PREDICT EMPLOYEE CHURN RATE USING MACHINE LEARNING TECHNIQUES

Poorva Agrawal[1]*, Seema Ghangale[2]

Abstract:
Employee Churn is a common occurrence in the modern business industry. An employee might unexpectedly leave a company or may be let go. This turnover of employees creates a lot of problems as a company invests a lot of valuable time and resources in preparing the employee for working in the company and then monitoring their performance. Employee Attrition also creates an opportunity for the business rivals as they can now hire the rejected employees who know secrets about their companies. This paper focuses on the research on this field and how supervised machine learning techniques have been used to predict the churn rate with maximum accuracy and which method is most suitable for the accurate prediction by comparing their analysis. The prediction of churn rate is based on various employee characteristics which include both technical and personal factors. We have further to tried to use the results of the predictions to come up with techniques on how to reduce the attrition of employees in the future.

Index Terms: Machine Learning, Artificial Intelligence, Support Vector Machine, Employee Attrition, Random Forest, Turnover, K-Nearest Neighbours

[1]*Symbiosis Institute of Technology, Symbiosis International (Deemed) University, Nagpur, India
Email: poorva.agrawal@sitnagpur.siu.edu.in
[2]Symbiosis Institute of Operations Management, Symbiosis International (Deemed) University, Nashik, India,
Email: seema.ghangale@siom.in

***Corresponding Author:** Poorva Agrawal
*Symbiosis Institute of Technology, Symbiosis International (Deemed) University, Nagpur, India,
Email: poorva.agrawal@sitnagpur.siu.edu.in

DOI: - 10.31838/ecb/2023.12.si5.0132

I. INTRODUCTION

Employees are the most important part of a company. Bill Gates once said "Take away our top 10 employees and overnight we become a mediocre company". Retention of employees is integral to the longevity of the company lifetime. The success of the company is reflected in the performance of the employees. Employees contribute to the quantity and quality of the company. Employee Attrition is easy to calculate for any given time period. The analyst needs to know the number of employees who have left the company and the average staff size for the same period of time. Take the number of employees who left the company and divide that by average size of the staff. We finally take that number and multiply it by 100 to get the attrition rate. The best way to describe the employee attrition and turnover rate is to call them Employee Churn. Even though Turnover and Attrition are used synonymously, they in fact mean two different types of employee churn.

Employee Attrition is normally voluntary like retirement or resignation. In contrast turnover is used to include both voluntary and involuntary departures. While turnover includes employees who leave off their own interests it also includes employees who are terminated or fired involuntary. In attrition, the HR is generally tasked with not replacing the empty positions while in turnover, it is the duty of the HR to replace the employees who have left the company.

Just as the explanations for attrition and turnover are different, the cost required for dealing with them are also different too. The method of decreasing labour costs is often seen as a part of attrition. The most severe part of attrition is layoff which is often deployed as a reduction technique for labour. In case of the natural leave of the employee, the empty spot need not be filled any longer and thus organisations can freeze hiring to reduce staffing costs. However, the two types of churn do not overlap with each other. The cost of fixing attrition can be as much as turnover. It is not cheap to fill the position required for replacing a voluntary resignation. The expenses in organisation add up drastically when the turnover or attrition rate add up drastically.

Computer Science has led to robust quantitative method for obtaining insight from statistics due to the prevalence of intelligent machine learning algorithms. The machine learning method used for the employee churn prediction model is Supervised Machine Learning. Supervised Learning generally involves the presence of a supervisor who acts as a teacher. A proper definition would be that supervised learning is a learning in which we train the model using labelled data which means some of the data is already tagged with the correct answer. After the learning, the model is provided a new set of data so that the learning algorithm observes the set of training examples and provides a correct outcome from labelled set.

As the information technology advances, numerous machine learning methods have been studied to improve the outcomes of human resources division.

It is crucial to maintain a permanent and promising workforce. This prediction model will help the organisation maintain their agility in a fast changing environment.

