



Prediction of Employee Attrition Using Machine Learning Approach

M.B. Shete¹, A. G. Patil²

¹M.Tech.(E&TC) student, Department of E&TC Engineering, PVPIT, Budhgaon, Sangli, India.

²Associate Professor, Head of Department of E&TC Engineering, PVPIT, Budhgaon, Sangli, India.

Email: medhas1993@gmail.com, Email: agpatil112@gmail.com

ABSTRACT

Employees are considered as backbone of an organization. Success or failure of the organization depends on the employees who work for an organization. The organizations have to face the problems when trained, skilled and experienced employees leave the organization for better opportunities and many other reasons. This project aims to predict whether an employee of a company will leave or not. The study was mainly undertaken to identify the dissatisfaction factor of employees and for what reasons they prefer to change their jobs. Once the dissatisfaction factor/s of employees has/have been identified, the organization can take actions accordingly and it may help them to reduce the attrition rate. This project based on build a system which will predict employee attrition based on Employee dataset from GitHub website. Following are the few algorithms that can be used for processing the data, K-Nearest Neighbor, logistic regression, decision Tree, random Forest, Support Vector Machine etc. Based on the accuracy measurement well performed algorithm will be used for this dataset. Which gives accurate solution of the problem statement. This system suggests reasons which optimize the employee attrition in any organization.

Keywords: Employees Attrition, Machine Learning, Support vector machine (SVM), KNN (K-Nearest Neighbor) Introduction:

1. Introduction:

Employee resignations are a reality for any business. However, if the situation isn't handled properly, key staff member's departures can lead to a downturn in productivity. The organization may have to employ new people and train them on the tool that is being used, which is time consuming. Most organizations are interested in knowing which of their employees are at the risk leaving. Employee Attrition is a reduction in manpower in any organization where employees resign. Employees are the valuable assets of any organization. It's necessary to know whether the employees are dissatisfied or are there any other reasons for leaving the respective job. These days for better opportunities, employees are eager to jump from one organization to other. But if they leave jobs unexpectedly, it may cause huge loss for organization. New hiring will consume money and time, also the freshly hired employees take time to make the respective organization profitable. Retention of skilled and hardworking employees is one of the most critical challenges faced by many organizations. Hence, by improving employee satisfaction and providing a desirable working environment, we can certainly reduce this problem significantly. The reasons of leaving the organization could be better-paying job outside, a bad relationship with boss, pursuing higher studies, relocating due to family reasons, fired from organization, job Dissatisfaction, salary not as per expectation, poor relationship with team members, poor working environment, lack of opportunity for career development, overtime, workload etc. In order to tackle this issue, this system developed that uses employee data to analyze reasons for employee attrition. This system is able to predict which employee may leave an organization with what reason, so that they can take several corrective actions in order to ensure that employees stay in the organization and can reduce the attrition. Some of the employee retention strategies to control attrition are motivating employees, expose employees to newer roles, taking constant feedback from employees, etc. Following are the few algorithms that can be used for processing the data Support vector machine (SVM), k-Nearest

Neighbor, Decision Tree, Random Forest, logistic regression etc. Based on the accuracy measurement well performed algorithm will be used for this dataset

1.1 Technologies used in the proposed system

Machine Learning is most important technology towards data analysis for quality prediction and evaluation. There are various algorithms in machine learning which are used to predict the appropriate class of new or unseen data. In our system we used different machine learning algorithms to find out the reasons for employee attrition. The machine learning algorithms which are used in system are described below:

2 SVM (Support Vector Machine)

Support Vector Machine is kind of classification technique. It is a model used for classification and regression problems. It can solve linear and non-linear problems. The idea of SVM is simple: The algorithm creates a line or a hyper plane which separates the data into classes [9]. When unknown data is given as input it predicts which class it belongs to. The margin between the hyper plane and the support vectors are as large as possible to reduce the error in classification

3 K-Nearest Neighbours

K-Nearest Neighbour is considered a lazy learning algorithm that classifies data sets based on their similarity with neighbours. It is one of the most fundamental and simple classification methods and one of the best choices for a classification study of the data [7]. The classification using KNN involve determining neighbouring data points and then deciding the class based on the classes of the neighbours.

