

Employee Attrition Model Using Machine Learning Approach

Avipra Amrute¹, Manasi Vartak², Anuja Gurav³

¹ Student, Computer Engineering, Pillai HOC College of Engineering and Technology, Maharashtra, India

² Student, Computer Engineering, Pillai HOC College of Engineering and Technology, Maharashtra, India

³ Student, Computer Engineering, Pillai HOC College of Engineering and Technology, Maharashtra, India

ABSTRACT

Employee attrition is a problem faced by many organizations, with many experienced, technical and valuable employees leaving the organization every day. The achievements and shortcomings of the organization relies upon employees who work for the business enterprise. When a productive employee leaves the organization, it results in huge financial loss and to recruit new employee's organization has to spend a lot of time. This leaves difficulty for Human Resource Manager (HRM) to choose the right and talented employee which suits for that particular job role. The primary concept of our model is to attain the effectiveness of the evaluation and the employee satisfaction within the enterprise, which can help lessen worker turnover. So, in this model various machine learning techniques such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Classifier (SVC) applied on data. This model will help organizations to know hidden reasons for employee attrition and accordingly management can take preventive measures which leads to economic growth.

Keyword :- attrition, organization, machine learning techniques, employee, economic growth

1. INTRODUCTION

In the business world, employee attrition is a very costly problem. It is a natural part of an organization's life cycle. A retention rate is the yearly or monthly loss in recurring revenue over a particular period of time. Attrition is often calculated on a monthly or annual basis. Research on data science and machine learning has begun being applied to this area so that employers can understand why and when employees are most likely to leave the company, allowing them to take actions to improve retention. This also allows employers to prepare in advance for new hiring opportunities. When skilled, talented and experienced employees leave organizations for better opportunities, they create a number of problems for the organization. The HR department needs to start thinking outside the box and out of its traditional ways and methods in order to effectively address HR challenges and issues going forward. Using People Analytics on a personal level, we like to think of it as when machine learning is applied to HR data. Through this analysis, we find out what type of employees leave their jobs, and what different parameters they possess, such as high salaries or low salaries.

Companies while conducting "Exit Interviews" try to identify trends in behavior of employee who is leaving and the reasons behind the same. Using that knowledge, the organization sets up retention strategies in order to make a positive impact on retention or at least have a positive impact on employee opinions about the organization. Most of the times, exit interviews ask what the employee thought of working for the organization, were they happy with the workplace environment, what is the reason for leaving job etc We should be seeking feedback from employees in an

active way, as an exit interview is usually too late. Employees may also leave the job for a variety of undisclosed reasons, such as unemployment, a lack of growth, the desire for a change, problems with senior management, etc.

Study in predicting employee attrition uses some of the classification models shown in the paper. The performance of models will depend on data analysis and transformation.

2. LITERATURE SURVEY

1. Adarsh Patel, Nidhi Pardeshi, Shreya Patil, Sayali Sutar, Rajashri Sadafule, Suhasini Bhat, [1] this paper resulted that here were a great deal of workers who aren't happy because of one in every of the aspects that isn't happy by the corporate which ends in high rate

2. Francesca Fallucchi Marco, Coladangelo Marco, Coladangelo Ernesto, William De Luca [2] here the results are presented using classical metrics, and the algorithm that produced the best results for the available datasets is the "Gauss Naive Bayes Classifier".

3. Namrta Bhatia, Sheetal Jannu, Purvika Shukla, Radhika Chapneri[3] this shows when feature selection for prediction is performed correctly, the accuracy rate is always better, in spite of which classification model is used, compared to classification without feature reduction.

4. Madara Pratt, Mohcine Boudhane, Sarma Cakula [4].It resulted that the proposed version is the quality set of rules for facts set describing worker pleasure and attrition with an precision of 85.12%.

5. Aseel Qutub, Asmaa Al-Mehmadi, Munirah Al-Hssan, Ruyan Aljohani, and Hanan S. Alghamdi [5]. The outcome display the lead of the linear version in phrases of precision, take into account and AUC.

3. DATASET

The dataset contains up to 14000 observations for employees. Here. the parameters used are Age, Business travel, Daily wage, Distance travel, Department in which employees working , Education qualification, Relationship Satisfaction of employees with their higher authority, Education field, Stock option level, Total working years at company, training time last year, Work life balance, Year in current role, Years at company, Years since last promotion, years with current manager. The target variable is 'Attrition' which means current status of employee.

4. METHODOLOGY

The steps include:

- 1.Data preprocessing
- 2.Data analysis
- 3.Model training
- 4.Model Validation
- 5.Model Predictions

1)Data preprocessing:

It's the first and most important stage in building a machine learning model. Because we can't give the Model raw data, it's critical to clean the data beforehand.

2)Data analysis:

A valuable insight can be derived from the results by manipulating, transforming, and visualizing the data. It takes 70% of the data for analysis.

3)Model training:

In order to train a machine learning algorithm (that is, a learning algorithm), training data must be provided. It is the model artifact created as a result of the training process that is called as an ML model. Training data must include the right response, also known as a target or target attribute.

4)Model Validation:

Validate the model against 30% dataset. If the model is validated then the model is implemented otherwise, we have to rebuild the model.

5)Model Predictions

Now the model is trained and tested, the model will predict the results.



