JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Understanding Predicting Employee Attrition using Machine Learning

Prof. Jyotsna A Nanajkar[#1],
[#] Information Technology Department,
Savitribai Phule Pune University,
Zeal College of Engineering and Research,
Pune, Maharashtra, India

Pratik Marathe[#2],
Nikhil Mandavkar[#3],
Shubham Marshettiwar[#4],
Gaurav Patil[#5],
[#] Information Technology Department

*Abstract— **The increasing popularity of machine learning in the business world is due to its ability to provide valuable insights and support data-driven decision-making. Consequently, organizations may make better judgments, work more effectively, and perform better as a whole.***

A major problem facing companies across a range of sectors is employee attrition. Attrition can be caused by several factors, such as poor management and inadequate working conditions, making it important for organizations to analyze their policies to increase employee retention.

The goal is to prevent or minimize employees leaving a company before hiring a replacement. Considering the recent influx of data, researchers in the field of artificial intelligence have also addressed this problem. Using a Random Forest Classifier and data that was made available to the public, this study addressed the problem of employee attrition. While evaluating the efficiency of a classification model, we employed a Confusion matrix, which consists of an N x N matrix, where N is the total number of target classes.

The model is accurately recognized with the aid of the confusion model and random forest classifier, and the accuracy of the employees who left the company is higher than other models.

*Keywords— **Machine learning; Employee Attrition; Random Forest(RF) classifier; Confusion Matrix; Decision Tree(DT);***

I. INTRODUCTION

Employee attrition is the term used to describe when employees leave a company for any of the following reasons: personal, work-related, financial, or bad business environment. Employee turnover is one of the key issues faced by HR managers today. Headcount reductions within an organization not only drive up the cost of employee training but also the cost of production, along with the ideal "knowledge maturity" of the organization and its employees. Organizations that can endure long-term relationships with their employees will survive in the market, while others will perish in the long run. The ultimate solution for the organizations are to stabilize business operations and the ensure the availability of an educated and mature workforce in abundance.

In this study, we propose a rapid and accurate strategy for forecasting employee attrition. The suggested technique aims to anticipate Employee Attrition early, eliminate false positives, and enhance the system's learning process. Employee attrition happens when your staff shrinks over time due to unavoidable circumstances like employee departure for personal or professional reasons. Employees are departing the job at a quicker rate than they are being hired, and this is sometimes beyond the employer's control.

By improving internal rules and methods, management will be able to respond more quickly when it may anticipate staff attrition. Where skilled workers or employees with the risk of leaving might be provided a number of incentives, such as a pay raise or appropriate training, to minimize their chance of leaving. Businesses that use machine learning algorithms can forecast staff turnover. One potential application of machine learning is in predicting employee turnover in organizations. By utilizing historical data stored in HR departments, analysts can create and maintain models that can forecast which employees are likely to leave the organization. The approach involves training the model to analyse similarities and differences between determined and terminated employees' characteristics. With this information, the model can identify patterns and develop a predictive algorithm that can help organizations proactively identify and address issues that may lead to high employee turnover. However, it's important to note that the accuracy of such models can be affected by a variety of factors, including the quality of data and the assumptions made during the model-building process. As such, it's important for analysts to carefully consider the limitations and potential biases of their models when using them to make critical business decisions.

II. LITERATURE REVIEW

Voluntary employee attrition is one of the major concerns for any company due to the severity of its impact. Talented employees are a major factor in business success and replacing such talent can be difficult and time-consuming [1]. Employee attrition is a problem that has been a focus of research for the last few decades. Much work mainly in the field of psychology, has tried to address this issue [2], [3]. The better the compensation, the lower the attrition rate. Whereas [1] found that money is not the only factor, as other combinations of factors, such as workload, performance pay, and a weak career plan, have increased the attrition rate in the retail industry.