4 Decision Tree

As the name implies all decision tree techniques recursively separate observations into branches to construct a tree for the purpose of improving the prediction accuracy. Decision tree is a conventional algorithm used for performing classifications based on the decisions made in one stage. This provides tree structured representation of the decision sets [10]

5 Implemented Algorithm

The proposed system consists of different machine learning algorithms. To build model, we take employee dataset which includes all past and present records of the employees, then we perform data preprocessing (Data Preprocessing is that step in which the data gets transformed, or encoded, to bring it to such a state that the machine can easily analyze it). We have divided dataset into two parts one is train data and second one is test data. Most of the data is used for training and smaller portion of data is used for testing (Train: 70%, Test: 30%). The aim of training is to make a prediction correctly as often as possible. The test data is used to see how well the machine can predict new answers and to validate machine learning model behavior. Afterward, using different machine learning algorithms we have built the model. After building model, user can give the new input data to the system. Furthermore, user can choose algorithm according to their choice and check the result. Output of the system is in two forms - one is graphical representation and other is in polar form that is 'Yes' or 'No' format. After evaluating result the reason behind the attrition is also given by the system

6 Architecture of Algorithm:

An algorithm based on Machine Learning domain is developed as follows which is illustrated in Figure [1]. Machine learning technique is

used to predict the employee turnover and to predict the risk of an employee leaving. Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so

1. Data Collection: The quantity & quality of your data dictate how accurate our model is. The outcome of this step is generally a representation of data which we will use for training. Using pre-collected data, by way of datasets from Kaggle, UCI, etc., still fits into this step
2. Data Preparation: Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.). Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data. Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis. Split into training and evaluation sets.

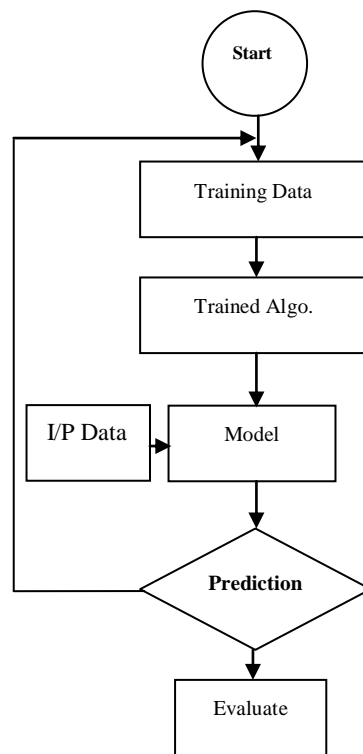


Figure [1] Block Diagram of Algorithm [1]

3. Choose a Model: Different algorithms are for different tasks; choose the right one.
4. Train the Model: The goal of training is to answer a question or make a prediction correctly as often as possible. Linear regression example: algorithm would need to learn values for m (or W) and b (x is input, y is output). Each iteration of process is a training step.
5. Evaluate the Model: Uses some metric or combination of metrics to "measure" objective performance of model. Test the model against previously unseen data. This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not). Good train/eval split? 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc.

7 Experimental Parameters:

In this section, the results of various machine learning approaches are illustrated. All the approaches are evaluated on the precision, recall, accuracy. Various performance parameters are described below (Alduayj & Rajpoot, 2018; Powers, 2011)

1-Accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad [\text{Eq. 1}]$$

2-Precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad [\text{Eq. 2}]$$

3-Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad [\text{Eq. 3}]$$

8 Conclusion:

This paper presented the effect of voluntary attrition on organizations, and why predicting it is important. It further outlined various classification algorithms based on supervised learning to solve the prediction problem

REFERENCE

- [1] Lopes, S. C., Guerra-Arias, M., Buchan, J., Pozo-Martin, F., & Nove, A. (2017). A rapid review of the rate of attrition from the health workforce. *Human Resources for Health*, 15(1). doi:10.1186/s12960-017-0195-2
- [2] Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 41(2), 321-330. doi: 10.1016/j.eswa.2013.07.046
- [3] Cotton J. and Tuttle, J.M., 1986. "Employee turnover: A meta-analysis and review with implications for research" *Academy of management review*, pp.55-70.
- [4] Liu, D., Mitchell, T.R., Lee, T.W., Holtom, B.C. and Hinkin, T.R., 2012. "When employees are out of step with co-workers; How job satisfaction trajectory and dispersion influence individual-and unit-level voluntary turnover". *Academy of management journal*, PP.1360-1380.
- [5] Heckert, T.M. and Farabee, A. M., 2006. "Turn over intentions of the faculty at a teaching-focused university". *Psychological reports*, pp.39-45.
- [6] Rish, Irina. 'An empirical study of the naïve bayes classifier', *IJCAI Workshop on EMPIRICAL METHODS in AI*.
- [7] David a. Freedman, *Statistical Models: Theory and practice*, Cambridge university press p.128
- [8] Rosenblatt, Frank, *Principles of Neurodynamic: Perceptions and the Theory of Brain Mechanisms*.
- [9] Fawcett, T., 2006. An introduction to ROC analysis. *Pattern recognition letters*, pp.861-874