International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Attrition Prediction and Retention of High Risk Employees Using Machine Learning and Explainable Artificial Intelligence

Sidharth Nigade, Amit Shinde, Nikhil Dethe, Ganesh Chormale, Vaishali Kandekar

Pune Institute of Computer Technology

ABSTRACT

'The Indian Express' reports that the attrition rate in the IT industry has surpassed 25%. The majority of Fortune 500 firms are currently looking for new employees. On the one hand, there are more opportunities, but according to a report by Employee Benefits News, companies are also losing a significant amount of money - about 33% of their annual income. The current era in which the IT industry serves as the foundation for the economies of the world's leading nations. It's interesting to note that 36% of these individuals don't even have their next job secured. The cost of attrition includes both monetary and non-monetary costs, like halted billing, delayed product launches, lack of reliability, etc. Our objective is to determine the causes of attrition and implement preventative actions as a result. The ideal answer was one-on-one interaction with HR, but as we are well aware, it is not scalable. The only option left to us is to use the Data to replicate this ideal state. Use the machine learning model to reliably predict the attrition rate based on all of the available data. Consequently, in this forthcoming modern period where the economics of the world's top-tier countries are centered on the IT industry, recognizing attrition is crucial. While discussing attrition, Bill Gates once said, "If the top 1% of skilled employees leave Microsoft, the company declines from an extraordinary to an average company. As a result, the goal of this paper is to provide customized software to predict the employee retention rate based on supervised and unsupervised machine learning algorithms by analyzing Glassdoor reviews as well as churn dataset to reduce the cost of Employee turnover, to find the potential factors responsible for voluntarily increase in attrition rate, to provide better work life for the employees, etc. Lime which is a library of XAI is used to predict the top 10 important features for the particular employees leaving the organization. Algorithms like Backward selection, logistic regression, decision trees and Random forest regressor were used. Along with this measures were taken to fill the positions or retain the employees using XAI methods.

***Keywords** - Machine Learning, React, Node, Random forest, Logistic Regression, eXplainaible Artificial Intelligence*.

Introduction

Human resource management is an important function in any organization, and one of its key challenges is employee attrition. Employee attrition has a negative impact on the organization, including the loss of valuable human capital, reduced productivity, and increased costs associated with hiring and training new employees. HR analytics is a growing field that uses data analytics and machine learning techniques to help organizations manage their human resources more effectively.

This paper summarizes and synthesizes how organizations can use data analytics to reduce employee turnover and improve their overall performance. One area of HR analytics that has received significant attention is attrition prediction, which involves predicting which employees are likely to leave an organization and taking steps to retain them. In recent years, several studies have investigated the use of machine learning algorithms for attrition prediction.

Literature Review

Here, each study used a different approach to forecast employee attrition utilizing a churn IBM supervised dataset. However, both supervised and unsupervised machine learning data could be used to analyze attrition. Glassdoor reviews can be scraped and added to a dataset.

The 34 features that have been provided have been broken down into their key elements in this essay. The correlation matrix has been produced using the RapidMiner program. Using the Decision Tree approach, attrition is predicted along with the finding of strongly associated components[1]. It learns the connections between the data it receives as input. It forecasts the outcome using the relationships between the features. The conclusion is that employee overtime is a significant factor in employee attrition[1]. Additionally, the group of individuals with low salaries has a higher likelihood of leaving the organization. The younger generation is looking for opportunities more actively[1]. It emphasizes how detrimental attrition is to an organization's environment, regardless of the size of an organization. rather than on managerial experience through valid data provided that can be beneficial in certain ways[1].

An adaptive probabilistic model predicts results with unquestionably improved accuracy when only a tiny subset of the total feature collection is taken into account [2]. Additionally, we may forecast the minimal and maximal for the attrition group by taking into account the group of variables[2]. The crucial attrition rate, attrition of different ages, and cities are clearly shown by the inter-group and intra-group. In addition to identifying the attrition rate, the model predicted in this dataset also selects simple points from the dataset[2].

Both monetary and non-monetary losses were taken into account when calculating the attrition rate. In the UAE, the attrition rate of employees was recognized, and ways to reduce it were developed[3]. This essay mostly focuses on non-financial reasons why employees leave their jobs. This essay seeks to determine the commonality and efficacy of the Big 4 non-financial retention techniques, including training, development, etc[3]. To understand the overall consensus of the informants, statistical analyses such as bar plots, pie charts, and measures of central tendency were taken. The three main causes of employee turnover were determined to be work-related stress, a lack of growth, and insufficient growth possibilities[3].

The key machine learning models used in this study are logistic regression, random forest, gradient descent, and extreme gradient adaptive boosting, among which extreme gradient descent performs well with an accuracy of 98%[4]. The dataset is divided into groups with and without turnover. It has characteristics like leave hours that improve the effectiveness of the GBT and XGB approaches. Precision and recall performance is greatly enhanced[4].