Some industries like call centers tend to have a high attrition issue compared to others, but in general, this affects all industries and organizations [4]. Various research studies have investigated the utilization of machine learning to anticipate employee actions. According to these studies, the job title has emerged as the most significant feature, while age does not appear to have a clear impact. In a specific study cited as [5], the authors examined multiple data mining algorithms to forecast employee churn or attrition. Their analysis was based on a dataset that included 1470 records and 35 features. The study [5] used multiple machine learning algorithms such as naive Bayes, SVM stands for support vector machines, (LR) logistic regression, decision trees, and random forests (RF). The research results recommend using a support vector machine (SVM), which has 84.12% accuracy.

In [6], different decision tree algorithms were explored, including C4.5, C5, REPTree (The reduced error pruning (REP) method can be applied to an existing dataset to construct a decision tree. This technique is an extension of the C45 algorithm, utilizing Reduced Error Pruning (REP) to enhance the pruning stage of decision tree construction.) and classification, and regression trees (CART). The researchers used a dataset of 215 employee records to train and test decision tree models through experiments. The dataset was sampled from a larger set of 1470 records. The study revealed that the C5 decision tree had the highest accuracy rate of 74% compared to other decision trees. The study revealed that the length of service and employee's salary were identified as essential variables in predicting employee turnover based on the analyzed data. These factors have a significant impact on employee retention. Authors in [7] used neural networks to predict the turnover rate for a small-west manufacturing company. Consequently, they developed the neural network simultaneous optimization algorithm (NNSOA) alongside 10-fold cross-validation, which predicted the turnover rate with 94% accuracy. Moreover, they were able to identify the most important, relevant 'Tenure of an employee on January 1' by using a modified genetic algorithm. In [8], a total of 6,909,746 employees' profiles available on the web were used to predict employee attrition. The employees' profiles included work experiences and education information along with company information. The researcher was able to train and evaluate an SVM model. The model prediction had 55% average accuracy, which is obviously not very high. The researcher recommended adding more personal features to the dataset, such as employee age, gender, and work environment, which could improve the trained model. Overall, the attrition issue is crucial. While there are many reasons for this, the three reasons mentioned below are significant:

[9] Cost implications: Employee resignation causes the company to lose productivity. There are other economic costs related to employee attrition, such as the company must pay the workers who are handling the leaving employee's work until the organization hires new ones. Companies need to spend money on job advertising, and interviewing potential substitutes, in addition to the fees related to the actual recruiting and of hiring an employee.

[10] Overall business performance: Employees in companies with high turnover rates are less productive and much less efficient than they might have been in a lower turnover environment. This also makes a company with a low retention rate less competitive and productive as they do not have a stable workforce. According to Huselid, high-performance work practices and firm

performance have an economically and statistically significant impact on both employee turnover and short and long-term measures of corporate financial performance.

[11] Challenging to control company environment: Research consistently displays that employees change jobs more often because of the work environment or antiemployee relations rather than because of the difficulty of the job. This is something that organizations have little or no control over as they cannot interfere with the employees' relationships or feelings. Shalley et al. conducted an extensive study and found that there is a direct link between job satisfaction and fewer instances of intentions to leave.

TABLE I. Literature Review

Year	Author	Paper Name	Algorithms Used	Acc. %
2019	Jyotindra Dubey, Anil Yadav	Understanding Employee Attrition using Decision Tree	Decision Tree, SVM, Logistic regression.	81.7 86.5 78.6
2022	G.R. Rajeswari, Murugesan, R Arunan	Predicting Employee Attrition through ML	SVM, Random Forest, KNN	84.1, 85.6, 82.7
2019	P. Sahu, S. Jena, P.Sarangi	Predicting Employee Attrition using SLC Models	Supervised Learning classification	82- 88.8
2021	T. Tsai, K. Lin, and Y. Hu	Employee Attrition Using ML And Depression Analysis	DT RF	82.5 83.5
2020	RSangeetha, R.Vidhya	Employee Attrition Prediction System	SVM Logistic Regression	86.4 81.3
2020	S Nagendra, Anbazhagan ,S.Ravichandran	Predicting Employee Attrition Using Machine Learning Approaches	DT, SVM, Logistic Regression	83 87.4 72.8
2020	S. Al-Ani, S. Alzahrani,, S. Almujaini	Analyzing Employee Attrition Using Machine Learning	Decision trees, Random forests, logistic regression	84.7 92.5 82.4

[16].

