European Economic Letters
ISSN 2323-5233
Vol 13, Issue 1s (2023)
http://eelet.org.uk

Predictive Model for Employee Attrition Risk Assessment

[1]Deepesh Mamtani, [2]Dr. Bharti Malukani
[1]Asst. Professor, [2]Asst. Professor
deepesh_mamtani@pimrindore.ac.in
Prestige Institute of Management & Research, Indore (M.P.), India
Prestige Institute of Management & Research, Indore (M.P.), India
bharti_malukani@pimrindore.ac.in

Abstract: In the contemporary business landscape, organizations face a significant challenge in the form of employee attrition. This phenomenon entails a reduction in the workforce, resulting from either voluntary resignations or management-driven terminations. The departure of employees not only leads to a loss of human resources but also diminishes the organization's expertise and revenue generation potential associated with those individuals. As a result, it becomes imperative for companies to proactively anticipate and mitigate the risks associated with attrition by implementing preventive measures and making well-informed decisions.

Within the scope of this research paper, we introduce a sophisticated model that leverages advanced Machine Learning techniques to accurately predict employee attrition. The model's development involved meticulous design and comprehensive evaluation, utilizing diverse company profiles. The findings demonstrate the model's efficacy in maximizing employee retention, empowering management to address attrition-related challenges proactively and make strategic decisions that ultimately benefit the organization.

Keywords: Classification, Supervised Learning, Feature Engineering, Data Preprocessing, Human Resources, Risk Management, Talent Management, Predictive Analytics, Employee Retention, Multivariate Analysis, Predictive Modeling, Logistic Regression, Data Mining

Introduction: Employee attrition, or the reduction in staff within a company, is a prevalent challenge in industries such as IT and BPO. The reasons for attrition can vary, including factors like divergent career goals, salary disparities, geographical preferences, and demographic differences Denve and McMahon (1992). Unpredictable attrition can result in significant losses for organizations. To mitigate such losses, this research paper proposes a machine learning approach for predicting employee attrition.

Controlling attrition is vital for the overall health and success of an organization. Human resources development constitutes a substantial portion of an organization's expenses. This includes expenditures on salaries, bonuses, training programs, and other employee benefits. When employees' aspirations are not met, it can lead to uneven attrition, which can have adverse effects on the company's growth.

To ensure accurate risk planning, it is essential to possess a comprehensive understanding of the data available, encompassing both company users and employees. Developing such systems involves the identification of anomalies within the data, which can stem from inconsistent values and missing data points. The complexity of anomaly detection techniques may vary, along with the requisite amount of data and relevant features.

This research paper introduces modeling techniques that aid management in effectively planning for the uneven attrition of employees within their organizations. These techniques have been successfully utilized in prediction systems employed by reputable services across diverse industries.

The proposed model focuses on providing precise predictions for employee turnover, necessitating a well-suited dataset for training and validation purposes. The implemented machine learning methods are thoroughly evaluated, and the resulting outcomes are collected. The primary objective of this paper is to design and implement a prediction model that accurately forecasts employee attrition within a company. Prior to addressing this problem, an extensive review of relevant literature is conducted. The next part presents the proposed architecture and methodologies employed to tackle

this issue, including pseudo-codes of the prediction and comparison algorithms. Furthermore, illustrate the behavior of the proposed model through a specific scenario. Finally, presents the results and conclusions derived from the proposed model, comparing it with other related systems.

Literature Review: The term "attrition" or "turnover" Price and Mueller (1981) refers to the ratio of employees who have left an organization over a specific time period, divided by the total number of employees working in the organization during that same period. Employee attrition is a widely studied phenomenon Schwab (1991), yet definitive conclusions have remained elusive. The departure of employees has significant implications for organizations, making it a pressing issue that demands attention. Researchers argue that high turnover rates, if not effectively managed, can have adverse effects on organizational growth Wasmuth and Davis (1993).

Controlling employee attrition poses a complex and challenging task. Mobley Mobely (1982) conducts an analysis to identify the reasons behind employee attrition. Understanding employees' withdrawal behavior in detail can assist management in finding solutions to this pervasive issue that affects organizations. It enables management to identify the factors that compel employees to willingly leave company. However, many managers struggle to comprehend and accept this situation within their organizations due to their limited perspective and insufficient understanding of its repercussions. Nevertheless, identifying the primary causes and quantifying the reasons for turnover can offer valuable insights to managers seeking to make a positive impact within their organization.