II. RELATED WORK

Opinions of top computer scientists led us to focus on the problems associated with employee turnover and attrition. The overall quality of a company drops down significantly as worker attrition causes a major setback. In most companies it is often difficult to find well trained and competent employees but it is even more complicated to replace an experienced employee. Both market expenses and workforce prices are affected due to these facts. There are often misconceptions regarding the literature surrounding this topic which has caused an ineffective treatment plan being made to deal with the churn rate. The motive of this paper is to build an adaptable framework for predicting the churn rate while considering different technical and non-technical attributes like education, potential, behaviour along with different classification techniques. HR department has many sets of responsibilities and a big quantity of knowledge to to generate on a regular time. The biggest responsibility of HR is to envision a proper way to find the correct replacement of the employees who have departed voluntarily or involuntarily.

In an organisation, an employee once trained and hone their skills in technical and managerial aspects often leave for a better opportunity. This leads to serious setbacks as replacing these workers means we have to teach the ropes again to a new employee which leads to interchanges within the company. The prediction model uses data gained from previous worker's experience to research the facts behind both attrition and turnover. We have applied twelve classification methods on the HR data. A feature choice approach needs to be considered based on the statistics and analysis of the results to combat the churn rate. It subsequently helps to reduce the prices which are a heavy result of attrition and turnover. There can be many reasons behind the departure of an employee with some causes being personal like family, health problems etc. or professional like equity, salary, work environment.

There is also the problem of relocation or transfer. In this endeavour, different supervised machine learning algorithms are explained, understood and measured in their potential to predict the employee churn. In the following paragraphs, we provide a general overview of the theory behind some of the algorithms.

1. Decision Tree (DT)

The most famous and powerful tool for building a prediction model. Decision Tree as the name implies has a tree like structure and flowchart design. There are three different parts of decision tree and the parts include an internal node which denotes a test on an attribute, a branch that represents the outcome of the test, and each leaf or terminal node that holds the labels of the class. The training of a decision tree is done by dividing the primary set into subsets on the basis of the features value test. A technique called Recursive Partitioning is repeated recursively on the subsets derived from the source. When the division no longer adds any predictive values, the recursion is completed and it is also finished when the target value is equal to the subset value at the node. To build a decision tree, the developer does not need any domain specific knowledge and the construction of the tree is appropriate for an exploration and discovery which gives a high accuracy. An advantage of decision tree is that it helps to build a classification model without involvement of big computations and helps in generating flexible rules. In order to find which attributes are most suitable for the churn prediction decision tree provides a significant contribution. However, decision trees are prone to errors and often require critical computation for training the data.

2. Random Forest (RF)

Random Forests supplements the decision tree structure by mixing a bunch of weak learning nodes to create a stronger learning node. The bagging method are used in the training of random forests. It is an ensemble method which uses divide and conquer approach. Random Forests are used for both classification and regression. Random forests have the primary objective of reducing the variance by being trained on different parts of the similar data sets. The process helps to boost the prediction model but leads to increase and decrease in bias and interpretability respectively.

3. Gradient Boosting Trees (GBT)

Gradient Boosting Trees are similar to random forests as they are both ensemble machine learning techniques used for regression and classification obstacles with the difference being that gradient boosting trees follow a more sequential operation. Normally weak prediction models are brought together in an ensemble to develop a prediction model with the help of decision trees. In this process, a series of trees are built to stabilise the mistakes made by the predecessor trees in the sequence. When the prediction model can no longer be enhanced, the incremental adding of trees are stopped sequentially. GBT does have some defects with the most prominent ones being that it is uneasy for interpretation, hard to visualise and and partial observation being made. Regardless, GBT is fast and memory efficient and the boosting helps to reduce over fitting through regularisation.

