International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Impact Analysis on Employee Attrition using Machine Learning Techniques

Roopak Krishna[1], Neeta Singh[2]

[1]Student, Gautam Buddha University, Greater Noida, India

[2]Assistant Professor, Gautam Buddha University, Greater Noida, India

Abstract: ***Employees are the most valuable resources for any organization but Employee attrition is the biggest problem for any organization, it is a rate at which more employees are leaving the organization than the rate at which they are getting hired. It is the primary and most challenging task for any organization regardless of their size. The cost associated with professional training, the developed loyalty over the years and the sensitivity of some organizational positions, all make it very essential to identify who might leave the organization. According to a report, this is the biggest problem in the call centers. It deteriorates the customer experiences. This problem comes under the Human Resource and for growth in the organization it is believed that the employee retention rate is the deciding factor. Attrition brings down HR management morale way down because it is their task to understand the employee needs and how to keep them by accounting all the factors. If an employee leave, team workflow pipeline get disrupted because the pressure comes down on the shoulder of other team members. Many reasons can lead to employee attrition. In this paper, several machine learning models are compared to automatically and accurately predict employee attrition. IBM attrition dataset is used in this work to train and evaluate machine learning models; namely Bagging Classifier, Random Forest, Logistic Regression, and K Nearest Neighbor models. The ultimate goal is to accurately detect attrition to help any company to improve different retention strategies on crucial employees and boost those employee satisfactions.***

Keywords: Data Analysis; Predictive Modeling; Machine learning; Employee Attrition

1. Introduction

In today's scenario, Data has answers to every question. In fact, Data is the key factor in the growth of the company. It enables us to perform analysis, exploration and helps to understand the insights to improve the operations of the company. We can modify the existing policies of the company to function more efficiently by understanding the data. It is believed that if the company wants to survive in the long run, it must possess analytical skills from the HR management (Sujeet et al.2016). Human Resource predictive analytics (HRPA) helps organizations to optimize the company performance and also improve the employee engagement.

According to the U. S. Bureau of Labor Statistics, 4 million Americans quit their jobs in July 2021. This problem has mostly seen in young generation. We can have an idea if we are talking about India in terms of population and also keeping in the mind that our workforce is mostly on younger side then there is a greater possibility of employee attrition rate is higher than anyone.

Employees are the important asset for any organization and keeping them in the organization is the biggest task for the HR of the company. Relation between the work and home life of an employee has an significant effect on the work productivity (Chu et al.2022). According to a live mint article, attrition rate of India in 2020 was 6% and in 2022, it was 20.3%. Due to the pandemic factor, it would have made the negative impact on the employees' psychological well being and this would have increased the attrition rate. It was seen that procrastination can also deteriorates the performance of the project in the company (A. Khan et al.2022). External interruptions, adaption and emotional are the categories that influence the employee productivity in the remote work environment (Bezerra et al.2020).

There are number of papers published related to employee attrition and (Guerranti et al.2023) shows that Machine learning will work as a support system for any organization. Ma - chine learning for predicting employee attrition could play an important part in decision making of the company. Some researchers have applied Boosting algorithms (Kakulapati et al.2023; Ganthi et al.2022), Artificial Neural Network with SMOTE (Soner et al.2022), and also Deep Learning (Arqawi et al.2022).

There are various parameters for employee attriton rate. It can also be voluntary and involuntary. It could be resignation, transfer, leaves, training gap, private life disturbed, salary problem, less or no hike, distance from home is more, job dissatisfaction, no promotions so there can be number of factors involved depending on the employee.

Data taken in this work is from Kaggle (IBM HR dataset) which have 1470 rows and 35 features. The main focus in this proposed paper is to optimize the algorithms through different methodology in order to build a better HR support system. Firstly, we have addressed the highly correlated features that are associated with the target variable. By analyzing the correlated features, we can understand the features that are responsible for the attrition of the employee.

