European Economic Letters
ISSN 2323-5233
Vol 13, Issue 4 (2023)
http://eelet.org.uk

Predicting Employee Attrition Using Machine Learning

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

Abstract: This paper delves into the necessity and significance of predicting employee attrition based on their individual circumstances. A key challenge addressed in this study pertains to the presence of noise within such datasets, as well as the inherent inaccuracies stemming from interdependencies among the data points. To tackle this challenge, the paper showcases the effectiveness of the LGBM algorithm in training models with noisy data while maintaining high accuracy levels. Comparative analyses between LGBM and alternative algorithms highlight the algorithm's superior performance, particularly in terms of accuracy in predicting termination.

In terms of future research directions, one potential avenue involves implementing automatic employee allocation. By leveraging the same algorithms and utilizing different datasets, it becomes possible to train models that can predict an employee's aptitude and specialization within the workplace. Consequently, this could facilitate automatic task allocation or work assignment processes.

Keywords: Machine Learning; Supervised Classification; Retention Prediction; Gradient Boosting; Naïve Bayesian; K Nearest Neighbour.

Introduction: Numerous facets concerning a company warrant discussion, such as the organizational milieu, workload, and the delicate equilibrium between professional and personal life. The absence of an adequate work-life balance or an excessive burden may prompt an employee to seek employment elsewhere. Organizations address this issue by endeavoring to prognosticate an employee's state of being based on their treatment and emotional well-being within the company. Such prognostication relies upon inputs provided by managers, team leaders, and the Human Resources (HR) department. However, the datasets created for this purpose are highly susceptible to containing substantial amounts of erroneous and imprecise data. Many companies have not prioritized investments in robust HR Information Systems (HRIS) technologies capable of effectively capturing comprehensive employee data throughout their tenure. This reluctance can be attributed, in part, to a flawed perception of the advantages and costs associated with such investments. Additionally, measuring the return on investment for HRIS implementations proves to be a challenging task. Consequently, the presence of noisy datasets significantly impedes the ability of these algorithms to generalize effectively.

This research paper delves into the issue of employee attrition and explores machine learning methodologies as potential solutions. Specifically, it focuses on leveraging data from the Human Resource Information System (HRIS) dataset. The paper culminates in a conclusion highlighting the superior accuracy of the Light Gradient Boosting Machine (LGBM) algorithm.

The paper is organized into several distinct sections. Section II elucidates the problem statement and highlights the imperative to address it. Section III delves into a comprehensive explanation of various supervised machine learning techniques. In Section IV, the paper delves into the experimental analysis, detailing the dataset, pre-processing methods employed, and the metrics employed to evaluate performance. Section V presents the results obtained. Finally, Section VI concludes the paper by offering recommendations based on the findings.

Literature Review: Employee attrition can be conceptualized as the departure of valuable intellectual capital from an organization's workforce. Such departures can be categorized as either accidental or voluntary, with this paper focusing specifically on voluntary turnover. Extensive research has identified several key variables crucial for estimating turnover, including overall work satisfaction, age, tenure, and salary. Additionally, studies have highlighted the significance of personal factors such as age, ethnicity, and education in predicting attrition. Noteworthy factors

influencing turnover prediction encompass remuneration, working conditions, supervisory practices, promotional opportunities, and job satisfaction. (Stoval and Bontis 2002; Singh et al. 2020; Finkelstein et al. 2013; Holtom et al. 2008; von Hippel et al. 2013; Peterson 2004; Sacco and Schmitt 2005; Allen and Griffeth 2001).

A high turnover ratio exerts a multitude of adverse effects on a company. The process of identifying and securing a new employee possessing the specific knowledge and skills necessary for the organization can prove arduous and time-consuming. Furthermore, the direct impact on productivity is undeniable. The recruitment of new employees entails substantial costs, encompassing the entire gamut of activities ranging from screening candidates based on requisite knowledge to training the selected individuals to achieve the desired competency level (Liu et al. 2012; Swider and Zimmerman 2010; Heckert and Farabee 2006).

In light of these challenges, companies resort to employing various machine learning and mathematical algorithms as preventive measures against attrition.

Research Methodology: In the field of machine learning, classification can be approached through two main methodologies: supervised learning and unsupervised learning. Supervised learning involves training the algorithm on a dataset where each data point is associated with a known output. The algorithm learns the patterns that lead to specific outputs and aims to generalize this knowledge with guidance. In the context of supervised learning, the dataset contains labeled output values to be predicted, and the algorithm learns how to make predictions through a process of analysis and generalization. On the other hand, unsupervised learning operates without knowledge of the specific categories to which each data point belongs. It seeks to uncover patterns within the data and generalize based on intrinsic features, ultimately assigning output labels during the training process.

