Predicting Employee Turnover: A Systematic Machine Learning Approach for Resource Conservation and Workforce Stability †

Parmod Kumar 1, Sagar Balu Gaikwad 2*, Shunmugavel Thanga Ramya 3, Tripti Tiwari 4*, Mohit Tiwari 5,* and Binod Kumar 6

1 Department of Electronics and Information Engineering, Jiangxi University of Engineering, Xinyu 338000, China; dr.p.kumar2019@gmail.com
2 Department of Management, MET Institute of Management, Mumbai 422003, India; s.balugaikwad1993@gmail.com
3 Department of Computer Science and Design, R.M.K. Engineering College, Kavaraipettai, Tiruvallur 601206, India; thangaramya@gmail.com
4 Department of Management Studies, Bharati Vidyapeeth Institute of Management and Research, Delhi 110063, India; tripti.tiwari@bharatividyapeeth.edu
5 Department of Computer Science and Engineering, Bharati Vidyapeeth’s College of Engineering, Delhi 110063, India; mohit.tiwari@bharatividyapeeth.edu
6 Department of Master of Computer Applications, JSPM'S Rajarshi Shahu College of Engineering, Pune 411033, India; binod.istar.1970@gmail.com
*
Correspondence: mohit.tiwari@bharatividyapeeth.edu
†
Presented at the International Conference on Recent Advances in Science and Engineering, Dubai, United Arab Emirates, 4–5 October 2023.

Abstract: A company’s most valuable resource is its workforce, which includes each worker. Because of the crucial role that employees play in the success of an organization, measuring employee turnover rate has become one of the most important metrics that businesses are concentrating on in the modern era. Attrition may occasionally arise owing to unavoidable circumstances such as moving to a distant place, retirement, etc. But when attrition begins creating holes in the pockets of an organization, it is necessary to monitor the situation closely. When hiring new staff, a company must use a significant quantity of its available resources. The process of rehiring employees needs to be eliminated, and a strong workforce needs to be maintained, so it is necessary to adapt the analysis of systematic machine learning models. From these models, a suitable model that gauges the risk of attrition may then be selected. This not only helps an organization save money by preserving its resources but also assists in preserving the status quo of its staff.

Keywords: employee attrition; machine learning; descriptive analysis; deep learning; random forest

1. Introduction

Any business can benefit from having more workers. When an individual begins working for a company, it is inevitable that they will quit at some point in the future for a variety of reasons. Attrition can be defined as the departure of any employee owing to events that may or may not be within their control, such as retirement, death, transfer, or the pursuit of better opportunities. When an employee is being hired, the hiring company invests a significant amount of time and a significant number of resources in the process [1]. When an employee’s departure begins to have a negative impact on a company, it becomes a source of concern for everyone in the company, but particularly for its human resources department. Such a business not only suffers the loss of its competent experts because of the departure of qualified employees but must also rehire and educate the individual who replaces them [2]. This results in a decrease in the company’s staff and has a negative impact on the company. There has been a tremendous uptick in opportunities across the board as a direct result of growing globalization, particularly in the period after the epidemic. An employee makes the decision to leave one company and join another to pursue new...
opportunities and advance their professional development [3]. The loss of employees due to attrition has a negative impact on a company’s operations for a limited amount of time. Incorporating artificial intelligence into the process of predicting attrition is one way to keep workforce size stable while also cutting expenditures.

This article presents a discussion of the numerous approaches that may be taken to forecast employee turnover, and it also includes an analysis of the most effective solution that was conducted by comparing different models. Figure 1 shows the various reasons why an employee may decide to leave an organization.

![Figure 1. Reasons for employee attrition.](image)

2. Literature Review

Several scholars have investigated the factors that lead to employee turnover as well as its consequences. According to one of the corresponding papers, ref [4], the upkeep of skilled and deserving employees is a crucial factor that HR departments need to pay attention to in order to be successful. The cited study identified the most pertinent metrics that could be of assistance in the endeavour of forecasting employee turnover. It was pointed out that an employee’s level of education and experience are directly proportionate to the number of work opportunities that are available to them. It was also claimed that some of the most pleasant things that help in sustaining a workforce include having a decent work–life balance, having healthy connections with co-workers, having better policies, and so on. According to the findings of another study of this kind [5,6], for a business to realize the greatest possible profit, it must treat its workforce with the utmost respect and show a great deal of concern. This can be accomplished by putting more of an emphasis on the creation of new opportunities and by introducing innovative technologies, both of which assist employees in keeping their interest in the company for which they work [7,8]. The findings of the cited study also emphasize how important it is for organizations to regularly host training programs, cultural events, and other types of gatherings. These sorts of activities help in lowering the barrier of communication, which, in turn, encourages interaction and growth for the individual. The primary purpose of this research was to provide an explanation of why it is essential for an organization to have a culture of transparent work, ensuring that every individual is kept fully informed about the nature of their work and the results of their efforts [9,10].