Some companies also use attrition as their strategy to move forward. Company freezes the hiring so that they increase the attrition rate in order to maintain the company balance in terms of financial. There are planned layoffs nowadays that we notice through social media, news, any article that big companies are laying of some amount of workers to keep them afloat.

If the company is getting restructured or reorganized or under new management, first they layoff the number of employees so that they organize the company through their perspective. Organization reputation got damaged if the rate is getting higher and higher. Pipeline of workflow get disturbed and employee burnout rate will be higher. Employee burnout has the negative effect on the organization. There are other dependent factors of burnout like the employee's might get stressed due to the heavy work load, their work life balance cycle might get disturbed, get depressed due to low energy and heavy work load. Overall their mental state directly or indirectly get effected due to this factor.

Through this, we can have an overall idea of various factors that affects the employee and what factors make them to leave the organization. So, we are going to try to understand the problem through the predictive analytics and build a predictive model. Results are expressed in numerical terms of Accuracy, Precision, Recall rate and com - pared with the previous studies. Logistic Regression achieved the best accuracy among the four algorithms.

The rest of the work is organized as follows. In Section 2, methodology of our work is discussed in brief, starting with data description, data pre - processing, data analysis, feature selection/importance, building model with multiple algorithms. In Section 3, we reported the results of our study and compared them with previous studies and also the comparison between the evaluation metrics of before hyper parameter tuning and after hyper parameter tuning. Finally, in Section 4 future scope of this work is discussed and then references.

2. Methodology

Data Description

The dataset has been taken from the Kaggle that has been provided by IBM for study purposes. The dataset consist of 35 features and 1470 rows.

Table 1: Complete features

Feature Name	Data Type
Age	Int64
Attrition	Object
Business Travel	Object
Daily Rate	Int64
Department	Object
Distance from Home	Int64
Education	Int64
Education Field	Object
Employee Count	Int64
Employee Number	Int64
Environment Satisfaction	Int64
Gender	Object
Hourly Rate	Int64
Job Involvement	Int64
Job Level	Int64
Job Role	Object
Job Satisfaction	Int64
Marital Status	Object
Monthly Income	Int64
Monthly Rate	Int64
Num Companies Worked	Int64
Over18	Object
Over Time	Object
Percent Salary Hike	Int64
Performance Rating	Int64
Relationship Satisfaction	Int64
Standard Hours	Int64
Stock Option Level	Int64
Total Working Years	Int64
Training Times Last Year	Int64
Work Life Balance	Int64
Years at Company	Int64
Years in Current Role	Int64
Years since Last Promotion	Int64
Years With Current Manage	Int64

The dataset contains target feature, identified by the variable Attrition: "No" represents an employee that did not leave the company and "Yes" represents an employee that left the company. Data has been imported in Jupyter environment through the local machine and all the work has been done by using Python.

Data Pre - processing

In the second step of machine learning process the data is pre - processed to check the quality which affects the outcome of the research in terms of accuracy, precision etc. We have checked missing values, outliers, redundancy and prepared a proper dataset. Below we have discussed all the steps in detail:

Missing Values Identification

We have imported missing no package in the environment for the visualization of each feature for the same.

Data Duplication Identification

Redundancy of data has been checked and it is observed that there is no redundant data.

Outliers Identification

Outliers are the extreme points from the mean position of the dataset that can be on higher and lower side. Though, they don't give errors in general but could have great effect on the outcomes as they are sometimes called as conceptual wrong data. So we can look for them manually too and also through visualizations.

Label Encoding/Mapping

As there are categorical features in the form of Gender [Male, Female], BusinessTravel [Travel_Rarely, Travel_Frequently], Education Field [Medical, Life Sciences, Other] etc. Thus all the feature are encoded and assigned values accordingly. Labels of features are: Over Time, Gender, Performance Rating, Attrition, Business Travel, Department, Education Field, Job Role, Marital Status, Over18 are encoded properly.

