Employee Attrition Using Machine Learning With Reinforcement Algorithm

Vanshika Pritmani
Department of Computer Engineering & Technology Dr. Vishwanath D. Karad
MIT World Peace University, Kothrud, Pune Maharashtra, India

Aditya Patil
Department of Computer Engineering & Technology Dr. Vishwanath D. Karad
MIT World Peace University, Kothrud, Pune Maharashtra, India

Shuhaab Safi
Department of Computer Engineering & Technology Dr. Vishwanath D. Karad
MIT World Peace University, Kothrud, Pune Maharashtra, India

Varun Agarwal
Department of Computer Engineering & Technology Dr. Vishwanath D. Karad
MIT World Peace University, Kothrud, Pune Maharashtra, India

Tanmay Jain
Department of Computer Engineering & Technology Dr. Vishwanath D. Karad
MIT World Peace University, Kothrud, Pune Maharashtra, India

Prof. Pramod Mali
Department of Computer Engineering & Technology Dr. Vishwanath D. Karad
MIT World Peace University, Kothrud, Pune Maharashtra, India

***Abstract*—** Employee attrition, the phenomenon of employees leaving a company, poses significant challenges for organizations, including loss of talent, increased recruitment costs, and decreased productivity. This paper gives an understanding of employee attrition. The main purpose of this paper is to give maximum accuracy as possible. Which is being done by using Random Forest algorithm. The algorithms such as KNN, decision tree and SVM are used. The main purpose is to show how Random Forest which gives better accuracy than ANN. The data set used is of IBM employee attrition. The accuracy of Random Forest algorithm is 95.74 %. Which is the highest of any other algorithms.

Keywords— Employee attrition, Decision tree, machine learning, K-Nearest Neighbors(KNN), Support vector machines(SVM), Random Forest.

1. Introduction

In today's organizations understanding the employee attrition has become a top-most priority for the businesses across industries which includes higher turnover rates not only disrupt workflow and reduce productivity, but they also incur significant costs like recruitment, training. To effectively navigate this challenge, organizations are increasingly turning to data-driven approaches that use analytics to extract actionable insights from massive repositories of employee data.Our investigation includes 2 parts: First, we use exploratory data analysis (EDA) to gain a thorough understanding of the dataset's structure. We hope to discern salient trends, identify potential predictors of attrition, and uncover different insights into the interplay of various factors influencing employee turnover using visualizations, statistical summaries, and correlation analyses.

We then secondly use predictive modelling techniques to create robust algorithms capable of accurately forecasting attrition risk. We use machine learning algorithms such as logistic regression, decision trees, and ensemble methods to build models that accurately predict employee attrition. By clarifying the complex relationship between various predictors and attrition outcomes, our research aims for providing organizations with actionable intelligence to address attrition challenges. Finally, our findings are intending to inform strategic decision-making, allowing businesses to implement targeted retention strategies, foster employee engagement, and create a work environment conducive to long-term organizational success.

In crux, this paper demonstrates data analytics' transformative potential for redefining how organizations understand, anticipate, and mitigate employee attrition through rigorous analysis and empirical validation.

2. Literature Survey/ Related Work:

Employee attrition is a major concern for firms that prioritize human capital in their sectors.Classification is crucial in data mining. Several studies have been undertaken on categorization algorithms. Efficient and scalable data mining methods are necessary for extracting information from massive datasets. Data mining methods rely on classification accuracy and training time to evaluate performance.

These factors aid in identifying the best algorithms for classifying tasks in data mining.

A study employing IBM's human resource analytic attrition dataset discovered an imbalance in the results. Correlation plots and histograms were utilized to depict the relationship between the model's continuous variables. Literature evaluations often aim to optimize machine accuracy and speed. The following literature surveys were reviewed:

1. Norsuhada Mansor, Nor Samsiah Sani (Main Reference): The writers aimed to identify potential job or company leavers. Classification Techniques: Decision Tree, SVM, and ANN. The author recommends ANN.

2. Srivastava, D. K., & Nair, P. - OBJ 2: Identifying predictors of employee turnover. Classification technique: artificial neural network. Recommended: ANN.

3. Research Gate - Demonstrates a technique for predicting employee turnover. Classification techniques include SVM, Random Forest, J48, Logit Boost, MLP, KNN, LDA, Naive Bayes, Bagging, AdaBoost, and Logistic Regression. Recommended: SVM.

4. Ozdemir, Coskun, Gezer and Gungor- Objective was to guess if an employee will quit soon and find out what makes them want to leave. Classification Technique- LR & XG Boost. Recommended- XG Boost.

