices. Particularly, in recent years, there has been a growing emphasis on the significance of HR as the quality and skills of the workforce increasingly become pivotal for organizational growth and competitive advantage. Initially embraced by sales and marketing departments, AI is now making significant inroads into HR management, aiming to guide decisions pertaining to employees based on objective data analysis rather than subjective assessments. The overarching objective is to delve into how tangible factors impact employee attrition, deciphering the primary drivers behind an employee's decision to depart from a company. By harnessing AI-driven analytics, organizations aspire not only to identify the root causes of attrition but also to forecast the likelihood of individual employees leaving the company. This endeavor represents a paradigm shift towards evidence-based decision-making in HR, empowering organizations to proactively address retention challenges and optimize workforce stability. Through the judicious utilization of AI technologies, organizations can cultivate a deeper understanding of employee dynamics, thereby fostering an environment conducive to talent retention and organizational resilience.

Indexed Terms- Artificial Intelligence, Human Resource, Employee Attrition, Knowledge Economy

I. INTRODUCTION

In the contemporary landscape of a fiercely competitive economy driven by technological advancements, the acquisition, study, and analysis of data have emerged as the bedrock of a burgeoning "knowledge economy." Information technologies serve not merely as repositories of data but as pivotal enablers for data analysis, facilitating the processing of vast data sets and the extraction of actionable insights. Data, once a mere byproduct, has now evolved into a strategic asset for organizations spanning diverse sectors, revolutionizing traditional business processes.

The advent of artificial intelligence (AI) has revolutionized decision-making across various domains within organizations. Human resources (HR), in particular, have garnered increased attention as companies recognize the pivotal role of employee quality and skill sets in gaining a competitive edge. Beyond its conventional applications in sales and marketing, AI is now revolutionizing HR management, empowering companies to base decisions on objective data rather than subjective evaluations.

Maximizing profits remains a primary objective for businesses, necessitating a nuanced approach to workforce management. While gig economy models may suffice for tasks of lower complexity, specialized roles demand a focus on employee specialization and continuity. The significance of skills, knowledge, and ongoing learning underscores the transformative potential of AI in HR, where predictive models harness historical data to optimize HR activities and mitigate critical issues.

Employee attrition poses a significant challenge for organizations, resulting in the loss of valuable resources invested in recruitment, training, and development. Mitigating attrition requires proactive measures informed by data-driven insights into employee motivations. This paper embarks on an analysis of the factors driving employee attrition, offering a classification model based on statistical evaluations. Leveraging a real dataset from IBM analytics, the study identifies key attributes correlated with attrition, with the Gaussian Naïve Bayes classifier emerging as the optimal algorithm for prediction.

In conclusion, the integration of AI-driven analytics holds immense potential for revolutionizing HR management, enabling organizations to retain talent, enhance productivity, and maintain a competitive edge in today's dynamic business landscape.

II. LITERATURE SURVEY

(i). Paper Title: "Predicting Employee Attrition Using Machine Learning: A Comparative Study"

For businesses in all sectors, employee attrition presents serious difficulties since it affects output, morale, and eventually the bottom line. Researchers have used machine learning approaches to create prediction models that can detect people who are at danger of leaving a business, realizing the value of proactive attrition management. Utilizing data-driven methodologies, like decision tree classifiers, support vector machines (SVM), and neural networks, enables firms to obtain significant insights into the elements that lead to employee turnover and to devise focused retention strategies that lessen its impact.

One seminal study in this field involved the analysis of the IBM Human Resource Analytics Employee Attrition and Performance dataset using various machine learning models. Through rigorous preprocessing steps, including data exploration, cleaning, and feature selection, researchers aimed to develop accurate predictive models for employee attrition. By optimizing parameters and applying regularization techniques to mitigate overfitting, the study compared the performance of different classifiers and identified SVM as the top-performing model, achieving an impressive accuracy rate of 88.87%. These findings underscore the potential of machine learning to revolutionize attrition management by enabling proactive interventions and strategic workforce planning.

(ii). Paper Title: "Understanding Employee Attrition: A Literature Review"

Moreover, literature reviews have highlighted the multifaceted nature of employee attrition and the diverse factors influencing it. Studies have emphasized the significance of demographic attributes, job-related factors, and organizational culture in shaping attrition rates. For instance, research has shown that compensation, job satisfaction, and career advancement opportunities play pivotal roles in employee retention. By analyzing historical data and identifying patterns, machine learning algorithms can effectively capture these complex relationships and provide actionable insights for decision-makers.