Data Analysis

In this section, Data analysis and data mining is done which is a important backend technique to understand the behaviour of data that we have. Model building is not only to apply algorithms but to understand the data through graphs, stats etc. A pattern of a particular feature can be observed through analysis and use them accordingly for the decision making and also to convert insights of data into

some meaningful information. There is no limitation for the analysis.

Feature Selection/Importance

If there are tons of features, the task is to reduce the unnecessary features that might impact negatively during the model building part. To know which feature is important or which to discard reflects in the model building as in good terms.

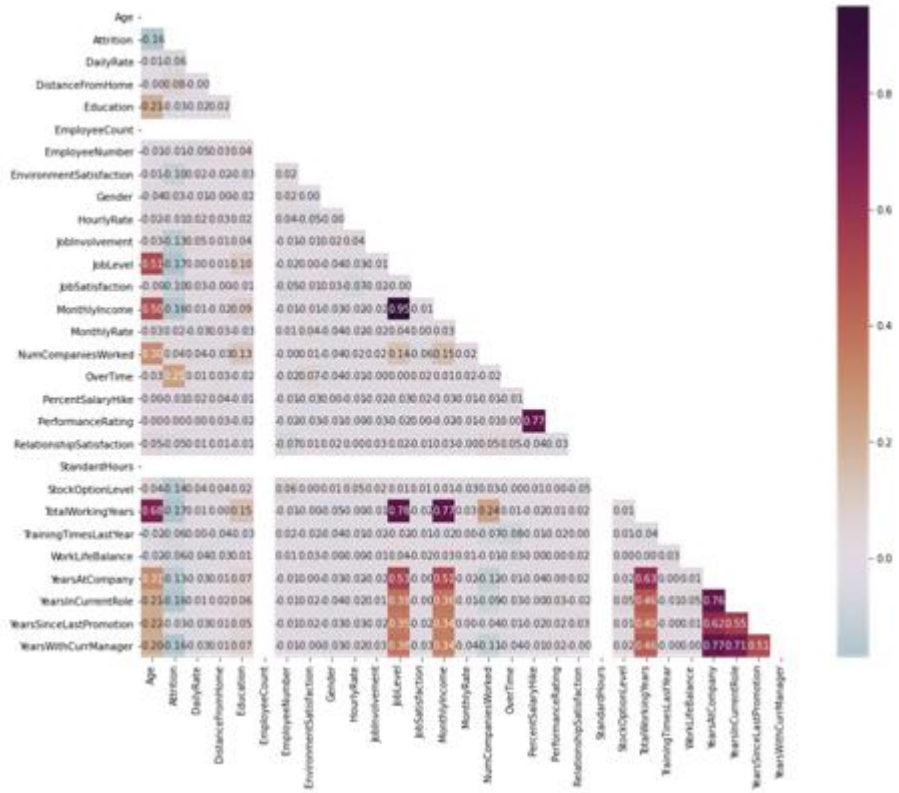

Figure 8: Correlation matrix

For also hypothesis testing purposes and statistical modelling, one has to understand the importance of feature selection. There could be an issue of multicollinearity. If there are two or more features are highly correlated either in positive or negative terms that means, only one correlated feature has to be kept because keeping high correlated features for the model does not give good outcomes. Aim is to keep only significant features for the model building.

Through Figure8, it is seen that MonthlyIncome and Job Level feature are highly correlated with the value of 0.95. So there is a high chance of Multi Collinearity issue. So, Job Level feature has been dropped from the dataset. Correlation matrix is also a best way to understand the relationship between various features.

There are few algorithm that python offers for understanding the feature importance. In this proposed work, Extra Trees Classifier algorithm has been used that is based on Random Forest algorithm and compute the Gini index for each feature by taking the account of target variable.

Model Building

Four different machine learning algorithms have been used to build the model: Logistic Regression, K Nearest Neighbor, Bagging Classifier, Random Forest.