Sri Ranjitha Ponnuru and colleagues predicted staff turnover using Machine Learning algorithms. IBM HR research informed the forecast. They made predictions using Logistic Regression and obtained 85% accuracy [11].
To aid HR recruiters in making better placement and recruiting choices in the real world, one should present a complete analytics platform. The proposed architecture begins with a single-hire-level local prediction technique for recruitment performance. In the second phase, a mathematical model is used to provide an organization-wide recruiting optimization technique that considers multiple levels of analysis [12].

The Extreme Gradient Boosting (XGBoost) methodology, developed by Rohit Punnoose and Pankaj Ajit, is more reliable than other methods due to its inclusion of a regularisation formulation [13]. Using data from a multinational retailer’s HRIS, we show that XGBoost is much more accurate at predicting employee turnover than six commonly used supervised classifiers.

To reduce employee turnover, Shikha N. Khera and Divya devised a model to forecast why workers leave their companies [14]. A support vector machine was used to create a predictive model (SVM).

To forecast employee turnover, Sarah S. Alduayj and colleagues conducted three primary studies using synthetic data generated by IBM Watson [15]. In the first experiment, ML algorithms such as SVM, KNN, and random forest were trained. In the second experiment, the adaptive synthetic (ADASYN) strategy was employed to correct for class imbalance, and then the machine learning models were retrained on the updated dataset. The third test aimed to achieve class-balance via the manual under-sampling of the data. Training on an ADASYN-balanced dataset using KNN (K = 3) yielded the best results (0.93 F1-score). Eventually, an F1-score of 0.909 was attained using feature selection and Random Forest using just 12 of the possible 29 characteristics.

In this study, DT-, RF-, LR-, and ensemble-technique-based classifiers were trained and evaluated using the IBM attrition dataset, which was collected and analysed by Aseel Qutub et al. [16].

Christopher Boomhower et al. investigated the kinds of jobs our model is applicable to and the kinds of characteristics connected to an employee’s decision to leave a company [17]. According to the conclusions of Mishra and Mishra’s research, one of the most significant strategic roles that HR will play in the future is in limiting attrition and reducing its negative impacts. In order to keep a high-performing talent pool within a business, the relevant elements must be identified, and effective retention methods must be implemented [18].

Setiawan and his group used logistic regression to study turnover rates in the workplace. Management may utilise the findings to determine what changes they can make to their workplace to keep the vast majority of its employees [19]. The cited study describes the methodology followed, the steps performed, the data used, and the results obtained in building an attrition risk model. In addition, Rupesh Khare et al. tried to identify the focus areas and best practices regarding employee retention at various points in an employee’s tenure with a business [20].

The advent of artificial intelligence has resulted in unprecedented expansion across all industries. It has been helpful in locating solutions to a wide variety of difficult challenges. The issue of employee turnover is one that is currently being discussed across the globe. Artificial intelligence has the potential to provide a robust solution to this challenge that may be implemented by a variety of companies. Companies all over the world are benefitting from the application of machine learning to forecasting employee turnover. Research of a similar nature has been carried out before, in which a variety of models, including Support Vector Machines, Random Forests, KNN classifier, and XG Boost, were put to the test. The information regarding these kinds of studies is presented in Table 1 below.