The turnover ratio of employees is influenced by a multitude of factors, yet there is no universally applicable theory or set of rules that comprehensively explains it. Extensive research efforts have been dedicated to exploring the determinants of employees' intentions to quit, examining various aspects such as their backgrounds (Nagadevara et al., 2008; Kevin et al., 2004; Saks, 1996). Employees leave organizations for diverse reasons, which are then utilized to predict their intentions to leave and actual job departures. Contributing factors to job turnover encompass (i) commitment, (ii) job dissatisfaction, (iii) performance appraisal system, (iv) unclear expectations from senior and peers, (v) job profile, (vi) salary, (vii) career advancement and (vii) location. The nature of attrition varies among different types of employees; for example, labor encounter challenges linked to environmental and managerial factors. These factors encompass elements such as (i) organizational ethos and values, (ii) managerial style, (iii) equitable compensation, (iv) mutual support among employees, (v) organizational trust, (vi) manageable workload, (vii) career development, and (viii) job satisfaction. Numerous studies have identified causes of employee turnover, ranging from job-related issues to factors directly under the employer's control, including dissatisfaction with working conditions, conflicts with supervisors, or inconsistencies in salary matters. Understanding the underlying causes of job turnover is vital for identifying concerns within organizations facing attrition risks, as these causes are within the direct influence of the employer.

Assessing attrition risk within an organization can be accomplished using machine learning techniques that analyze multiple features Khare et al. (2015). A logistic regression-based model has proven successful in accurately predicting turnover by fitting a logistic curve to the inherently random data. This model utilizes a distinct dataset containing employee information to establish a risk equation, which is subsequently employed to assess attrition risk among the current employee cohort. High-risk clusters can be identified, and further analysis can be conducted to uncover the reasons behind attrition, empowering management to develop strategies for mitigating this risk.

In a similar vein, a novel algorithm called Data Mining Evolutionary algorithm (DMEL) Wai et al. (2003) has been developed for a carrier company to predict both the likelihood of a customer switching to another company and the specific probability of such an occurrence. The algorithm revealed that when a customer is at risk of leaving, the carrier selects a set of loyalty programs with special offers and services at significantly reduced rates to retain them. DMEL has demonstrated accurate results by generating insightful classification rules and diverse churn rates when applied to real-time employee data.

The subsequent section of this paper provides a comprehensive account of the approach employed to develop model, including process flows, relevant data sources, and the resulting outputs.

Research Methodology:

The Design: Outlined below is the standard O.S.E.M.N (Obtain, Scrub, Explore, Model, and Interpret) methodology employed in this research, serving as the foundation for the analysis:

- Obtain the data: The dataset used in this study was sourced from Kaggle's website. This benchmarked dataset comprises 17,790 employees and includes nine distinct features.
- Scrub or clean the dataset: Raw datasets obtained from external sources often contain errors, missing values, or require preprocessing to ensure suitability for meaningful analysis. Failure to adequately clean the dataset prior to analysis or utilizing it as input for machine learning algorithms can lead to inaccurate outcomes, resulting in irrelevant insights and hindering the decision-making process. Common errors encountered in datasets include:

  - Duplicate records
  - Missing ranges
  - Raw data format
  - Incorrect numerical scales
  - Incorrect data formats

- Explore dataset: Step involves various operations aimed at gaining insights into the statistical properties of the dataset and understanding its features. Techniques such as visual graphs are employed to analyze boundaries and distributions. The following types of graphs are used:

  - Box Plot
  - Line chart
  - Histogram
  - Scatter Plot

In certain cases, when dealing with datasets that possess a large number of independent variables, it is referred to as a high-dimensional dataset representation. Therefore, dimensionality reduction becomes crucial for dimensions reduction to a more manageable number. In this paper, feature selection is employed, which is carried out using logistic regression.

The dependent variable in this study is the feature ‘Inactive' which denotes the number of employees who have left the organization. The remaining nine variables are treated as independent variables, including:

- satisfaction_level
- evaluation_score
- number_of_projects
- average_working_monthly_hours
- number_of_years_spent_at_organization
- accident_at_work
- got_promotion
- department
- salary_level

To assess the relationships between variables, correlation matrices and heat maps are utilized. These visualizations enable the identification of positive or negative correlations among the variables under consideration. For instance, variables such as 'evaluation_score', 'average_working_monthly_hours,' and 'number_of_projects,' demonstrate positive correlations, indicating that employees who work longer hours tend to receive higher evaluation scores. On the other hand, the variable 'satisfaction_level' exhibits a negative correlation with the 'Inactive' variable, suggesting that employees with lower satisfaction levels are more inclined to leave the organization.