To determine the likelihood of the methods employing ROC and AUC for analysis, Support Vector Machines are combined with a variety of machine learning techniques, including Logistic Regression, KNN, and Naive Bayes[5].

In order to forecast employee retention, machine learning models were used to build classification and performance rating models[6]. Due to the limited data available, the main focus was on data cleansing and preprocessing, which also helped minimize HR costs and the rate of employee churn. Machine learning models are used to analyze the data coupled with feature selection on the dataset[7].

Proposed methodology

Employee Retention and turnover affect the organization in monetary as well as non-monetary ways. Monetary ways include paused billing, product launch delays, rehiring costs, etc. Along with this, non-monetary ways include credibility, cultural stability, trust, etc. Therefore, we identify the causes such as salary, promotion, learning, manager .etc, and mitigate by taking corrective actions. Data collection, visualization and pre-processing is a vital step in order to gauge the employee attrition rate of the program as well as for the particular individual of the organization.In the pre-processing step, we have added the feature of pregnancy which is also one of the prime reasons for the young working ladies to leave the organization. For this, we have added a few new rows along with changing the values of previously assigned variables. In such cases, we can’t do anything but at least it gives a fair idea of vacant positions created which are to be filled.

Due to the fact that data is limited, integration of both supervised and semi-supervised machine learning algorithms was an important factor. For the supervised data, the IBM dataset developed by the U.S. data scientist team of IBM company came in handy. However, only these features weren’t alone satisfactory at all. Therefore, referencing semi-supervised data was necessary. As a result, web scraping of anonymous reviews on the Glassdoor website by the former as well as current employees were also taken into consideration.

After web-scraping, the data was pre-processed before performing the Sentiment Analysis. The ‘Pros’ and ‘Cons’ listed by the employees of IBM in the U.S. region were analyzed. Using the VADER(Valence Aware Dictionary for Sentiment Reasoning) and Roberta(A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network) sentiment analysis was performed. However, Roberta's model yielded better accuracy as compared to the Vader model. Hence the Roberta Pros from the sentimental analysis of Pros and Roberta Neg from the sentimental Analysis of the Cons were added to the scraped dataset.

In order, to integrate both supervised and semi-supervised dataset, the standard column is necessary. Therefore, the JobRole column of IBM which consisted of Sales Executive', 'Research Scientist', 'Laboratory Technician', 'Manufacturing Director', 'Healthcare Representative', 'Manager', 'Sales Representative', 'Research Director', 'Human Resources’ and Role column of the scraped dataset were together assigned standard features as

- **AESP (Assistant Engineering & Scientific Personnel)**: E.g., Lab Technicians, Technical Support/Specialists, etc.
- **Corporate**: E.g., Admin Executive, Procurement, Human Resources, Business Partners, etc.
- **Director**: E.g., Vice President, Senior Vice President, Chief Technology Officer, etc.
- **ESP (Engineering & Scientific Personnel)**: E.g., Data Scientists, Engineers, Member of Technical Staff, Programmers, etc.
- **Manager**: E.g., Team Leads, Project Managers, Service Delivery Managers, etc.
- **Sales**: E.g., Sales Support, Client Representative, Account Representative, etc.

Then the average sentiments of Pros and Cons for the specific role were added to the IBM dataset.

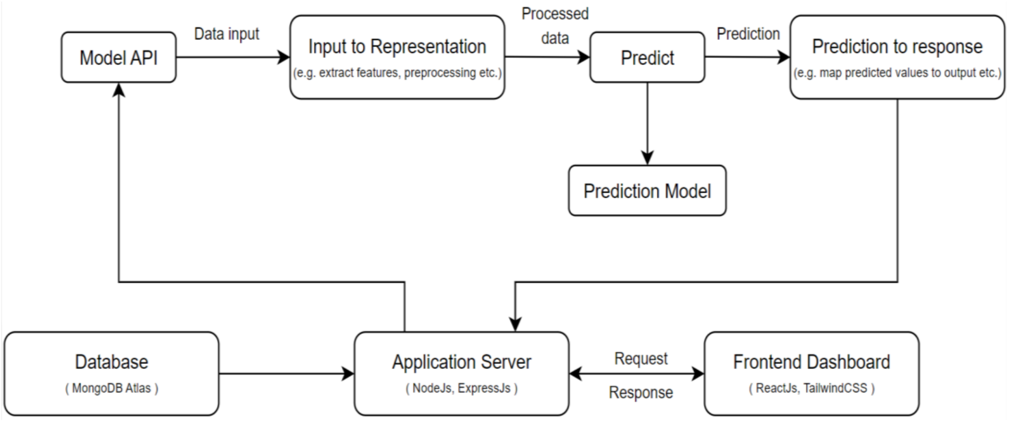
Once, the dataset was created, data was pre-processed and visualized using numpy, pandas, seaborn, matplotlib in-order to get the insights of how data is behaving. After adding the sentimental scores, clustering of the dataset was performed using the KMeans clustering algorithm. An elbow plot was selected in order to find out the number of clusters formed. The total number of clusters formed was then taken as an input feature in the dataset.