III. PROPOSED METHODS

In this research, we aim is to understand the factors that contribute to employee turnover in an organization. The research methodology used in an employee attrition study typically involves the following steps:

1. Data collection: Collect information on employee attrition, demographics, job satisfaction, pay, and other pertinent variables. This can be accomplished using surveys, interviews, or information already in the company.
2. Data analysis: Analyze the data using Machine learning models such as Random Forest, SVC, K-Nearest Neighbors, and Decision tree.
3. Model development: Develop a model to predict employee attrition based on the findings from the data analysis.
4. Validation: Validate the model by using a holdout sample or cross-validation techniques.
5. Recommendations: Based on the results of the study, provide recommendations for reducing employee attrition and improving employee retention in the organization.

It's important to note that the specific methods used in an employee attrition research study may vary based on the research design, data sources, and specific objectives of the study. However, the above steps provide a general overview of the process involved in conducting employee attrition research.

A. CLASSIFICATION

In this paper, we have used an existing machine learning classification model to classify unseen data, and below we will introduce the classifier used in this research.

The Random Forest algorithm is a popular and robust technique for supervised machine learning used to generate classifications and regressions. It employs multiple decision trees to train the data, as outlined in reference [12]. Each tree casts a vote for a classification label for a given dataset. Afterward, the RF model selects the class with the highest number of votes from the decision trees, as elaborated in reference [13].

B. CORRELATION MATRIX

The correlation matrix gives us insights into how attributes are connected. By analyzing the matrix, we can observe that certain aspects are independent of other features since they are not related to other features. Such attributes are shown in TABLE II While certain attributes are strongly correlated with others, for example, the job level and monthly income as well as years spent in the current role and years with the company. While certain attributes are related to one another, the correlation between them is not very significant. For example, years since the previous promotion is connected with years at the company and years in the current role, whereas age is correlated with job level and monthly income.

TABLE II. List of top 12 features in synthetic balanced data

1. Overtime	7. Stock Option Level
2. Total Working Years	8. Business Travel
3. Job Level	9. Job Role
4. Monthly Income	10. Job Involvement
5. Marital Status	11. Job Satisfaction
6. Years with Current Manager	12. Environment Satisfaction

C. FEATURE SELECTION

A significant variety of traits may be present in real-world datasets. Some of these characteristics are considered noise and may have no beneficial impact on training machine learning algorithms. Incorporating all available characteristics increases model complexity, which has an impact on model performance and training time [14].

There are several techniques for evaluating and ranking all characteristics. In this study, we will employ the confusion matrix to compute Recall, Precision, Specificity, Accuracy, and most crucially. The confusion matrix formula is written as follows [15]:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

IV. CONCLUSION

The exploration of HR practices and waste within the organization was a valuable learning experience. It offered insight into the intricacies of the HR department's role in an organization and identified key reasons behind employee turnover. The study also brought attention to the unique contributions of individual workers. Along with the theoretical framework, hands-on training was provided, which improved HR skills and deepened understanding of HR operations. This comprehensive approach to learning combined both theoretical and practical aspects of design, making it a highly effective educational experience.

There has been a significant amount of knowledge gained in the field of employee retention, including the fundamentals of attrition and retention, calculating attrition, the cost and consequences of high attrition rates, and the reasons for employee turnover. Through the study, researchers were able to better understand the various factors that contribute to employee turnover and identified that opportunities for development and growth are critical for workers in an organization. It was also noted that there is a crisis in human capital management.