Furthermore, the analysis encompasses comparisons between the 'salary_level' and the 'Inactive' variable, different departments and the 'lnactive' variable, as well as the 'number_of_projects' and the 'lnactive' variable. These comparisons offer insights into the relationship between these variables and the likelihood of employees leaving the company.

Modeling the dataset: Following an exploration of the dataset's extreme boundaries and features, the subsequent step involves training the dataset using suitable algorithms. In classification problems, algorithms such as Decision trees, Support Vector Machines, Random Forests, and Naive Bayes classifiers are commonly employed. However, when the objective is to predict discrete values within a given domain, regression modeling is preferred. Notably, this paper addresses both aspects of predicting the attrition rate and classifying employees into those likely to leave the company or remain.

Interpreting the results: Results obtained from modeling the dataset using appropriate algorithms necessitate analysis and processing to facilitate decision-making. For instance, the 'Inactive%' obtained from Equation 1 is merely a numerical value unless it is associated with a meaningful interpretation, such as a risk factor as discussed in [9]. Graphical representations derived from exploratory analysis are comprehensively interpreted to provide insights into the overall problem statement. To assess the accuracy and effectiveness of the utilized algorithms, this paper presents the results in the form of tables displaying Precision, Recall, and F1-score. These parameters aid in summarizing the solution effectively.

Methodology: The following is the pseudo-code for the machine learning models developed in this study:

Logistic Regression:

Given, {[xi, yi]} where i goes from 1 till m.
Initialize variable d=<1,...1>T
Repeat the steps until algorithm converges:
for each j=0,...,n:
for dj'=dj + Σi(yi-ha(Xi))xji for each j= 0,...,n:
dj = dj
Output d

Decision Tree:

Decsion_tree_algo(Sample S, Attribute_list A)
Create a node N
Samples if belong to same class C; label the node N with C and terminate
A if is null; label N with most common class C in majority voting
Select a belongs to A, with the highest information gain; Name N with n
For value t of n:
Grow a branch from N with the condition n=t; Assume S to be the subset of samples in S with n=t
If St is empty; join a leaf labelled with most common class in S
Else attach node generated by Decision_Tree(St, A-n)

After applying logistic regression to our testing data, we obtained the following equation:

$$\text{Inactive \%} = -3.7 * \text{satisfaction\_level} + 0.20 * \text{evaluation\_score} + 0.170 * \text{number\_of\_years} + 0.18 \quad (1)$$

In Equation 1, the constant value of 0.18 represents the cumulative impact of other independent variables that were not included in our proposed model.

To illustrate Equation 1, let's consider an example using values from the test dataset, which produced the following

results:

$$\text{lnactive\%} = -3.7 * \text{satisfaction\_level} + 0.20 * \text{evaluation\_score} + 0.170 * \text{number\_of\_years} + 0.18 \quad (2)$$

The values for the different parameters from the data set are: (i) satisfaction_level = 0.9, (ii) evaluation_score = 0.7 and (iii) number_of_years = 5; Substituting the values in (2) we get:

$$\text{lnactive\%} = -3.7*0.9+0.20*0.7+0:17*5+0.18 \quad (3)$$

$$\text{lnactive \%} = -3:33+0:14+0.85+0:18 \quad (4)$$

$$\text{lnactive \%} = -2.16 \quad (5)$$

final value of employee retention is predicted by using the following equation:

$$\text{final}=\exp(a)/1 + \exp(a) \quad (6)$$

where a = lnactive% (7)

final= 0.23 (8)

This means that the employee has a chance of 23% chance of leaving the company. According to Khare at al. (2015) the risk factor analyzed. It is found that there is not much risk and the employee will be loyal to the company.

Results: This research paper introduces the application of Logistic Regression as a means of assessing the risk of employee attrition and predicting turnover within a company. The employee dataset is extensively analyzed using our proposed model, which is implemented using the 'NumPy' numerical Python library. To provide a comprehensive comparison, we evaluate our Logistic Regression model against other established supervised learning algorithms including Decision Tree, Random Forests, and Adaboost.

The analysis reveals a strong correlation between salary levels and the likelihood of employee attrition. Employees with low to medium salary levels demonstrate a higher propensity to leave the organization. Furthermore, an examination of departments in relation to the 'lnactive' variable identifies Sales and Technical departments as experiencing the highest rates of employee attrition.

We delve deeper into the relationship between the number of projects employees handle and their propensity to leave. Results indicate that a significant majority of employees who left the company were handling more projects. Notably, employees managing 7 projects had the highest attrition rate, suggesting that a higher workload increases the likelihood of turnover.