(iii). Paper Title: "Addressing Class Imbalance in Employee Attrition Prediction: A Comparative Study

In addition to traditional machine learning techniques, recent studies have explored innovative approaches to address the challenges of attrition prediction. For example, one study proposed the use of adaptive synthetic (ADASYN) sampling to overcome class imbalance issues in the dataset. By generating synthetic minority samples, ADASYN effectively balanced the dataset and improved the performance of machine learning models. Similarly, another study applied manual undersampling techniques to achieve a more balanced representation of classes, leading to enhanced predictive accuracy.

(iv). Paper Title: "Enhancing Employee Attrition Prediction with Neural Networks: A Comparative Analysis"

Furthermore, the emergence of neural network techniques has expanded the scope of attrition prediction, offering high-speed processing and scalability for analyzing large datasets. By leveraging feedforward and feedback propagation networks, researchers have achieved significant improvements in predictive accuracy, surpassing traditional machine learning models in some cases. These advances highlight the potential of neural networks to address the complexities of attrition prediction and facilitate more informed decision-making in human resource management.

(v). Paper Title: "Ethical Considerations in Predicting Employee Attrition: Safeguarding Rights and Privacy"

Despite these advancements, challenges remain in effectively predicting employee attrition and implementing actionable strategies to mitigate its impact. Issues such as data imbalance, model interpretability, and ethical considerations pose significant hurdles for researchers and practitioners alike. Addressing these challenges requires interdisciplinary collaboration and ongoing innovation to develop robust and reliable predictive models.

Moreover, the ethical implications of attrition prediction cannot be overlooked, as the use of predictive analytics raises concerns about privacy, fairness, and bias. Organizations must ensure transparency and accountability in their use of machine learning models, taking proactive steps to mitigate potential risks and safeguard employee rights.

III. RELATED WORKS

Existing System

Employee attrition, the natural reduction of employees within an organization, occurs due to various unavoidable factors.

This attrition leads to substantial losses for companies, with the SHRM estimating the average cost-per-hire to be USD 4129.

Recent statistics indicate a 57.3% attrition rate in 2021. Addressing the causes of employee attrition and developing a predictive model are crucial for organizational stability.

Drawbacks

The complexity of existing models, especially compared to decision trees, can pose challenges in decision-making.

Training time is prolonged due to this complexity, as each decision tree must process input data individually for predictions.

Proposed System

Data sourced from Kaggle is initially pre-processed to extract pertinent features such as Monthly Income, Last Promotion Year, and Salary Hike, which are indicative of employee attrition.

Exploratory Data Analysis is conducted to summarize data characteristics and predict employee terminations. The proposed system utilizes the random forest technique, incorporating logistic regression and word vector formation for analysis.

By focusing on improving employee morale and creating conducive working environments, the proposed system aims to mitigate attrition significantly.

IV. METHODOLOGY

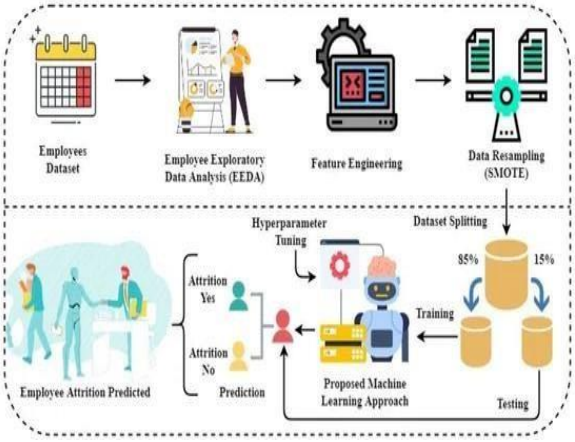

Fig.1 Architecture diagram

- Dataset collection
- Data Cleaning
- Feature Extraction
- Data training
- Data testing
- Performance Evaluation
- Prediction

a. Dataset Collection

The first step in implementing predictive analytics for employee retention is gathering relevant historical data. This dataset encompasses diverse factors such as employee demographics, job roles, performance metrics, satisfaction surveys, attendance records, and salary information. Obtaining this data from multiple sources provides a comprehensive understanding of the organization's workforce dynamics over time. Platforms like Kaggle dataset archives offer valuable repositories of historical data, facilitating the collection process.

