**PREDICTING EMPLOYEE ATTRITION USING MACHINE LEARNING**

**A PROJECT REPORT**

***Submitted by***

**ABINAYA.M 813814104001**

**AISHWARYALAKSHMI.S 813814104003**

**AKSSHAYA.K 813814104005**

**GANESH.V 813814104018**

***in partial fulfillment for the award of the degree***

***of***

**BACHELOR OF ENGINEERING**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

**SARANATHAN COLLEGE OF ENGINEERING, TRICHY**

**ANNA UNIVERSITY: CHENNAI 600 025**

**APRIL 2018**

**ANNA UNIVERSITY: CHENNAI 600 025**

**BONAFIDE CERTIFICATE**

Certified that this project report “**PREDICTING EMPLOYEE ATTRITION USING MACHINE LEARNING**” is the bonafide work of

**ABINAYA.M 813814104001**

**AISHWARYALAKSHMI.S 813814104003**

**AKSSHAYA.K 813814104005**

**GANESH.V 813814104018**

who carried out the project work under my supervision.

**SIGNATURE SIGNATURE**

Dr.S.A.Sahaaya Arul Mary, M.E., Ph.D. Ms. R Thillaikarasi, M.Tech,

**HEAD OF THE DEPARTMENT SUPERVISOR**

**PROFESSOR ASSOCIATE PROFESSOR**

Computer Science and Engineering, Computer Science and Engineering,

Saranathan College of Engineering, Saranathan College of Engineering,

Panjappur, Panjappur,

Tiruchirapalli – 620 012 Tiruchirapalli – 620 012

**VIVA – VOCE EXAMINATION**

**PREDICTING EMPLOYEE ATTRITION USING MACHINE LEARNING**

*Submitted by*

**ABINAYA.M 813814104001**

**AISHWARYALAKSHMI.S 813814104003**

**AKSSHAYA.K 813814104005**

**GANESH.V 813814104018**

The Viva – Voce Examination of this project work done as a part of B.E. Computer Science and Engineering was held on \_\_\_\_\_\_\_\_\_\_\_.

**INTERNAL EXAMINER EXTERNAL EXAMINER**

**ACKNOWLEDGEMENT**

Gratitude is not only the greatest of virtues, but the parent of all others. We take this opportunity to thank all the encouragers and supporters of this project.

First and foremost, we thank the Almighty who has been an unfailing source of strength and comfort for showering his blessings throughout this study.

We sincerely thank our beloved Director **Prof. V.Nagarajan, B.E.,** and **Dr. D.Valavan, M.Tech., Ph.D.,** Principal of Saranathan College of Engineering, for providing all the necessary facilities to do our work.

We would like to extend our deep sense of gratitude to **Dr. S.A.Sahaaya Arul Mary, M.E., Ph.D.,** Head of the department, Computer Science and Engineering, and **Ms. R Thillaikarasi, M.Tech,**Project Coordinator, for supporting us throughout our venture.

We are obliged to our supervisor **Mr. M.Anbazhagan, ME.,**Assistant Professor, for being our internal guide and facilitating us with his valuable support and guidance.

We are thankful to all the teaching and supporting staff members of Computer Science department for the help rendered by them in completion of this project. We are also thankful to our parents and friends who have been encouraging and morally supportive.

**ABSTRACT:**

Bill Gates was once quoted as saying,

“You take away our top 20 employees and we [Microsoft] become a mediocre company”.

His statement cuts to the core of a major problem: employee attrition. It is a major cost to an organization. Some costs are tangible such as training expenses and the time it takes from when an employee starts to when they become a productive member. However, the most important costs are intangible, such as new product ideas, great project management, or customer relationships.

To reduce the cost of attrition, organizations need to ensure that employees’ aspirations are met. Employee attrition control is critical to the long term health and success of any organization. An organization is only as good as its employees, and these people are the true source of its competitive advantage.

Accurate predictions enable organizations to take action for the retention of employees. This project aims to use different supervised classifiers to make predictions, and chooses the most accurate one.