The evaluation score emerges as a significant factor influencing the 'lnactive' variable. Employees with either very high or very low evaluation scores are more likely to leave the company.

Additionally, the analysis of average working monthly hours reveals a U-shaped relationship with attrition. Employees with significantly fewer or significantly more working hours tend to leave the company more frequently.

An interesting observation arises from examining clusters of evaluation scores and satisfaction levels. Many employees with good evaluation scores express dissatisfaction with their work, while those with low evaluation scores and low satisfaction levels also show a higher likelihood of leaving. Conversely, as evaluation scores and satisfaction levels increase, the likelihood of an employee leaving decreases.

Furthermore, we employ the Decision Tree algorithm to identify the most influential features impacting employee attrition. The analysis highlights satisfaction level, number of years, and evaluation score as more significant than other independent variables.

The predictions obtained from our logistic regression models are presented in Tables 1, and the attributes of these tables are thoroughly documented for reference.

	Precision	Recall	F1-score
Active	0.90	0.76	0.82
Inactive	0.48	0.73	0.58
Average–score	0.80	0.75	0.76

Table 1 Accuracy of Logistic Regression

Precision: It is also known as positive predicted value. It is given by:

TPV/ (TPV + FPV)

TPV– true positive value.

FPV–false positive value.

FNV –false negative value.

Recall: It is also known as sensitivity. It is given by:

TPV/ (TPV + FNV)

F1- Score: weighted average of the precision and recall.

active: the employee has not left the organization.

Inactive (already mentioned above): the employee has left the organization.

The comprehensive analysis yields the following key findings:

- Employees, who were underworked, typically working less than 150 hours per month, exhibited a higher likelihood of leaving their positions.
- Conversely, employees who were overworked, typically working more than 240 hours per month, also showed a tendency to leave the company.
- Individuals with extremely high or low evaluation scores were more prone to attrition within the organization.
- Employees with low or medium salaries were predominantly factor responsible for leaving the company.
- Employees handling more projects had a higher likelihood of leaving the company.
- Satisfaction level emerged as the most crucial factor influencing attrition.
- Employees who had been with the company for 4 to 5 years contributed significantly to the uneven attrition patterns observed.

Conculsion: This research paper employs Logistic regression as a valuable approach to assess the risk of employee attrition and predict turnover within the company. By utilizing this model, the organization can maximize employee retention and gain valuable insights into the underlying causes of attrition, empowering management to make informed decisions. While Logistic regression yielded an accuracy of 80%, the model's strength lies in its adaptability to changing feature sets within the dataset, enhancing its superiority. Notably, the analysis highlights the crucial role of level of satisfaction, number of years of service, and evaluation score as primary factors influencing the attrition within organizations.

References

1. Denver, A., McMahon, F.: Labour turnover in London hostels and the cost-effectiveness of preventive measures. Int. J. Hosp. Manag. 11–2, 143–540 (1992)
2. Kevin, M.M., Joan, L.C., Adrian, J.W.: Organizational change and employee turnover. Pers.Rev. 33(2), 161–166 (2004)
3. Khare, R., Kaloya, D., Choudhary, C.K.: Employee attrition risk assessment using logistic regression analysis. In: IIMA International Conference on Advanced Data Analytics, Business Analytics (2015)
4. Mobley, W.H.: Employee Turnover: Causes, Consequences, and Control. Addison-Wesley Publishing, Philippines (1982)
5. Nagadevara, V., Srinivasan, V., Valk, R.: Establishing a link between employee turnover and withdrawal behaviours. Appl. Data Mining Tech. Res. Pract. Hum. Resour. Manag. 16 (2), 81–99 (2008)
6. Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., Feinstein, A.R.: A simulation study of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. 49, 1373–137 (1996)
7. Price, J.L., Mueller, C.W.: A causal model of turnover for nurses. Acad. Manag. J. 24, 543–565 (1981)
8. Saks, A.M.: The relationship between the amount of helpfulness of entry training and work outcomes. Hum. Rel. 49, 429–451 (1996)
9. Schwab, D.P.: Contextual variables in employee performance-turnover relationships. Acad. Manag. J. 34, 966–975 (1991)
10. Wai, H.A., Chan, K.C.C., Yao, X.: A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Trans. Evol. Comput. 7(6), 532–545 (2003)
11. Wasmuth, W.J., Davis, S.W.: Managing employee turnover: why employees leave. Cornell HRA Q., 11–18 (1993)