Once, the final dataset after pre-processing, one-hot encoding and clustering were ready, models were trained using Logistic Regression with backward selection, decision trees, and Random Forest Regressor. Out of these, Random Forest Regressor suits to be the best algorithm till now.

Once the model was trained, we segregated the dataset in order to find out the 'High Risk Employees' based on various factors like JobRole, Experience, Overtime status, Years of working with current manager .etc. High risk employees are termed as employees which are valuable to the organization. If they leave, this might create a vacuum. Our model tends to predict the high risk employee's attrition rate with the top five factors responsible for attrition.

Explainable Artificial Intelligence commonly known as XAI was used to predict the top five factors for each of the high-risk employees contributing to attrition. After considering the top five factors, HR can decide if there is a specific change in those factors or a comparative increase in market salary then what's the attrition status of an individual. Along with this HR can also monitor the ideal candidate required to fill the vacant position of the high risk employee.

System architecture



The system architecture consists of a frontend, backend and machine learning model. The pre-trained machine learning model is built for predicting employee attrition with factors contributing to attrition which will be deployed and the API will be called in an application server. The application server of the system consists of NodeJS and ExpressJS. The system is designed using ReactJS and Tailwind CSS for the front end. MongoDB Atlas is used as a database for storing the employees & hr data.

Results:

The IBM dataset's characteristics combined with the positive and negative sentiments from Glassdoor reviews has proven to be more effective than earlier models based on supervised machine learning algorithms.

The accuracy of the final model after backward selection training is 84.4508%, and the k_score is 89.212%. The top 10 features for the total dataset that contribute to the rise in attrition rate are listed here. Similar to backward selection, the model's accuracy when employing logistic regression is 87.301%, which is rather high. However, 35 employees are predicted by logistic regression to be on the verge of leaving the company. When compared, the random forest, decision trees, and logistic regression algorithms all yield higher precision rates but lower accuracy. However, they both correctly anticipated 76 and 89 employees leaving the company, with respective accuracy rates of 77.551% and 84.526%.

Conclusion and future scope:

As a result of this project, we can predict the attrition rate of an organization. Further more we can determine factors that contribute to attrition. Monetary losses of organizations will be reduced significantly by doing the right changes in the working atmosphere, salary, work routine or any other causes predicted by the machine learning model. By using explainable AI we can figure out five attrition factors for an individual as well.

References

[1] Towards Understanding Employee Attrition using a Decision Tree Approach Saadat M Alhashmi College of Computing and Informatics University of Sharjah Sharjah, UAE 19653360, 2020

[2] Abhiroop Nandi Ray; Judhajit Sanyal "Machine Learning Based Attrition Prediction" 2019 (GCAT) IEEE

[3] Vimoli Mehta, Shrey Modi, "Employee Attrition System Using Tree Based Ensemble Method", 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4), pp.1-4, 2021.

[4] A Review of Employee Motivation Theories and their Implications for Employee Retention within Organizations Sunil Ramlall, Ph.D., University of St. Thomas, Minneapolis, MN

[5] Nikita Tresa Cyriac, Kamaladevi Baskaran A Study on the Effectiveness of Non-Monetary Retention Strategies in UAE 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)

[6]R. S. Shankar, J. Rajanikanth, V. V. Sivaramaraju and K. VSSR Murthy, "PREDICTION OF EMPLOYEE ATTRITION USING DATA MINING," 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA), Pondicherry, 2018, pp. 1-8.

[7] S. Yadav, A. Jain and D. Singh, "Early Prediction of Employee Attrition using Data Mining Techniques," 2018 IEEE 8th International Advance Computing Conference (IACC), Greater Noida, India, 2018, pp. 349-354

[8]D. Wilson, "Predicting Employee Churn with Data Mining - CDO Advisors,"2017. Predicting-employee-churn-datamining [Accessed: 06-Jan-2019].

[9] Shobhit Aggarwal, Mugdha Sharma "Employee Attrition Prediction Using Machine Learning Comparative Study" IMAES pp 453–466 2021.

[10] W. H. Mobley, "Intermediate linkages in the relationship between job satisfaction and employee turnover.," J. Appl. Psychol., vol. 62, no. 2, p. 237, 2016.

[11] Sarahi Aguilar-Gonzalez, Leon Palafox "Prediction of Student Attrition Using Machine Learning" MICAI 2019.

[12] A. Frye, C. Boomhower, M. Smith, L. Vitovsky, and S. Fabricant, "Employee Attrition: What Makes an Employee Quit?," SMU Data Sci. Rev., vol. 1, no. 1, p. 9, 2018.