1. **INTRODUCTION**
   1. **EMPLOYEE ATTRITION - AN OVERVIEW**

Human resource is the most important asset for a company to be competitive. Thanks to liberalization on the labor market, it has become possible for an employee to leave his job. However, having excess employees leave their jobs will influence the morale of the companies.

The loss of good employees can diminish a company’s competitive advantage and furthermore lead to a reduction in output and quality. High employee attrition has a significant negative effect on an organization by virtue of lost productivity, increased training and recruitment costs.

Employees voluntarily leave an organization for various reasons, such as new opportunities, limited or no professional growth in current position, unhappiness with compensation, personal reasons, etc. By taking proactive action to retain its top employees, a company can thus reap substantial benefits, thereby increasing its top and bottom line.

* 1. **METHODS**

This project discusses different classification algorithms of supervised learning and feature selection algorithms. This section gives a summary of each of these machine learning algorithms.

* + 1. **Logistic Regression:**

Logistic regression is a statistical method for evaluating a dataset which consists of one or more independent variables that determine an outcome. The outcome is measured with a variable which takes on one of only two possible outcomes. The goal of logistic regression is to find the best fitting model that describes the relationship between a set of independent (predictor or explanatory) variables and the dichotomous characteristic of interest (dependent variable = response or outcome variable). Logistic regression generates the coefficients of a formula that predicts a logit transformation of the probability of the presence of the characteristic of interest. The logit transformation is defined as the logged odds:

=

where p is the probability of presence of the characteristic of interest.

And logit(p) =

* + 1. **k-Nearest Neighbors:**

K-nearest neighbors algorithm is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) method used for [classification](https://en.wikipedia.org/wiki/Statistical_classification), where the input consists of the k closest training examples in the [feature space](https://en.wikipedia.org/wiki/Feature_space) and the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors measured by a distance function (k is a positive [integer](https://en.wikipedia.org/wiki/Integer), typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. The distance function used is Hamming distance, given by:

DH =

kNN is a type of [instance-based learning](https://en.wikipedia.org/wiki/Instance-based_learning), or [lazy learning](https://en.wikipedia.org/wiki/Lazy_learning), where the function is only approximated locally and all computation is deferred until classification. A useful technique is used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

* + 1. **Random Forest:**

Random forest is an ensemble learning method that gives an improved performance using divide and conquer technique.  The main principle behind ensemble methods is that a group of “weak learners” can come together to form a “strong learner”. As the name suggest, this algorithm creates the forest with a number of trees. In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results. It operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes of the individual trees. When the training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This [oob (out-of-bag) data](https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#ooberr) is used to get a running unbiased estimate of the classification error as trees are added to the forest.

After each tree is built, all of the data are run down the tree, and [proximities](https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#prox) are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalized by dividing by the number of trees. Proximities are used in replacing missing data, locating outliers, and producing illuminating low-dimensional views of the data.

* + 1. **SelectKBest:**

SelectKBest is a feature selection algorithm that scores the features of a dataset using a score function and then removes all but the k-highest scoring features. It takes as a parameter the score function, which must be applicable to a pair of data from the training set (X) and test set (y). The score function returns an array of scores, and SelectKBest simply retains the first k features of training set with the highest scores.

The score function used is Chi-Square (chi2). It measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent. The test compares the observed data to a model that distributes the data according to the expectation that the variables are independent.

SelectKBest computes the chi2 statistic between each feature of X and y (assumed to be class labels). A small value will mean the feature is independent of y. A large value will mean the feature is non-randomly related to y, and so likely to provide important information. The formula for the chi2 statistic is given by:

χ2 = =

where Oi is the observed value and Ei is the expected value.

