IMPROVEMENT IN EMPLOYEE RETENTION THROUGH MACHINE LEARNING TECHNIQUES

Prateek Jaiswal[I] **and Deepankur Kansal**[I]

I. ABSTRACT

Employee attrition is one of the major problems faced by many companies,start-ups nowadays. As loss of an employee incurs huge cost to the company in terms of lost productivity,recruitment and training costs. The biggest challenge faced by many organisations today is how to retain their employees.This research paper aims to provide the solutions for the retention of an employee or to increase the tenure of that employee. Firstly,we will predict the expected tenure of the currently working employees in the company by analysing the employee attrition data , predicting the value of the employee to the company and then so as to retain the employees ,we will update each feature and generate new value of the employees, subject to the constraint of the resources.Finally, we will predict the updated tenure of the employee corresponding to the newly generated value.

II. INTRODUCTION

Employees are the backbone of an organization. Hence, the retention of the employees is important in keeping the organization on track. In order to retain the best talents, strategies aimed at satisfying employee's needs are implemented, regardless of global companies or small-sized firms[1].

Employee attrition incurs huge cost in terms of lost productivity,recruitment and training costs.Between costs associated with separation, loss of productivity, recruitment, interviewing, training, and on boarding, the loss of a single employee is estimated to cost the company one third of that individual's annual salary[2].There could be various factors responsible for employee leaving the company.The major challenge faced by many organisations is to retain their employees.In this research paper,we will create models that predict the expected tenure , value of an employee to the company.We will treat each employee as an individual entity and suggest preventive attrition measures at an individual level.we will provide the appropriate measures to be taken to Retain the employee at the predicted tenure. Using these proposed solutions,we will predict the new expected tenure of the employee. So,In this way we will solve the problem of employee retention.

III. Literature Review

In Lucas (2013),It is reported that employers don't understand the expense of high employee turnover. Recruiting new staff is costly due to advertising and administrative expenses, time and resources for on boarding and training as well as loss of productivity.[3]

Omer and Laura (2015) said successful employee retention is essential to an organization's stability, growth and revenue. Organizations can achieve employee retention by developing four strategies.[4]

In Regresion Analysis (2012), an effective human resource management practices namely employee empowerment, training and development, appraisal system compensation are the main factor for the success of a firm on employee retention. By using a multiple regression analysis, training and development, appraisal system compensation are significant to employee retention.[1]

In the book, "Factors Affecting Employee Attrition and Predictive Modelling Using IBM HR Data", establishing a predictive model for employee attrition involves data preprocessing with chi square versus logistic regression for feature selection, machine learning models and their comparison using the confusion matrix, precision, recall and f1-scores based on IBM HR Data Set.[6]

Analysis of RF, ANN, and SVM Regression Models are compared by Beijing Research Center accounting for the application of various regressive models for a specific use case of estimation.(2017) [5]

IV. Data

The Data taken is developed IBM R and D lab to provide a standard for the HR management issues incorporated by a medium scale enterprise. It is developed to uncover the factors that lead to employee attrition and explore important questions and ideas leading to a fruitful outcome containing data features pertaining to every aspect of a employee concerning his/her professional life.

The Data itself is complete and is divided into two parts separating attrition and retention subsections of the employee. The initial training is done on attrition data set in the ratio of 80 percent train data and 20 percent test data and predicts values for the retention data set.

V. Methodology

A. *Human Resource Management Models*

A1. Remaining Tenure Prediction Model

A predictive regression model is trained by ensemble learning by combining different regressive models (B1-B6) and generating the smallest mode range for the specific employee using the employee attrition data set(from above) and then the expected tenure of the employees currently working in the company is predicted using the regression model on the retention data set along with acting as the test condition for the model. As, the number of attrition-ed employees will always be less than currently working employees in a specific period of time therefore the test condition is that the Remaining predicted mode from the model for the currently working employees should be greater than 0 for the validation of the model.

$TenurePred(E_i) > 0.$

Let us assume x_1,....,x_n be the features of the employees from the data set, using which we are predicting the expected tenure(let us assume it to be $x_n + 1$) and E_i represents the i^{th} employee.

A2. Employee Value Prediction Model

The employees are clustered into many clusters using the K Means algorithm, and each cluster is assigned a range mid-point of class size 5 denoting the value of the employee to the company and therefore defining the value of k to be 20. For clustering,the value is generated considering the weighted n+1 features: where w_1,.....,w_n,w_{n+1} are the weights corresponding to each of the n+1 features x_1,....,x_n,x_{n+1} respectively and the n attributes(x_1,.....,x_n) are from the attrition-ed employees data set and the predicted tenure forms the $(n + 1)^{th}$ attribute used for the value-generationn in the employee value prediction model.

A3. Resource consumption Model

This model evaluates a function $A3_i$ depending on all the attributes of an employee $x_1...x_n$ in a weighted manner representing the resources being used according to the current attributes.

A4. Retention Technique Selection Model

The above calculated value is considered and for the employees with more than 50 percent importance value will be considered for this model. This model in turn carries various Retention Techniques which are unique for each x_i, updates the features x and re-

evaluates the cluster in the model $A2$ for $A2_i$ generating a different importance value. Also the function $A3_i$ has to be minimized while updating and recurring for the suitable solution. This continues till all the combination of the techniques is exhausted and generates the best combination of technique possible for the each employee $A2_i$ being considered along with the most optimal Resource possible. Therefore the argument is to $max(A2_i/A3_i)$ where the Res_i represents the Resource consumption an employee i. This can be written as:-

$$max(A2_i/A3_i) = max(\frac{A2_i|x_1...x_{n+1}}{A3_i|x_1...x_n})$$

$$= max((\frac{A2_i|x_1...x_n)*(x_{n+1}|x_1....x_n)}{A2_i|x_1...x_n})$$

$$= (x_{n+1}|x_1....x_n)) * max(\frac{A2_i|x_1...x_n}{A3_i|x_1...x_n})$$

as x_{n+1} represents the predicted tenure which is considered to be constant for the whole iterative process considered and therefore can be taken out.

$$= (A1) * max(\frac{A2_i}{A3_i})$$

A5. Updated Tenure Prediction Model

After calculating the optimal solution for the attributes of the employee we generate the updated predicted tenure according to our Model A1 i.e. x_0. Now, the difference between $x_0 - x_{n+1}$ is calculated and represents the Final Result for our Algorithm along with the Resource consumption value and the updated attributes which can be used to reflect the actual retention strategies used for each employee.