b. Data Cleaning

Once the dataset is compiled, the process of data cleaning becomes indispensable to ensure its accuracy and reliability. This phase involves meticulous identification and rectification of inconsistencies, errors, duplicates, and missing values within the dataset. Data scientists employ a variety of statistical analysis and visualization techniques to meticulously explore the dataset, detecting anomalies and executing necessary cleaning operations. The integrity of the data is foundational as the efficacy of predictive models hinges on the quality of the data upon which they are constructed.

c. Feature Extraction

Feature extraction assumes a pivotal role in reducing the dimensionality of the dataset while preserving pertinent information for analysis. By extracting informative and non-redundant features from the initial set of measured data, this process facilitates pattern recognition and enhances the efficiency of machine learning algorithms. Techniques such as PCA and feature selection methods aid in identifying the most influential attributes contributing to employee retention. Feature extraction not only expedites the training process but also enhances the accuracy of predictive models by focusing on key predictors.

d. Data Training

Armed with the refined and feature-engineered dataset, the subsequent step involves training machine learning algorithms to predict employee attrition. Supervised classification algorithms like logistic regression, decision trees, and random forests are commonly employed for this purpose. During the training phase, the model learns to correlate input features with the corresponding output labels, determining whether an employee is likely to stay or leave the organization. Iterative processes such as model fitting optimize the algorithm's parameters to enhance predictive accuracy.

e. Data Testing

Following the training of the predictive model, it is crucial to assess its performance using a separate test dataset. This dataset comprises unseen examples that were not utilized during the training phase, providing an unbiased evaluation of the model's effectiveness. By comparing the model's predictions against the actual outcomes, data scientists can evaluate its accuracy, precision, recall, and other performance metrics. Rigorous testing ensures that the predictive model generalizes well to new data and can reliably forecast employee attrition in real-world scenarios.

f. Performance Evaluation

Performance evaluation is a critical aspect of predictive analytics, enabling organizations to assess the effectiveness of their retention strategies. Metrics such as F1 score, accuracy, precision, recall, and confusion matrix analysis offer insights into the model's predictive capabilities. In instances of suboptimal performance, iterative refinement techniques, including algorithm optimization and feature engineering, can be employed to enhance the model's accuracy and robustness. Continuous monitoring and evaluation ensure that the predictive model remains aligned with evolving organizational dynamics.

g. Prediction

The ultimate goal of predictive analytics in employee retention is to forecast the likelihood of attrition and take proactive measures to mitigate it. Leveraging the trained and validated machine learning model, organizations can identify at-risk employees and implement targeted interventions to improve retention rates. Predictive analytics enables HR professionals to anticipate potential turnover trends, allocate resources effectively, and foster a culture of employee engagement and satisfaction.

B. Algorithm Used

Logistic Regression

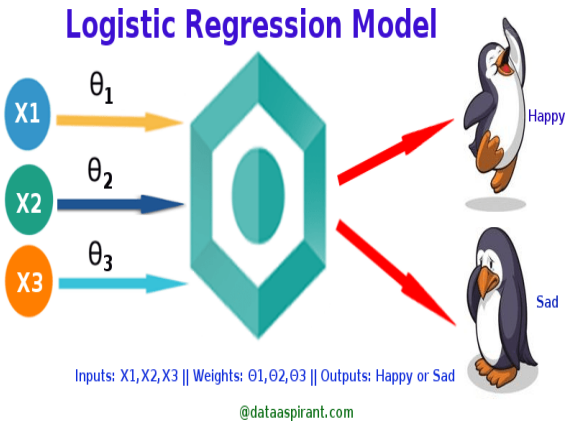

Fig.2 Logistic Regression

Logistic regression emerges as a pivotal tool in regression analysis when dealing with a dichotomous (binary) dependent variable. In contrast to traditional regression analyses, logistic regression primarily serves as a predictive model. Its application extends to both describing data and elucidating the intricate relationship between a single dependent binary variable and one or more independent variables of nominal, ordinal, interval, or ratio-level measurement.