Furthermore, to optimize the model, data normalization technique MinMaxScaler () has been used. It maintains the range of the numerical features. It brings all the features in between the [0, 1] range where 0 is for minimum and 1 is for maximum. Best part about MinMaxScaler () is that it scales the values without any change in the original distribution of the dataset. But it is only for numerical features.

Algorithms like Random Forest does not work well with the features that have labels. Beside MinMaxScaler () for numerical features, categorical values have been converted into indicator values (0, 1) /dummy variables.

Dataset is divided with 70% of data into training purpose and rest 30% of data into testing purpose. During the segregation of dataset shuffling of data is kept TRUE, so that there is no chance of getting same data.

3. Results

Many researchers have worked on the area where machine learning models can support human resource management. Below we have discussed comparison of some of the papers (Francesca et al.2020; Qutub et al.2021; Raza et al.2022; Guerranti et al.2023) using various machine learning algorithms.

Table 6: Comparison of accuracy with previous work

Algorithms/Years	2020	2021	2022	2023
Logistic Regression	87.5	88.43	74	87.96
K Nearest Neighbour	85.2			
Random Forest	86.1	85.03		82.9

In Table 7, the comparisons of the given algorithms have been shown. We have identified the best parameters in for the each algorithm through the Grid Search Cross Validation method and again rebuild the model with those specific parameter to optimize the previous machine learning models and to give more accurate results.

Table 7: Comparison of various metrics

Algorithms	Accuracy	Precision	Recall
LR	88.44	90.81	96.31
KNN	85.71	87.64	97.1
Bag. class.	86.62	87.81	95.78
RF	86.17	87.29	99.47

4. Future Work

This paper compare various machine learning models to identify the key indicators which can find the accuracy that an employee will leave the company or not. Four base model were trained. In future, various data standardization technique and ensemble machine learning algorithms can also be employed.

References

[1] Sathe, C. A. and Panse, C. (2022), "Analyzing the impact of agile mindset adoption on software development teams productivity during COVID - 19", Journal of Advances in Management Research, Vol. ahead - of - print No. ahead - of - print. https: //doi. org/10.1108/JAMR - 05 - 2022 - 0088

[2] Balazs Aczel, et al. (2021), "Researchers working from home: Benefits and challenges", https: //doi. org/10.1371/journal. pone.0249127

[3] Khan, H. Zada, M. Tahir (2022), "Achieving Project Performance through Work from Home during the COVID - 19 Pandemic: A Mediating Role of Procrastination", Jinnah Business Review, Vol.10, No.2, pp.25 - 34

[4] Yan, Binxin and Stuart, Logan and Tu, Andy and Zhang, Qingquan, "Analysis of the Effect of COVID - 19 on the Stock Market and Investing Strategies" (March 28, 2020). http: //dx. doi. org/10.2139/ssrn.3563380

[5] Bezerra, C. I., de Souza Filho, J. C., Coutinho, E. F., et al. (2020), "How human and organizational factors influence software teams' productivity in COVID - 19 pandemic: a Brazilian survey", Proceedings of the 34th Brazilian Symposium on Software Engineering, pp.606 - 615

[6] Bloom, N., Bunn, P., Mizen, P., Smietanka, P. and Thwaites, G. (2020), "The impact of COVID - 19 on productivity", National Bureau of Economic Research, Vol.1 No.1, w28233, doi: 10.3386/w28233

[7] Jakub Prochazka, Tabea Scheel, Petr Pirozek, Tomas Kratochvil, Cristina Civilotti, Martina Bollo, Daniela Acquadro Maran (2020), "Data on work - related consequences of COVID - 19 pandemic for employees across Europe", https: //doi. org/10.1016/j. dib.2020.106174

[8] Alexander Brem, Eric Viardot, and Petra A. Nylund (2020), "Implications of the coronavirus (COVID - 19) outbreak for innovation: Which technologies will improve our lives?", doi: 10.1016/j. techfore.2020.120451