3. METHODOLOGY

Attrition- Attrition is a process wherein the body of workers dwindles at a enterprise, following a duration wherein some of humans retire or renounce, and are not changed.

A. Data Pre-processing

1. Information Description: data pre-processing is one of the critical step in any statistics analysis. This entails remodeling uncooked data right into a layout this is appropriate for similarly analysis or modelling.

2. Outlier Detection: Outlier detection is a vital issue of statistics preprocessing, specially in facts evaluation and s ystem gaining knowledge of. Outliers are facts points that appreciably deviate from the relaxation of the records, doubtlessly indicating anomalies, errors, or uncommon activities. managing outliers is crucial to make certain the accuracy and reliability of the evaluation or version.

3. Visualization: Facts visualization is robust device used to represent information graphically. It includes the creation of visible representations of data visualization, starting from simple charts and graphs to more superior interactive visualizations to be had for data visualization, basically includes easy charts and graphs to more superior interactive visualizations.

4. Statistics cleansing and reduction: The attributes which are irrelevant and do not contribute to the goal of this examine ought to be discarded. Features consisting of 'EmployeeCount', 'StandardHours', and 'Over18' can be eliminated because they have got a cardinality/difference of '1', indicating they have the identical values in the course of the records. moreover, 'EmployeeNumber' is deemed beside the point for the modeling and prediction procedure and may be excluded from the dataset. No spelling inconsistencies had been detected, as such inconsistencies may pose problems in later merges or variations.

5. Normalization and Discretization: during the pre-processing stage, information transformation involves making use of characteristic scaling or normalization. Normalization standardizes the variety of impartial variables or functions of the information, making sure that they fall within various 0 to 1. This method eliminates dependency on the choice of size devices forattributes. moreover, records cleaning and reduction have been executed, along with the discretization procedure and changing the attribute type from numerical to nominal. four attributes had been removed based at the findings, leaving 30 attributes. No outliers had been detected after regenerating the interquartile clear out.

6. Feature selection: characteristic choice is the subsequent step in pre-processing for device gaining knowledge of, aimed toward selecting relevant features inside the facts and doing away with inappropriate and redundant information to reduce dataset dimensionality. This process allows improve accuracy, reduce overfitting, lower sc hooling time, and discover the maximum crucial and predictive fields for evaluation.

7. Test records: on this test, the statistics is divided into two units - education and checking out facts - the usage of the resample filter out feature, with a split ratio of 80:20.

8. Addressing Imbalanced records: The facts exceptional record highlighted an imbalance inside the magnificence distribution, with 237 times predicted as 'yes' and 1233 times expected as 'No.' magnificence imbalance is a not unusual challenge in classification issues, wherein one elegance is drastically more prevalent than the others. This imbalance frequently ends in poorer predictions for the minority elegance, that's generally the focal point of hobby in type obligations. coping with imbalanced facts calls for techniques that can take care of varying misclassification costs.

Getting to know type Algorithms:

This segment explains the 3 algorithms which can be used in this examine:

1. LOGISTIC REGRESSION:

Logistic Regression is a statistical approach used for binary classification duties, making it a suitable desire for predicting worker attrition, which commonly includes figuring out whether an worker will depart (1) or stay (zero) primarily based on diverse capabilities.

The output which can be generated from logistic regression is between 0 to 1. It is then transformed the usage of a threshold (commonly 0.5) to make binary predictions.

2. SECTOR VECTOR MACHINE: Sector Vector Machines (SVM) comes underneath supervised set of rules in system mastering, specifically inside the context of worker attrition datasets. SVM serves as each a classifier, categorizing statistics into one of a kind classes, and as a regression characteristic, estimating numerical values based totally on a linear combination of capabilities, relevant to both linear and non-linear records.

The SVM version is skilled at the dataset to generalize the enter records primarily based on their functions and make predictions. The SVM machine gaining knowledge of technique effects in a version that predicts the goal values of the test statistics. The fundamental idea in the back of SVM is to separate training with a maximum margin created by using hyperplanes.

Tuning parameters in SVM consist of the kernel, regularization parameter (C parameter), and gamma. Polynomial and exponential kernels rent kernel hints to calculate separation traces in a better size, facilitating the classification of non-linearly separable data.

3. K-NEAREST Neighbor (KNN)- KNN is some other popular algorithm for classification obligations, along with predicting worker attrition. It is a easy but powerful method that classifies data factors based totally on the general public magnificence amongst their ok nearest friends in function area.

