EMPLOYEE ATTRITION RATE PREDICTION

Medha Gupta*, Nimish Mandowara+, Kavita Namdeo*, Rahul Patel*

*1,2 Student, Department of Computer Science and Engineering, Acropolis Institute of Technology and Research, Indore, Madhya Pradesh, India.

*3,4 Assistant Professor, Department of Computer Science and Engineering, Acropolis Institute of Technology and Research, Indore, Madhya Pradesh, India.

ABSTRACT

Most important characteristics of any organization is productivity and strength, which stands on the legs of its employees. Human Resources are one of the most important assets of any organization. Maintaining regular employees is a great deal for all organizations in the competitive world. It incurs a high cost such as training expenditures and the duration it takes for an employee to become a profitable member and give fruitful results.

The attrition can be defined as the rate of recruitment and termination criteria of the company. Employee Attrition is one of the biggest business problems in Organization. Companies invest a lot in the training of the employees keeping in mind the returns they would get in future. If an employee leaves the company, it is the loss of opportunity and cost to the company. This study interprets the employees’ attrition rate through the related attributes like Job Role, overtime, job level affect the attrition primarily. Hence, organizations attempt to reduce the attrition rate. It is necessary to determine the factors that may affect attrition in the company and thereby, take action upon those factors and reduce the attrition rate.

KEYWORDS: Attrition Rate, Employee Turnover, HR, Classifier, Preprocessing, Employment Features.

I. INTRODUCTION

The outcome of many investigations proves that the most appreciated asset and essential resource in any organization are its employees. Nowadays, the demand for skilled and easy to go employees has increased exponentially and hence proficiency governs the attrition rate. The employee attrition is a serious issue for any organization. The cost of recruiting and training employees is very high. A firm needs to search, hire and train fresher employees. Lack of experienced staff, especially experts, is a tedious job to manage and negatively impact the performance of organizations and deviate them from their goals. The study emphasises on the variables that may lead to control of the attrition rate of the employees.

Employees are considered as the most valuable assets of an organization. But when they quit their jobs unexpectedly, it may incur a considerable cost to any organization. Employee Attrition is known to be a reduction in the workforce in any firm where employees may voluntarily leave the organization or maybe retired. Employee turnover is defined as the number of currently working employees replaced by new employees for a specific period of time. It causes vast expenditure on human resources, including new recruitment, training, and development of the freshly appointed employees, also performance management. Again, attrition which is voluntary is considered unavoidable. Hence, by enhancing employee morale and providing a healthy and competitive working environment, we can certainly reduce this problem significantly.

II. LITERATURE SURVEY

The approaches using various data mining techniques are collected to determine and analyze the employee attrition rate at multiple stages. The study associated with data mining for extracting the employees’ attrition rate used in various models and the comprehensive literature review of numerous researchers’ works are detailed below:

Qasem A, A. Radaideh and Eman A Nagi, have implemented a data mining algorithm to build a classification model to forecast the performance of workforces [5]. They adopted the CRISP-DM data mining methodology [6] in their model. The Decision tree was the prime data mining tool implemented to build the classification

[1009]
model, where numerous classification rules were generated. They corroborated the generated model, and several experiments were conducted using real data collected from several companies. The model is intended to be used for predicting new applicants’ performance.

**Support Vector Machine (SVM):**

A Support Vector Machine is one of the types of supervised learning algorithms that uses the similar beliefs as that of statistical learning models [1] and can solve linear and non-linear binary classification problems. A support vector machine paradigms a hyperplane in higher dimensional space for achieving various class separations. The perception here is that a good separation is accomplished by the hyper-plane that has the significant distance to the nearest training data points of any class the more significant the margin, the lower the generalisation error of the classifier.

**Linear Discriminant Analysis (LDA)**

In this type, the discriminant analysis involves generating one or more discriminant functions to maximize the variance among the groups relative to the difference with the groups. Linear Discriminant Analysis is deriving a variant or z-score, which is a linear grouping of two or more independent parameters that will discriminate best between two (or more) different groups.