[9] Lingfeng BAO, Tao LI, Xin XIA, et al. (2022), "How does working from home affect developer productivity? — A case study of Baidu during the COVID - 19 pandemic", https: //doi. org/10.1007/s11432 - 020 - 3278 - 4

[10] E. Jeffrey Hill, et al. (2003), " Does it matter where you work? A comparison of how three work venues (traditional office, virtual office, and home office) influence aspects of work and personal/family life", doi: 10.1016/S0001 - 8791 (03) 00042 - 3

[11] Clara De Vincenzi, et al. (2022), " Consequences of COVID - 19 on Employees in remoteWorking: Challenges, Risks and Opportunities An Evidence - Based Literature Review", https: //doi. org/10.3390/ijerph191811672

[12] Francesca Fallucchi, et al. (2020), " Predicting Employee Attrition Using Machine Learning Techniques", doi: 10.3390/computers9040086

[13] Sujeet N. Mishra, et al. (2016), " Human Resource Predictive Analytics (HRPA) for HR Management in Organizations", International Journal of Scientific & Technology Research, 5 (5), 33 - 35

[14] Ayittey, F. K., Ayittey, M. K., Chiwero, N. B., Kamasah, J. S., & Dzuvor, C. (2020). Economic impacts of Wuhan 2019-nCoV on China and the world. Journal of medical virology, 92 (5), 473.

[15] Tušl, M., Brauchli, R., Kerksieck, P., & Bauer, G. F. (2021). Impact of the COVID - 19 crisis on work and private life, mental well - being and self - rated health in German and Swiss employees: A cross - sectional online survey. BMC Public Health, 21 (1), 1 - 15.

[16] Matli, W. (2020). The changing work landscape as a result of the Covid - 19 pandemic: insights from remote workers life situations in South Africa. International Journal of Sociology and Social Policy.

[17] Platts, K., Breckon, J., & Marshall, E. (2022). Enforced home - working under lockdown and its impact on employee wellbeing: a cross - sectional study. BMC Public Health, 22 (1), 1 - 13.

[18] Chu, A. M., Chan, T. W., & So, M. K. (2022). Learning from work - from - home issues during the COVID - 19 pandemic: Balance speaks louder than words. Plos one, 17 (1), e0261969.

[19] IBM HR analytics employee attrition dataset: https: //www.kaggle. com/datasets/pavansubhasht/ibm - hr - analytics - attrition - dataset

[20] Qutub, A., Al - Mehmadi, A., Al - Hssan, M., Aljohani, R., & Alghamdi, H. S. (2021). Prediction of employee attrition using machine learning and ensemble methods. Int. J. Mach. Learn. Comput, 11 (2), 110 - 114.

[21] Kakulapati, V., & Subhani, S. (2023). Predictive Analytics of Employee Attrition using K - Fold Methodologies. DOI: 10.5815/ijmsc.2023.01.03

[22] Guerranti, F., & Dimitri, G. M. (2023). A Comparison

of Machine Learning Approaches for Predicting Employee Attrition. Applied Sciences, 13 (1), 267.

[23] Soner, S., Hussain, A. A., Khatri, R., Kushwaha, S. K., Mathariya, S., & Bhayal, S. (2022). Predictive Deep Learning approach of employee attrition for imbalance datasets using SVMSMOTE algorithm with Bias Initializer.

[24] Arqawi, S. M., RUMMAN, M. A. A., ZITAWI, E. A., RABAYA, A. H., SADAQA, A. S., ABUNASSER, B.

[25] S., & ABU - NASER, S. S. (2022). Predicting Employee Attrition and Performance Using Deep Learning. Journal of Theoretical and Applied Information Technology, 100 (21).

[26] Ganthi, L. S., Nallapaneni, Y., Perumalsamy, D., & Mahalingam, K. (2022). Employee Attrition Prediction Using Machine Learning Algorithms. In Proceedings of International Conference on Data Science and Applications: ICDSA 2021, Volume 1 (pp.577 - 596). Springer Singapore.