* + 1. **Recursive Feature Elimination:**

Using an external estimator that assigns weights to features (e.g., the coefficients of a linear model), Recursive Feature Elimination selects features by considering smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a coef\_ attribute or through a feature\_importances\_ attribute of the estimator. Then the least important features are pruned from current set of features. Until the desired number of features is selected, this procedure is recursively repeated on the pruned set. The estimator used here is the linear model of Logistic Regression.

|  |  |
| --- | --- |
| **SelectKBest** | **RFE** |
| Age | Department |
| Daily Rate | Environment Satisfaction |
| Distance From Home | Gender |
| Monthly Income | Job Involvement |
| Monthly Rate | Job Level |
| OverTime | Job Satisfaction |
| Total Working Years | Marital Status |
| Years At Company | Overtime |
| Years In Current Role | Stock Option Level |
| Years With Current Manager | Work Life Balance |

*Comparison of features selected by SelectKBest and Recursive Feature Elimination (RFE)*

* + 1. **XGBoost:**

XGBoost stands for eXtreme Gradient Boosting. This algorithm goes by lots of different names such as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines. Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. XGBoost is an implementation of gradient boosting machines, engineered for efficiency of compute time and memory resources.

Every parameter in XGBoost has a significant role to play in the model's performance. Its parameters can be divided into three categories:

1.General Parameters - Controls the booster type in the model which eventually drives overall functioning

2.Booster Parameters - Controls the performance of the selected booster

3.Learning Task Parameters - Sets and evaluates the learning process of the booster from the given data

1. **PROPOSED SYSTEM**

A fictional dataset created by IBM data scientists is used for analysis. It has 35 features and 1470 observations. There are two class labels for the feature “Attrition” – Yes and No. The dataset includes various important features such as Age, DailyRate, DistanceFromHome, Overtime, EnvironmentSatisfaction, JobLevel, JobSatisfaction, MonthlyIncome, WorkLifeBalance, etc. There are 34 independent features and 1 dependent feature (Attrition). Out of the thirty four features, six are categorical and the remaining are numeric.

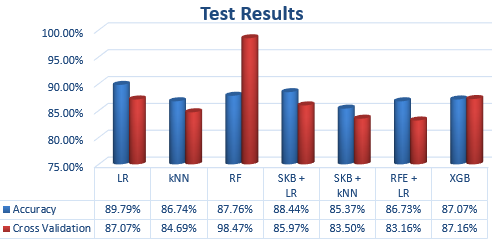
* 1. **DATA PREPROCESSING**

The missing data from the dataset is handled by Interpolation, which is a mathematical technique to estimate the missing values in some interval, when a number of observed values are available within that interval. By default, linear interpolation is performed at the missing data points.

All the categorical values in each column are converted to numerical values using LabelEncoder. It is used to assign ordinal levels to categorical data. It encodes the labels with values between 0 and (number of classes - 1). For example, the feature MaritalStatus with labels “Divorced”, “Married” and “Single” is encoded as 0, 1 and 2 respectively.

* 1. **CLASSIFICATION**

The dataset is split into training and test data in the ratio 80:20. Different classification algorithms [as mentioned in Section 3] are used to train the training set. Furthermore, feature selection algorithms are also combined with classification algorithms to train the training set. The trained model is used to make predictions on the test data, and is also applied on the training data for cross validation. The model with the highest accuracy is used to make the final prediction.



**SOFTWARE REQUIREMENTS SPECIFICATION**

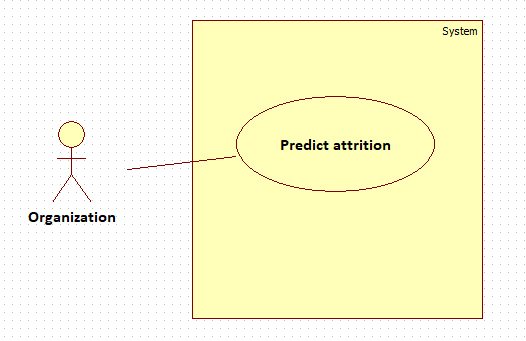
* 1. **INTRODUCTION**
     1. **Purpose:**

Employee turnover reflects an organization’s internal strengths and weaknesses. Organizations face difficulties in retaining the employees as well as attracting potential employees. All this has a significant impact on the strength of a company in managing their business in a competitive environment. Hence, it has become critical for the companies to satisfy their employees in order to retain them.