4. REINFORCEMENT-

Reinforcement Learning (RL) is a kind of system studying paradigm in which an agent learns to make choices by way of interacting with an surroundings. in the context of predicting worker attrition, RL might not be the maximum common approach in comparison to supervised getting to know techniques like logistic regression or okay-nearest pals. But, RL may want to nevertheless be applied in positive scenarios or aspects of employee attrition prediction.

RL can permit organizations to tailor interventions for character employees based on their precise traits and behavior patterns. by way of getting to know from beyond interactions and outcomes, the enterprise can

make more powerful selections approximately a way to engage and aid employees to prevent attrition.

B. Task Results

By taking look at 4 algorithms, LR, SVM, KNN, and RF. The common measures taken into consideration are rate of accuracy, error rate, RMSE score. The formula for predicting accuracy is- correct prediction percentage divided by total number of predictions. The main thing predicted by RMSE is the measure of the fitness of dataset which is absolute. So as RMSE decreases it is better or healthier for the research. On initial level the first model of project was performed on training dataset. The usage is every parameter that was default was classified. The assessment which is performed of classifier is in Table II. In the desk the initial procedure of models has been disguised:

1. highest accuracy end result at 95.74% is proven by using RF at the same time as the lowest at 59.91 % is confirmed by means of SVM.
2. The bottom value of RMSE which is 0.2061 is shown by RF

TABLE II. COMPARATIVE RESULT BETWEEN CLASSIFIERS

Performance Measure	LR	SVM	KNN(k_n=4)	RF
Accuracy	61.74%	59.91%	77.93%	95.74%
Error Rate	38.25%	40.08%	26.31%	4.25%
RMSE	0.6185	0.6330	0.4697	0.2061

4. CONCLUSION

We would like to conclude by saying our utilization of reinforcement learning algorithms has proven to be highly effective in achieving superior accuracy compared to traditional artificial neural network approaches. As we continue to delve deeper into the realm of reinforcement learning, there's ample potential for further advancements, promising even greater levels of precision and efficiency in future endeavours. Reinforcement learning algorithms outperform traditional artificial neural network (ANN) approaches in several aspects. They excel in dynamic environments, adapt their strategies based on feedback, and balance exploration and exploitation efficiently.

5. ACKNOWLEDGMENT

We extend our heartfelt appreciation to Mr. Pramod Mali for his unwavering support and guidance during the development of our project on the employee attrition system. His mentorship and encouragement were instrumental in shaping our research.
Additionally, we would like to thank Dr. Vishwanath D. Karad and MIT World Peace University for providing us with a platform to explore and present our findings. Their support enabled us to undertake this endeavor and contribute to the field. We are deeply grateful for their invaluable assistance.

6. REFERENCES

1. https://thesai.org/Downloads/Volume12No11/Paper_49-machine_Learning_for_Predicting_Employee_Attrition.pdf
2. https://www.ijresm.com/Vol.2_2019/Vol2_Iss8_August19/IJRESM_V2_I8_48.pdf
3. H. Jiawei and M. Kamber, Data Mining: Concepts and Techniques, Burlington, MA: Morgan Kaufmann, 2001.
4. Z. A. Othman, A. A. Bakar, N. S. Sani, and J. Sallim, "Household Overspending Model Amongst B40, M40 and T20 using Classification Algorithm," International Journal of Advanced Computer Science and Applications, vo1. 11(7), pp. 392-399, 2019.
5. S. O. Akinola and O. J. Oyabugbe, "Accuracies and training times of data mining classification algorithms: An empirical comparative study," J. Softw. Eng. Appl., vol. 8, pp. 470–477, September 2015.
6. https://www.researchgate.net/publication/361522993_Predicting_Employee_Attrition_Using_Machine_Learning_Approaches
7. https://www.researchgate.net/publication/370471266_A_STUDY_ON_EMPLOYEE_ATTRITION_AND_RETENTION_WITH_REFERENCE_TO_EVRON_IMPEX
8. https://ijariie.com/AdminUploadPdf/Employee_Attrition_ijariie11888.pdf
9. https://hmct.dypvp.edu.in/Documents/research-papers-publication/Resarch-publications/60.pdf
10. https://ieeexplore.ieee.org/document/8605976\
11. https://www.mdpi.com/2073-431X/9/4/86
12. https://www.researchgate.net/publication/356809874_Machine_Learning_for_Predicting_Employee_Attrition
13. https://hmct.dypvp.edu.in/Documents/research-papers-publication/Resarch-publications/60.pdf
14. https://www.hindawi.com/journals/cin/2022/7728668/