This research paper specifically focuses on classification using supervised learning, where two distinct labels, namely "Inactive" and "active," are assigned to the data points.

A. Logistic Regression: Logistic regression is a fundamental linear classification algorithm commonly used for differentiating between linear models. It operates on the sigmoid mathematical function and is particularly effective for predicting binary and categorical classes. Regularization is often applied in logistic regression to mitigate overfitting.

B. Naïve Bayes: Naïve Bayes is a widely used classification technique known for its simplicity. This algorithm relies solely on probabilities for prediction. It assumes that each variable is independent of the others, requiring only a small portion of the dataset to estimate the mean and variance.

The Bayes' rule is defined as follows: The target function is represented as $P(Y|X) = X \rightarrow Y$. The training data is utilized to estimate $P(X|Y)$ and $P(Y)$. Using these probabilities and Bayes' rule, new X values can be classified into different labels.

C. LightGBM: LightGBM is a high-performance gradient boosting algorithm primarily used for ranking and classification tasks. It is based on the decision tree algorithm.

Unlike other algorithms, LightGBM grows trees in a vertical manner instead of a horizontal one. This means that LightGBM grows trees leaf-wise, while other algorithms grow them level-wise. It avoids converting to one-hot coding and offers significantly faster processing times compared to one-hot coding.

The Design: The selected dataset represents the distribution of employees across various locations in the United States. It consists of two distinct labels: "Inactive" (0) and "Active" (1). Initially, all employees are labeled as active (0) and remain in the company for a period of four months. After this duration, employees leave the company, resulting in a change of their class label to Inactive (1).

The dataset utilized in this research is sourced from Kaggle and comprises numerous features such as pay, age, and team-related characteristics, among others, which are employed for prediction purposes. The dataset encompasses a total of 33 features, including 27 numeric and 6 categorical variables.

A. Data Preprocessing: The initial step in data preprocessing involved cleaning the dataset by eliminating erroneous and noisy data. Missing numerical values were replaced with zeros, while rows with missing categorical data were removed. Zero values were assigned to fields like the number of promotions for employees with missing data to ensure the model trained with enhanced accuracy. Subsequently, categorical features were encoded using one-hot encoding, transforming them into binary fields.

B. Model Training: The dataset was divided into an 80:20 ratio, with 80% allocated for training and 20% for testing purposes. Regularization techniques were employed, and penalty hyperparameters were set for each algorithm. The training dataset was utilized to train the models with their respective optimal configurations. Subsequently, the trained models were employed to make predictions on the remaining 20% of the data.

C. Evaluation: The performance evaluation of different algorithms was based on prediction scores or accuracy achieved under their optimal training conditions. The models were tested on the dedicated testing dataset, which represents 20% of the complete dataset. The accuracy achieved on this dataset provided the evaluation metrics for comparing the performance of different algorithms. Additionally, a confusion matrix was generated to facilitate a comparative analysis among the various models.

Results: The dataset comprises a diverse group of employees from an organization, encompassing individuals with varying ages, genders, pay levels, team assignments, and backgrounds. These employees have all worked for a minimum period of four months before departing the company, either voluntarily or due to external factors. The dataset, sourced from the Kaggle website, serves as a valuable resource for training models capable of predicting employee attrition rates.

To determine accuracy from all the classes (positive and negative), how many of them have predicted correctly. Following formula has been used.

$$\frac{TP+TN}{(TP+TN+FP+FN)} \qquad \text{eq } \ldots\ldots\ldots(1)$$

Algorithm	**Prediction Score/ Accuracy**
Logistic Regression	87.17
Naïve Bayesian	89.22
LightGBM	95.20

Table 1 – Prediction Score Table

Confusion Matrix

		Predicted Values	
True Label	**Didn't Leave Job**	3261 **(TP)**	167 **(FN)**
	Left Job	410 **(FP)**	662 **(TN)**
		Didn't Leave Job	**Left Job**

European Economic Letters
ISSN 2323-5233
Vol 13, Issue 4 (2023)
http://eelet.org.uk

Fig 1 – Logistic Regression Confusion Matrix

Confusion Matrix

		Predicted Values	
True Label	Didn’t Leave Job	3310 **(TP)**	118 **(FN)**
	Left Job	367 **(FP)**	705 **(TN)**
		Didn’t Leave Job	**Left Job**

Fig 2 – Naïve Bayesian Confusion Matrix

Confusion Matrix

		Predicted Values	
True Label	Didn’t Leave Job	3390 **(TP)**	38 **(FN)**
	Left Job	173 **(FP)**	899 **(TN)**
		Didn’t Leave Job	**Left Job**

Fig 3 – LightGBM Confusion Matrix

As observed from the results presented in Table 1, all three algorithms have performed well but LGBM algorithms demonstrate significantly superior performance compared to the other two approaches. Conversely, LGBM leverages the power of boosting techniques, enabling effective training on noisy datasets and accurate classification of data points within such datasets.