* + 1. **Project Scope:**

This project can help an organization identify the employees who are vulnerable to quitting their jobs.

* 1. **SYSTEM FEATURES**
     1. **Use Case:**

****

**Name:** Predict attrition

**Input:** Employee dataset

**Output:** Attrition rate

**Pre-condition:** An employee dataset with necessary attributes

**Steps:**

1. Preprocessing:

The missing data in the dataset is fixed.

The non-numeric attributes are encoded to contain numeric values.

1. Classification:

Various classification algorithms are used to find the number of employees who would leave an organization.

**Post-condition:** The employee attrition rate for the given dataset is displayed

1. **LITERATURE SURVEY**

Rohit Punnoose et al [1] explored the application of Extreme Gradient Boosting (XGBoost) technique which is more robust because of its regularization formulation. Data from the HRIS of a global retailer was used to compare XGBoost against six historically used supervised classifiers and its significantly higher accuracy for predicting employee turnover was demonstrated.

Vidya Sunil et al [2] conducted a study to find the main causes behind the increase in employee attrition in software industries and to find out the ways to control attrition. The study was carried out in software companies in Pune. The survey of 100 employees revealed that those having average age of 24-28 years and experience between 2-4 years have a higher percentage of attrition. The attrition rate increased because of dissatisfaction with pay, relationship with boss, lack of career advancement and compensation.

Sunil Kumar Dhal [3] et al conducted a study to find out the main causes which increase the employee turnover in BPO companies and to find out the way to control attrition. This study was conducted in BPO companies at Bhubaneswar.

Moninder Singh et al [4] describe a framework for using analytics to proactively tackle voluntary attrition of employees. This approach uses data mining for identifying employees at risk of attrition and balances the cost of attrition/replacement of an employee against the cost of retaining that employee (by way of increased salary) to enable the optimal use of limited funds that may be available for this purpose, thereby allowing the action to be targeted towards employees with the highest potential returns on investment.

Santoshi Sengupta [5] presented an approach to determine what and how job-related and demographic variables are associated with employee satisfaction of the BPO employees.

Ankita Srivastava et al [6] proposed a model to identify the root causes of attrition and retention in BPOs, analyzing the level of employee motivation, satisfaction and involvement, generate a model for maximizing sustenance of employees in the organization and come up with concrete recommendations.

Vijay Anand et al [7] carried out a research in BPO companies in which the opinions of 120 respondents (both ex-employee and current employee) were taken. A structured questionnaire was used for collecting data, and Percentage analysis, Weighted average method, Chi-square Test and ANOVA have been incorporated for analysis purpose.

Hsin-Yun Chang [8] proposed a method that could select subsets more efficiently. In addition, the reasons why employers voluntarily turnover were also investigated in order to increase the classification accuracy and to help managers to prevent employers’ turnover. The mixed feature subset selection used in this study combined Taguchi method and Nearest Neighbor Classification Rules to select feature subset and analyze the factors to find the best predictor of employer turnover.

Neeraj Pandey et al [9] presented an approach to explore the factors behind the high attrition in Indian ITeS call centres. A focussed group discussion (FGD) was conducted with a group to discuss the variables for attrition. Semi-structured interviews were conducted to validate the responses received during FGD. The key questions asked during the interview explored the reasons for joining and also reasons behind leaving the ITeS call centre jobs.

Rahul Yedida et al [10] discussed the application of the k-Nearest Neighbours (KNN) algorithm as a method of predicting employee attrition, with evaluation of employee performance, average monthly hours at work and number of years spent in the company, etc as features. Other approaches include the use of ANNs, decision trees and logistic regression. The conclusion is reached by comparing the performance of the KNN classifier against other techniques.

Rupesh Khare et al [11] presented the application of logistic regression technique to predict employee attrition risk in an organization based on demographic data of separated employees. The demographic information of both separated and existing employees was used to develop a risk equation, which was later applied to assess attrition risk with current set of employees. Post this assessment, high risk cluster was identified and focus group discussions were initiated to find out the reasons and their requirements and hence action plan was created to minimize the risk.