In the realm of Machine Learning, logistic regression stands as a borrowed statistical analysis method, strategically employed when confronted with a dichotomous or binary dependent variable. This

scenario typically involves outcomes characterized by two distinct possibilities, such as survival or fatality in an accident, or passing or failing an exam. In essence, logistic regression mirrors the principles of linear regression but is uniquely tailored to predict probabilities in classification problems, navigating the intricacies of binary outcomes with precision and efficacy.

C.WORKING PRINCIPLE

a. Data Preparation Phase:

Begin by structuring the data in a tabular format where each row corresponds to a distinct observation and each column signifies a different variable. It's imperative that the target variable, which is the variable intended for prediction, is dichotomous in nature, representing binary outcomes such as yes/no, true/false, or 0/1.

b. Model Training Stage:

Initiate the model training process by exposing it to the training dataset. This entails iteratively adjusting the model parameters to minimize the error between predicted and actual values within the training data.

c. Model Evaluation Step:

Gauge the model's efficacy by subjecting it to an independent test dataset. This evaluation phase provides insights into the model's performance when confronted with unseen data, thereby validating its predictive capabilities.

d. Prediction Utilization:

Once the model has undergone rigorous training and evaluation, leverage its learned patterns and relationships to make predictions on new, previously unseen data instances. By applying the trained model to fresh data, anticipate and forecast outcomes with a heightened level of accuracy and reliability.

V. RESULTS AND DISCUSSIONS

The evaluation approach for the Multiple Linear Regression (MLR) model closely mirrored that of other Machine Learning (ML) models. Accuracy assessments produced outcomes akin to those observed with ML models, mirroring the distribution of job retention and attrition within the training dataset. Conversations surrounding dataset partitioning and the influence of input variables remained pertinent during MLR model testing. Multicollinearity, evidenced by high correlations among independent variables, likely affected the outcomes, complicating the differentiation of each variable's individual impact on the predicted outcome. Fluctuations in the assigned weights of independent variables across tests indicated variations in their influence on the predicted outcome. Particularly, the variable "community," representing employees' workplace camaraderie, displayed inconsistency in its predictive impact across tests. This inconsistency challenges established research highlighting the importance of community in job retention decisions, suggesting potential inaccuracies in the model's predictions.