In particular, LGBM algorithm effectively addresses the challenge of overfitting through its inherent regularization mechanisms, mitigating the risk of excessively fitting the training data.

Conclusion: This paper delves into the necessity and significance of predicting employee attrition based on their individual circumstances. A key challenge addressed in this study pertains to the presence of noise within such datasets, as well as the inherent inaccuracies stemming from interdependencies among the data points. To tackle this challenge, the paper showcases the effectiveness of the LGBM algorithm in training models with noisy data while maintaining high accuracy levels. Comparative analyses between LGBM and alternative algorithms highlight the algorithm's superior performance, particularly in terms of accuracy in predicting termination.

In terms of future research directions, one potential avenue involves implementing automatic employee allocation. By leveraging the same algorithms and utilizing different datasets, it becomes possible to train models that can predict an employee's aptitude and specialization within the workplace. Consequently, this could facilitate automatic task allocation or work assignment processes.

References:

1. Alao, D., & Adeyemo, A. B. (2013). Analyzing employee attrition using decision tree algorithms. Computing, Information Systems, Development Informatics and Allied Research Journal, 4.
2. Allen, D. G., & Griffeth, R. W. (2001). Test of a mediated performance – Turnover relationship highlighting the moderating roles of visibility and reward contingency. Journal of Applied Psychology, 86(5), 1014-1021.
3. Finkelstein, L. M., Ryan, K. M., & King, E. B. (2013). What do the young (old) people think of me? Content and accuracy of age-based metastereotypes. European Journal of Work and Organizational Psychology, 22(6), 633-657.
4. Heckert, T. M., & Farabee, A. M. (2006). Turnover intentions of the faculty at a teaching-focused university. Psychological reports, 99(1), 39-45.
5. Holtom, B., Mitchell, T., Lee, T., & Eberly, M. (2008). Turnover and retention research: A glance at the past, a closer review of the present, and a venture into the future. Academy of Management Annals, 2, 231-274.
6. Hong, W. C., Wei, S. Y., & Chen, Y. F. (2007). A comparative test of two employee turnover prediction models. International Journal of Management, 24(4), 808.
7. Jantan, H., Hamdan, A. R., & Othman, Z. A. (2011). Towards Applying Data Mining Techniques for Talent Managements. 2009 International Conference on Computer Engineering and Applications, IPCSIT vol.2, Singapore, IACSIT Press.
8. Liu, D., Mitchell, T. R., Lee, T. W., Holtom, B. C., & Hinkin, T. R. (2012). When employees are out of step with coworkers: How job satisfaction trajectory and dispersion influence individual-and unit-level voluntary turnover. Academy of Management Journal, 55(6), 1360-1380.
9. Marjorie, L. K. (2007). Predictive Models of Employee Voluntary Turnover in a North American Professional Sales Force using Data-Mining Analysis. Texas, A&M University College of Education.
10. Nagadevara, V., Srinivasan, V., & Valk, R. (2008). Establishing a link between employee turnover and withdrawal behaviours: Application of data mining techniques.
11. Peterson, S. L. (2004). Toward a theoretical model of employee turnover: A human resource development perspective. Human Resource Development Review, 3(3), 209-227.
12. Punnoose, R., & Ajit, P. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. International Journal of Advanced Research in Artificial Intelligence, 5(9).
13. Sacco, J. M., & Schmitt, N. (2005). A dynamic multilevel model of demographic diversity and misfit effects. Journal of Applied Psychology, 90(2), 203-231.
14. Singh, S. P., Singh, P., & Mishra, A. (2020). Predicting potential applicants for any private college using LightGBM. In 2020 International Conference on Innovative Trends in Information Technology (ICITIIT), IEEE.
15. Stoval, M., & Bontis, N. (2002). Voluntary turnover: Knowledge management – Friend or foe? Journal of Intellectual Capital, 3(3), 303-322.
16. Swider, B. W., & Zimmerman, R. D. (2010). Born to burnout: A meta-analytic path model of personality, job burnout, and work outcomes. Journal of Vocational Behavior, 76(3), 487-506.
17. von Hippel, C., Kalokerinos, E. K., & Henry, J. D. (2013). Stereotype threat among older employees: Relationship with job attitudes and turnover intentions. Psychology and aging, 28(1), 17.