<table>
<thead>
<tr>
<th>Sr no.</th>
<th>Reference</th>
<th>Object of Study</th>
<th>Recommend Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>[3]</td>
<td>Employee attrition prediction</td>
<td>KNN classifier</td>
</tr>
<tr>
<td>2.</td>
<td>[4]</td>
<td>Early attrition prediction</td>
<td>Random Forest</td>
</tr>
</tbody>
</table>
3. Methodology

3.1. Design, Architecture, and Dataset

As per the architecture shown in Figure 2, the system that has been proposed was built using several different types of machine learning. Each model makes use of the same dataset to make a prediction regarding attrition. The collection is made up of a variety of different personnel records (both past and present) [10]. The incoming dataset is initially subjected to cleaning and preprocessing, which involves the management of all missing values, NaN values, etc., as well as the removal of unwanted columns. After that follows the process of model construction, which involves considering several different models to make a prediction. The dataset is then divided into a training dataset and a test dataset, with the training dataset being the one that is utilized to train each model employed. Following a comparison of all the predictions based on the evaluation measures, the most effective model is proposed.

![Figure 2. Machine learning architecture.](image)

The information that pertains to employees is included in the open source dataset. All non-numerical values were each assigned a letter (A1, A2, A3, etc.), and all unnecessary factors were eliminated from consideration. Some of the factors that went into making the forecast are illustrated in Figure 3.

3.2. Machine Learning Algorithms

A. Logistic Regression:

It is generally agreed that Logistic Regression is one of the most useful statistical models available. In addition, it is a well-known method of data mining that scientists and other researchers employ for the examination of both proportional and binary types of datasets. One of the things that sets logistic regression apart from other types of regression is the fact that it may be used to analyse data from more than one class [8]. It is one of the most frequently used classification algorithms across the globe.

B. Decision Tree:

The tree structure is visualized whenever the term “tree” is used in the vernacular of computer science. Root, branches, and leaves make up the components of a decision tree. It is common practice to refer to the root node as the parent node. Nodes are used to represent each characteristic, while branches are used to indicate the connections that are...
made between the nodes [9]. The rules or choices are contained within these branches as depicted in Figure 4. It is expected that the leaf will serve as the result or outcome. CHAID, ID3, and CART are examples of some of the decision tree algorithms that are used most frequently [9]. This algorithm is applied to problems involving classification and can work fluidly with both continuous and categorical information.

**Figure 3.** Features used in prediction.

**Figure 4.** Decision Tree.

KNN is an algorithm for supervised machine learning that can be applied to issues involving classification as well as regression. The KNN algorithm makes predictions about the output by making use of knowledge about the input [16]. The input is then divided into the appropriate categories. The algorithm tends to look for the position that will provide the best results for a new datapoint to fit in. Following an analysis of the data points that were provided as input, a decision is made regarding the location of the new point. The algorithm that was employed in this study is as follows:

1. Step 1: The first step is to choose the number of K, i.e., the neighbourhood;
2. Step 2: The second step is to compute the distance (Euclidean);
3. Step 3: Using, positions of data points, locate the K individuals that are geographically closest to a given position;
4. Step 4: The fourth step is to tally the number of points earned in each category;
5. Step 5: The fifth step involves assigning the newly acquired points to a category in which the surrounding points are greater in number;

D. Support Vector Machines (SVMs):

Another popular variety of supervised machine learning models is called a support vector machine (SVM). Its primary application is in addressing classification difficulties, although it can also be utilized to address regression settings. The fundamental concept behind the method is to draw a line or establish a border that delineates n distinct groups or classifications inside the overall space [18]. As a new data point is added to this space, it can quickly locate its proper location within the categories that have been formed. A hyperplane is another name for the line that cuts across the middle of these classes. A classification problem is said to have a linear algorithm when it can be solved by drawing a single straight line. It is referred to as a non-linear support vector machine (SVM) when a straight line is insufficient, in which case a curved line is obtained instead [15].

E. Random Forest:

A machine learning technique, Random Forest is utilized for solving problems involving regression and classification. It is something that has been passed down from the idea of ensemble learning. It is comparable to the use of decision trees. The dataset is split up into numerous subsets so that this technique can consider the many different types of trees. Because of this procedure, several outcomes are combined into a single result, which is then calculated as an average of all the component results. The accuracy of the algorithm improves in proportion to the number of trees and sub datasets that are taken into consideration. As a result of this behaviour, it can manage an extremely large number of datasets [10]. The steps of the algorithm are shown below:

1. Step 1: The first step is to pick K datapoints at random from the training set;
2. Step 2: Constructing decision trees for each subset is the second step;
3. Step 3: Choose the number N that will represent the number of decision trees;
4. Step 4: Execute S1 and S2 once again;
5. Step 5: According to the predictions made by each tree, assign each new datapoint to the appropriate category.