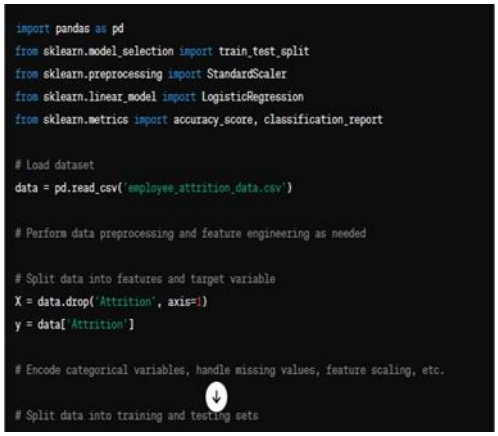

Fig.3 Libraries

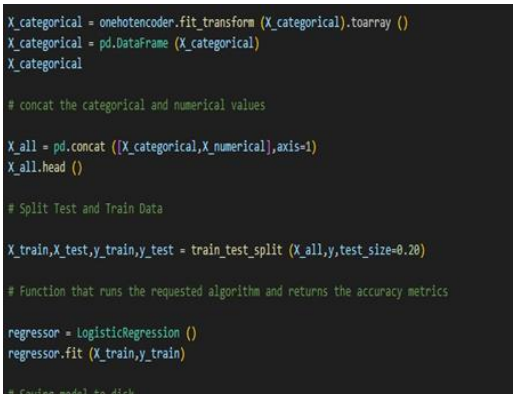

Fig.4 Prediction

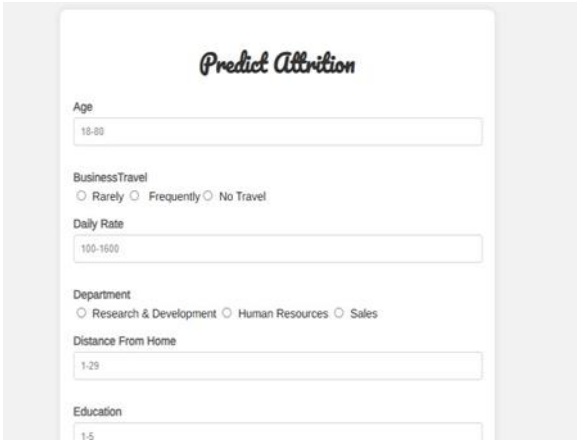

Fig.5 Predict Attrition

VI. CONCLUSION AND FUTURE WORKS

This research endeavors to anticipate employee attrition by employing logistic regression alongside feature selection techniques. Three distinct feature selection methods—information gain, select k-best, and recursive feature elimination (RFE)—are compared against logistic regression classification. Notably, logistic regression sans feature selection achieves a commendable accuracy of 0.865 and an AUC of 0.932. Although feature selection fails to surpass this accuracy, its implementation streamlines data training. Among the feature selection methodologies, RFE emerges as the most effective in predicting employee attrition when applied alongside logistic regression, particularly when selecting the top 20 features. This approach yields an accuracy of 0.853 and an AUC of 0.925, signifying excellent classification. These top 20 features, determined by RFE, are identified as the most influential variables.

Future Works

It is imperative to delve deeper into identifying models accurate enough for real-world implementation. Additionally, a comprehensive understanding of pertinent variables and data for predicting employee attrition warrants further investigation. Exploring the impact of such models on individual employees—whether positive or negative—presents an intriguing avenue for research. Moreover, extending this technique to predict other phenomena such as sick leave, motivation, and salary is an area ripe for exploration.

REFERENCES

[1] Society for Human Resource Management. Human Capital Benchmarking Report; Technical Report; Society for Human Resource Management: Alexandria, VA, USA, 2016. [Google Scholar]

[2] Ongori, H. A review of the literature on employee turnover. Afr. J. Bus. Manag. 2007, 1, 49–54. [Google Scholar]

[3] Bennett, N.; Blum, T.C.; Long, R.G.; Roman, P.M. A firm-level analysis of employee attrition. Group Organ. Manag. 1993, 18, 482–499. [Google Scholar] [CrossRef]

[4] Alao, D.; Adeyemo, A. Analyzing Employee Attrition using Decision Tree Algorithms. Comput. Inf. Syst. Dev. Informatics Allied Res. J. 2013, 4, 17–28. [Google Scholar]

[5] Punnoose, R.; Ajit, P. Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. Int. J. Adv. Res. Artif. Intell. 2016, 5. [Google Scholar] [CrossRef] [Green Version]

[6] Shaw, D. Jason. 2010. Turnover rates and organizational performance: Review, critique, and research agenda. Organizational Psychology Review1(3). 187-213. doi:10.1177/2041386610382152

[7] Ajit, P. & Punnoose, R. 2016. Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. A case for Extreme Gradient Boosting. International Journal of Advanced Research in Artificial Intelligence 5(9). 22-26. DOI: 10.14569/IJARAI.2016.050904

[8] Babor, F. Thomas; Stenius, Kerstin; Pates, Richard; Miovský, Michal; O`reilly, Jean & Candon, Paul. 2017. Publishing Addiction Science: A Guide for the Perplexed.

[9] Miles, Jeremy & Gilbert, Paul. 2005. A Handbook of Research Methods for Clinical and Health psychology. Oxford University Press.

[10] Shah, K. Sonali & Corley, G. Kevin. 2006. Building Better Theory by Bridging the Quantitative- Qualitative Divide. Journal of Management Studies 43(8). 1821-1835.

https://doi.org/10.1111/j.1467-6486.2006.00662.x

[11] Baxter, Pamela & Jack, Susan. 2008. Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. The Qualitative Report 13(4). 544-559. https://nsuworks.nova.edu/tqr/vol13/iss4/2/

[12] Jason, S.D. (2010). Turnover rates and organizational performance: Review, critique, and research agenda. Organizational Psychology Review, 1(3), 187-213. https://doi.org/10.1177/2041386610382152

[13] Punnoose, R., & Ajit, P. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. A case for Extreme Gradient Boosting. International Journal of Advanced Research in Artificial Intelligence, 5(9), 22-26. DOI: 10.14569/IJARAI.2016.050904

[14] Babor, F.T., Stenius, K., Pates, R., Miovský, M., O`reilly, J., & Candon, P. (2017). Publishing Addiction Science: A Guide for the Perplexed. Miles, J., & Gilbert, P. (2005). A Handbook of Research Methods for Clinical and Health psychology. Oxford University Press.