In order to tackle the current crisis in human capital management and maintain a highly skilled and high-performing workforce in the long run, significant reforms are necessary. It's important to acknowledge that financial incentives alone are not enough to motivate most employees. While organizations have historically relied on monetary rewards to retain their employees, even highly-paid high-tech workers are known to switch jobs frequently despite enjoying some of the fastest salary growth in any profession motivated by monetary compensation. The study found that job content, salary, relationship with the supervisor, and opportunities for development are all important factors that contribute to employee attrition. The research has also debunked the misconception that salary is the primary reason for employee attrition, revealing that opportunities for development are actually of greater importance to employees. However, salary still remains a crucial factor and needs to be competitive with market rates to retain employees and prevent them from leaving for better-paying competitors. Additionally, the research has highlighted that employees are willing to change jobs even if their salary is high, provided that better development opportunities are available elsewhere.

To reduce attrition rates and retain employees, organizations must address the two critical factors behind attrition, namely opportunities for development and growth, and salary. Employee training and recognition programs need to be given more attention, and salaries must be made more competitive. The significant factors essential for retaining employees, including competitive salaries, opportunities for growth and development, challenging job roles, and effective managerial guidance. As we face a shortage of skilled workers, employers must work smarter and provide opportunities for employees to work smarter.

V. REFERENCES

[1] S. Kaur and R. Vijay, "Job Satisfaction – A Major Factor Behind Attrition or Retention in Retail Industry," Imperial Journal of Interdisciplinary Research, vol. 2, no. 8, 2016.

[2] W. H. Mobley, R. W. Griffeth, H. H. Hand, and B. M. Meglino, "Review and conceptual analysis of the employee turnover process.," Psychol. Bull., vol. 86, no. 3, p. 493, 1979.

[3] W. H. Mobley, "Intermediate linkages in the relationship between job satisfaction and employee turnover.," J. Appl. Psychol., vol. 62, no. 2, p. 237, 1977.

[4] P. S. Budhwar, A. Varma, N. Malhotra, and A. Mukherjee, “Insights into the Indian call centre industry: can internal marketing help tackle high employee turnover?” J. Serv. Mark., vol. 23, no. 5, pp. 351–362, 2009.

[5] G. K. P. V. Vijaya Saradhi, "Employee churn prediction," Expert Systems with Applications, vol. 38, no. 3, pp. 1999-2006, 2011.

[6] D. A. B. A. Alao, "Analysing employee attrition using decision tree algorithms," Computing, Information Systems, Development Informatics and Allied Research Journal, no. 4, 2013.

[7] R. S. Sexton, S. McMurtrey, J. O. Michalopoulos and A. M. Smith, "Employee turnover: a neural network solution," Computers & Operations Research, vol. 32, no. 10, pp. 2635-2651, 2005.

[8] Z. Ö. KISAO˘GLU, Employee Turnover Prediction Using Machine Learning Based Methods (Thesis), MIDDLE EAST TECHNICAL UNIVERSITY, 2014.

[9] M. A. Huselid, “The impact of human resource management practices on turnover, productivity, and corporate financial performance,” Acad. Manag. J., vol. 38, no. 3, pp. 635–672, 1995.

[10] C. E. Shalley, L. L. Gilson, and T. C. Blum, “Matching creativity requirements and the work environment: Effects on satisfaction and intentions to leave,” Acad. Manag. J., vol. 43, no. 2, pp. 215–223, 2000.

[11] McKinley Stacker IV, “SAMPLE DATA: HR Employee Attrition and Performance – IBM Analytics

[12] T. K. Ho, "Random decision forests," in proceedings of the third international conference on Document Analysis and Recognition, 1995.

[13] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.

[14] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of machine learning research, vol. 3, pp. 1157-1182, 2003.

[15] W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm and J. S. Kovach, "Detection of cancer-specific markers amid massive mass spectral data," Proceedings of the National Academy of Sciences, vol. 100, no. 25, pp. 14666-14671, 2003.

[16] Literature Review of Table I, referred from all previous papers which is used to compared all the models’ accuracy.