Fig.4.1 Methodology

5.DATA MODEL

5.1.SVM

A support vector machine (SVM) is a supervised machine learning technique that can be used to deal with classification and regression problems. It is, however, mostly employed to solve categorization difficulties.

In this algorithm, every statistics object is represented as a factor in n-dimensional space (wherein n is the range of features). Each function has a price that corresponds to one of the co-ordinates. Then, with the aid of using finding the hyper-plane that high-quality distinguishes the 2 classes, class changed into accomplished.

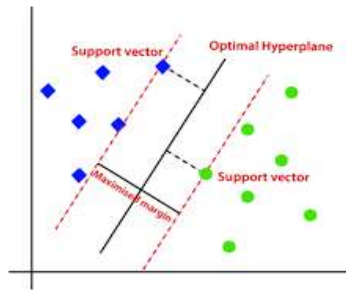


Fig.5.1 SVM

5.2 Decision Tree

The root node of the Decision Tree algorithm is the original set S . As the algorithm iterates through each unused property in S , it measures the entropy $H(S)$ (or information gain $IG(S)$) of that property.

When choosing a characteristic, the one with the lowest entropy (or greatest information gain) value is chosen. The data set S is used to select subsets based on several attributes (e.g., Age less than 50, Age between 50 and 100, and Age greater than 100).

All subset undergoes the same procedure, taking into consideration only traits that have never been selected before.

5.3 Random Forest

Random Forests produces a huge number of categorization trees. Use the input vector to categorize one new object from an input vector by moving the vector down individual tree in the forest. Each tree is assigned a categorization, and this is called its "vote." The classification that receives the highest votes is selected (over all the trees in the forest) by the forest as a whole.

1. Approximately $1/3^{\text{rd}}$ of the instances are excluded from the pattern for the contemporary tree while the education set is drawn the usage of sampling with substitute. When a tree is delivered to the forest, this OOB (out-of-bag) information is used to get a continuous, independent estimate of the class error. It is likewise used to calculate an estimate of the significance of a variable..

2. After every tree is constructed, all the information is administered down the tree, and proximities for every pair of instances are computed. When instances proportion a terminal node, their distance is extended with the aid of using one. The proximities are normalised at the belief of the run with the aid of using dividing with the aid of using the range of trees. Proximities are used to update lacking information, discover outliers, and create enlightening low-dimensional information displays..

5.4 Logistic regression

Under the supervised learning approach, one among the foremost vital Machine Learning algorithms is logistic regression. It's a way for predicting a categorical variable from a group of freelance variables.

A specific structured variable's output is anticipated the usage of logistic regression. As a end result, the end result should be a discrete or specific value. It may be yes or no, 0 or 1, true or false, and so on, in this model it will give result as yes for employee who is leaving company and no for employee who is not leaving the company.

Table 1. Comparison of classification of models

Models	Accuracy	
	Train	CV
1.Logistic Regression	89.89	87.66
2.Random Forest Classifier	100	85.33
3.Decision Tree	100	78.62
4.SVC	87.76	86.1

6.CONCLUSION

By using different classification algorithms, it will be helpful to find out what type of people leave their job and what different parameters are affecting for employee attrition after training and testing the models. This project will aid in the development of a reliable model for predicting employee attrition and retention methods. Decision tree and Random Forest algorithm gives 100% accuracy.

7.ACKNOWLEDGEMENT

We take great pleasure in expressing our heartfelt gratitude and genuine thanks to all members of our department's staff for their continuous guidance and advice throughout the project work, We have no words to express our profound gratitude for all of our department's teaching and non-teaching staff members, as well as our parents and friends, who assisted us in completing the project successfully.

8. REFERENCES

- [1] Adarsh Patel, Nidhi Pardeshi, Shreya Patil, Sayali Sutar, Rajashri Sadafule, Suhasini Bhat, "Employee Attrition Predictive Model Using Machine Learning", (IRJET) e-ISSN: 2395-0056 Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072
- [2] Francesca Fallucchi Marco, Coladangelo Marco, Coladangelo Ernesto, William De Luca, "Predicting Employee Attrition Using Machine Learning Techniques", October 2020; Accepted: 29 October 2020; Published: 3 November 2020.

[3] Namrta Bhatia, Sheetal Jannu, Purvika Shukla, Radhika Chapneri,” Employee Attrition Prediction Using Classification Model”, 2019 5th International Conference for Convergence in Technology (I2CT) Pune, India. Mar 29, 2019

[4] Madara Pratt, Mohcine Boudhane, Sarma Cakula,” Employee Attrition Estimation Using Random Forest Algorithm”, Baltic J. Modern Computing, Vol. 9 (2021), No. 1, 49-66

[5] Moninder Singh, Kush R. Varshney, Jun Wang and Aleksandra Mojsilovic, “An Analytics Approach for Proactively Combating Voluntary Attrition of Employees”, IEEE, 2019.

[6] Aseel Qutub, Asmaa Al-Mehmadi, Munirah Al-Hssan, Ruyan Aljohani, and Hanan S. Alghamdi, “Prediction of Employee Attrition Using Machine Learning and Ensemble Methods”, International Journal of Machine Learning and Computing, Vol. 11, No. 2, March 2021.