B. Machine Learning Algorithms

B1. Support Vector Regression Algorithm[7]

As a supervised-learning approach, SVR trains using a symmetrical loss function, which equally penalizes high and low misestimates.One of the main advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space.

In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM in incorporated.
The expected values are:-

$$y = \sum_{i=1}^{N}(a_i - a_j).K(x_i, x) + b$$

The kernel function is:-

$$k(x_i, x_j) = exp(-\frac{|x_i - x_j|^2}{2\sigma^2})$$

B2. Random Forest Algorithm[8]

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$ bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples.After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$f = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

B3. Principal Component Regression [12]
The PCR method may be broadly divided into three major steps:

1. Perform PCA on the observed data matrix for the explanatory variables to obtain the principal components, and then (usually) select a subset, based on some appropriate criteria, of the principal components so obtained for further use.

2. Now regress the observed vector of outcomes on the selected principal components as covariates, using ordinary least squares regression (linear regression) to get a vector of estimated regression coefficients (with dimension equal to the number of selected principal components).

3. Now transform this vector back to the scale of the actual covariates, using the selected PCA loadings (the eigenvectors corresponding to the selected principal components) to get the final PCR estimator (with dimension equal to the total number of covariates) for estimating the regression

coefficients characterizing the original model.

B4. Lasso Regression Algorithm[11]

Under lasso, the loss is defined as:-

$Lasso(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{m} |\beta_j|$

B5. Cox Regression Algorithm[9] The hazard function for the Cox proportional hazards model has the form

$\lambda(t|X_i) = \lambda_0(t)exp(\beta_1 X_{i1} + ...\beta_p X_{ip}) = \lambda_0(t)exp(X_i\beta)$

B6. Elastic Net Regression Algorithm[10]

Elastic Net aims at minimizing the following loss function:-

$L_{enet}(\beta) = \frac{\sum_{i=1}^{n}(y_i - x_i\beta)^2}{2n} + \lambda(\frac{1-\alpha}{2}\sum_{j=1} m\beta^2 + \alpha\sum_{j=1}^{m}|\beta_i|)$ where α is the mixing parameter.

B7. K-Means Clustering[13]

Given a set of observations $(x_1, x_2, ..., x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into $k \leq n)$ sets $S = S_1, S_2, ..., S_k$ so as to find:-

$arg_S \quad min \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{x,y \in S_i} ||x - y||^2$

VI. Results

At the end of evaluation we expect that we can get a significant increase in the predicted tenure of a substantial percentage of employees by utilising minimum amount of extra resources to be given in accordance of the predicted retention strategy.

VII. Conclusion

Therefore, we would show that in enterprises our predicted model will provide its extended help to the Human Resource Management Department in taking better and determined decisions towards a specific employee and therefore benefit the company by a significant margin by retaining the existing hired workforce and reducing a large amount of resources to be spent on scouting, hiring and training.

References

[1] Eric Ng Chee Hong ,Lam Zheng Hao,Ramesh Kumar,Charles Ramendran ,Vimala Kadiresan.An Effectiveness of Human Resource Management Practices on Employee Retention in Institute of Higher learning: - A Regression Analysis. International Journal of Business Research and Management (IJBRM), Volume (3) : Issue (2) : 2012.

[2] Matthew O'Connell and Mavis (Mei-Chuan) Kung. "The Cost of Employee Turnover." In: Industrial Management (2007), pp. 14–19.

[3] Lucas, S. (2013). How much employee turnover really cost you. Retrieved from Inc.:

[4] Omer Cloutier,Laura Felusiak,Calvin Hill,Enda Jean Pemberton-Jones.The Importance of Developing Strategies for Employee Retention.Journal of Leadership, Accountability and Ethics Vol. 12(2) 2015.

[5] Yuan, H.; Yang, G.; Li, C.; Wang, Y.; Liu, J.; Yu, H.; Feng, H.; Xu, B.; Zhao, X.; Yang, X. Retrieving Soybean Leaf Area Index from Unmanned Aerial Vehicle Hyperspectral Remote Sensing: Analysis of RF, ANN, and SVM Regression Models. Remote Sens. 2017, 9, 309.

[6] Khan, Emad Afaq; Hayat Khan, Sumaira Muhammad.Factors Affecting Employee Attrition and Predictive Modelling Using IBM HR Data. Journal of Computational and Theoretical Nanoscience, Volume 16, Number 8, August 2019, pp. 3379-3383(5).

[7] Awad M., Khanna R. (2015) Support Vector Regression. In: Efficient Learning Machines. Apress, Berkeley, CA

[8] Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[9] Cox, D. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2), 187-220.

[10] Hui Zou and Trevor Hastie(2005). Regularization and variable selection via the Elastic Net.Journal of the Royal Statistical Society, Series B. 301-320

[11] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267-288

[12] Jolliffe, I. (1982). A Note on the Use of Principal Components in Regression. Journal of the Royal Statistical Society. Series C (Applied Statistics), 31(3), 300-303.

[13] Forgy, Edward W. (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics. 768-769.