### III. EASE OF USE

**A. Abbreviations and Acronyms:**

- HR - Human Resources
- DFD - Data Flow Diagram
- LDA - Linear Discriminant Analysis
- SVM - Support Vector Machine
- UML - Unified Modeling Language

**B. Equations:**

1. **Random Forest Algorithm:**

   The Gini Impurity of a node is a probability that an arbitrarily chosen sample in a node would be incorrectly labelled if it was tagged by the distribution of examples in the node.

   \[ I_G(n) = 1 - \sum_{j=1}^{J} (p_j)^2 \]

2. **Naïve Bayes Algorithm:**

   Bayes theorem delivers a technique of calculating posterior probability \( P(c|x) \) from \( P(c) \), \( P(x) \) and \( P(x|c) \). Look at the equation below:

   \[
P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}
   \]

   \[
P(c \mid X) = P(x_1 \mid e) \times P(x_2 \mid e) \times \cdots \times P(x_n \mid e) \times P(e)
   \]
IV. ALGORITHMS USED

1. Random Forest Algorithm:

Random forest is a kind of supervised learning algorithm implemented for both classifications as well as regression techniques. However, it is primarily used for classification problems. As we know that a forest is made up of a collection of trees and more number of trees means a more dense forest. In the same way, the random forest algorithm generates decision trees on data samples and then gets an estimate from each of them and finally picks the best solution out of them. It is a collaborative method which is better than a single decision tree since it reduces the over-fitting by determining an average of the result.

We can understand the implementation of the Random Forest algorithm with the help of following steps –

- Step 1 – The First Step is to start by selecting random samples from a given dataset.
- Step 2 – Next step is to construct a decision tree for every sample. Then it will predict the result from each decision tree.
- Step 3 – In this step, voting will be performed for every predicted result.
- Step 4 – Finally, select the most voted prediction result as the final prediction result.

2. Naïve Bayes Algorithm:

Naïve Bayes classifiers are one of the collection of classification algorithms based on Bayes’ Theorem. It is a family of algorithms where all of them consists of a common principle, i.e. every pair of features being classified is independent of each other.

We can understand the working of Random Forest algorithm with the help of following steps –

- Step 1: We need to calculate the prior probability for given class labels
- Step 2: Next step is to find Likelihood probability with each attribute for each class
- Step 3: Then put these values in Bayes Formula and calculate posterior probability.
- Step 4: Finally, See which class has a higher probability, given the input belongs to the higher probability class.

V. FIGURES AND DIAGRAMS

1. Data Flow Diagram:

A data flow diagram (DFD) is defined as a graphical illustration of the flow of data through an information system, modeling its process. It shows data is processed by a system in terms of inputs and outputs.

We need to make sure that for each data flow, at least one of the endpoints (source and or destination) must exist in a process. The refined representation of a process can be done in different data-flow diagrams, which subdivides this process into various different sub-processes.

The data-flow diagram is considered as a part of the structured-analysis modeling tools. When using UML, the activity diagram takes over the role of the data-flow diagram.
2. Deployment Diagram:

Deployment diagrams are the diagrams that are used to picturize the topology of the physical components of a system on which the software components are deployed. The deployment diagram for the proposed system is shown below. It shows the physical or the hardware components that are needed to proper function. The physical components include the Server, Client, Windows JVM, and the Database.

3. Hardware Requirements:

The minimum configuration required on computer:

- Processor: 1 gigahertz (GHz) or faster processor.
- RAM: 1 gigabyte (GB) for 32-bit or 2 GB for 64-bit.
- Hard disk space: 16 GB for 32-bit OS 20 GB for 64-bit OS

4. Software Requirements:

The minimum configuration required on computer:

- IDE: Eclipse Luna or above
- Database: MYSQL
- Language: Java
- Platform: Windows 7 Professional or above
VI. CONCLUSION

Human Resource is the main pillar of any organization. The growth level, as well as market penetration, duly depends on the strength of the employees. Now a day due to increased population and people with high competency makes great success for any rm. But the prime issues which are normally addressed in any organization are only the attrition. This is a great challenge as well as retention is also the prime task.

This system can help to implement employee attrition prediction rate in the respective organization. The analysis is done by considering some features like Monthly Income, Last Promotion Year, Current Role in Company, Salary Hike and etc. It understands the key variables that influence the employee attrition rate using data mining. Here we are using a random forest algorithm to build a prediction model for identifying the various reasons for employee turnover.

VII. REFERENCES