F. Naive Bayes:

The Naive Bayes algorithm is a well-known supervised machine learning approach that was developed by applying the Bayes Theorem in its formulation [7]. Because of its probabilistic character, the operation of this algorithm is predicated on the likelihood that an item will be found. Although it is mostly utilized for problems requiring text categorization, it is adaptable enough to be utilized for a wide variety of classification issues [11].

The Bayes Theorem can be expressed using the following formula:

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}
\]

Whilst the likelihood probability is symbolized by the symbol \(P(B|A)\), the posterior probability is represented by the symbol \(P(A|B)\). Priority probability, abbreviated as \(P(A)\), is contrasted with Marginal Probability, abbreviated as \(P(B)\).

3.3. Dataset

For predicting employee turnover, one can consider using the “IBM HR Analytics Employee Attrition & Performance” dataset. This dataset contains various features related to employees’ demographics, job satisfaction, performance, work environment, and other factors that can potentially influence turnover. It includes a binary target variable indicating whether an employee has left a company (1) or stayed with it (0). This dataset is widely used
in machine learning research and has been made available on platforms like Kaggle, making it easily accessible for experimentation and analysis. Using this dataset, one can build predictive models to identify factors that contribute to employee turnover and develop strategies to retain valuable employees.

4. Results and Discussion

Implementation and Results

The information regarding some of the features concerning attrition is shown in a graphical format in Figure 5.

![Figure 5. Promotion vs. attrition.](image)

The relationship between promotion and attrition is depicted in the graph represented in Figure 6. It is clear from looking at the graph that an employee is more likely to remain with a company if they have been given a promotion during their time there. The effect that gender has on the rate of employee turnover was also analysed, revealing that male candidates have a greater chance of remaining employed by an organization compared to female candidates.

![Figure 6. Graphical representation of results.](image)

The Table 2 shows the test accuracy of each model. From the table, it can be gleaned that Logistic Regression performed best, as it has the greatest accuracy, followed by Random Forest.
Table 2. The accuracy obtained using various algorithms.

<table>
<thead>
<tr>
<th>MODEL</th>
<th>ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>87.71%</td>
</tr>
<tr>
<td>KNN Classifier</td>
<td>59.22%</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>86.59%</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>83.24%</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>80.45%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>83.24%</td>
</tr>
</tbody>
</table>

Tables 3 and 4 give an overview of the classification report of the two best models obtained from our dataset.

Table 3. Classification report of Random Forest algorithm.

<table>
<thead>
<tr>
<th>Label</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stay</td>
<td>0.88</td>
<td>0.91</td>
<td>0.89</td>
<td>118</td>
</tr>
<tr>
<td>Leave</td>
<td>0.81</td>
<td>0.75</td>
<td>0.78</td>
<td>61</td>
</tr>
</tbody>
</table>

Table 4. Classification report of Logistic Regression algorithm.

<table>
<thead>
<tr>
<th>Label</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stay</td>
<td>0.91</td>
<td>0.90</td>
<td>0.91</td>
<td>118</td>
</tr>
<tr>
<td>Leave</td>
<td>0.81</td>
<td>0.84</td>
<td>0.82</td>
<td>61</td>
</tr>
</tbody>
</table>

5. Conclusions

The goal of this research was to identify the algorithm that, when applied to the specific dataset that was selected, produces the most accurate forecasts of employee turnover. An open source dataset was used, and a total of six different machine learning algorithms were applied to it. The output that was received was then provided. Based on the results of the analysis, it is possible to deduce that the logistic regression model performed the best on the dataset, followed by the Random Forest algorithm. There are many more characteristics that may be included depending on the requirements of an organization; the attributes that were discussed in the paper are some of the most important causes of employee turnover. The purpose of this study was to compare some of the machine learning models that are the most extensively used so that they can assist different kinds of organizations in maintaining their workforce and reducing the rate of employee turnover.

Author Contributions: All authors have contributed equally in all the section of manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset used in this research is publicly available on Kaggle. With [https://www.ibm.com/communities/analytics/watson-analytics-blog/watson-analytics-use-case-for-hr-retaining-valuable-employees/](https://www.ibm.com/communities/analytics/watson-analytics-blog/watson-analytics-use-case-for-hr-retaining-valuable-employees/).

Conflicts of Interest: The authors declare no conflicts of interest.

References


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.