4. Logistic Regression (LR)

It is a classification algorithm built in a more traditional sense with the incorporation of linear discriminants. It was used initially in the biological field in the twentieth century before being developed for social science applications. LR is used when categorisation is required for the dependant variable. An instance of utilisation of logistic regression is to predict whether an email is spam or whether the tumour is malignant or not in cancer diagnosis. Logistic regression is brought in the field when linear regression is not suitable for classification due to its less bounded nature. There are three types of logistic regression which are-

4.1. Binary Logistic Regression
4.2. Multinomial Logistic Regression
4.3. Ordinal Logistic Regression

5. Support Vector Machine (SVM)

Support Vector Machine is often considered the second stepping stone of supervised machine learning. A SVN is a classifier with the ability to recognise and identify distinctions with accuracy by a separating dimension. When training data in labelled format is given, the algorithm gives a result in the form of an optimal hyperplane with new examples categorised. A hyperplane is a subspace whose dimension is one less than that of its ambient space. All machine learning experts should have a SVM in their arsenal as it can be used for both classification and regression task.

6. Neural Networks (NN)

A neural network is a sequence of algorithms that goes on to understand the relationships beneath a set of data by a process that imitates the way the human mind works. In this fashion, neural networks refer to the mechanism of neurons, either organic or unnatural. Neural networks have the ability of adapting to modifying input, the network cre-

ates the most accurate answers without the necessity required to recreate the output requirements.

III. DESIGN METHODOLOGY

Customer churn is an infamous issue and concern in the revenue industry. Customer churn basically involves customers moving from one product to another product or service. They replace their preferred brand with another brand of their choice which may be better or cheaper. Due to these reasons, churn prediction models are used for analysing the customer churn rate. But churn prediction and analysis is not just limited to one specific field. Employee Churn is similar to Customer Churn in terms of a problem statement and being a burden for the organisation but the model predicting the employee churn rate is much more complex than the model for customer churn rate. The ticket to running a good company is to retain its skilled workforce which ensures that the company operates in a smooth manner. In employee churn, the employees often leave the company either voluntary (Attrition) or involuntary (Turnover). Employee churn comprises of employee turnover and attrition which leads to obstacles like financial loss, dissatisfied customers, replacing experienced workers with newbies and teaching them the ropes again. Most companies focus on employee attrition as employees often leave the company in search of better opportunities, higher salary, good work environment or due to personal reasons like family issues, health problems, etc. There are also some more reasons which are often responsible for negativity like distrusts between co-workers, conflict of interest with the supervisors and lack of promotion. In this study we have focused on both voluntary and involuntary churn. Employee churn creates a vacuum in the company which makes it necessary to identify the factors responsible respectively for attrition and turnover. This research falls under the human resource analytics or people analytics.

We need more independent variable for making a good and simple prediction model. The module to properly classify the risk rate of an employee leaving the company or absence of risk is critical.

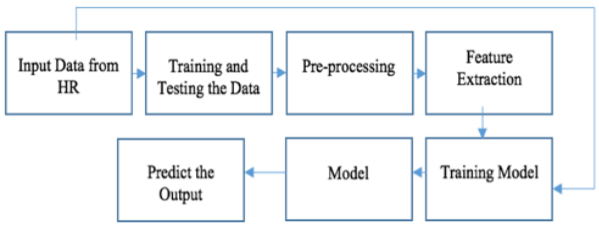

Figure 3.1: Prediction Model

It is also necessary to categorise the damage the company suffers when the employee leaves unexpectedly and come up with a good strategic plan. HR plays a crucial role in the establishment of a tech company. The prediction model helps to mitigate the effects when an employee quits. Employee Churn Prediction obtains the inside data from Human Resources Department which is mixed with statistical supervised machine learning methods to observe and generate a list of employees that will most likely resign in the foreseeable future. A Prediction Confident Score is also created for every employee. As the score increases, the more likely it is that the employee quits. A retention plan is also inferred which states the reason for the employee leaving the company and how to retain them.

In this experiment, the training dataset contained various technical and non-technical attributes.

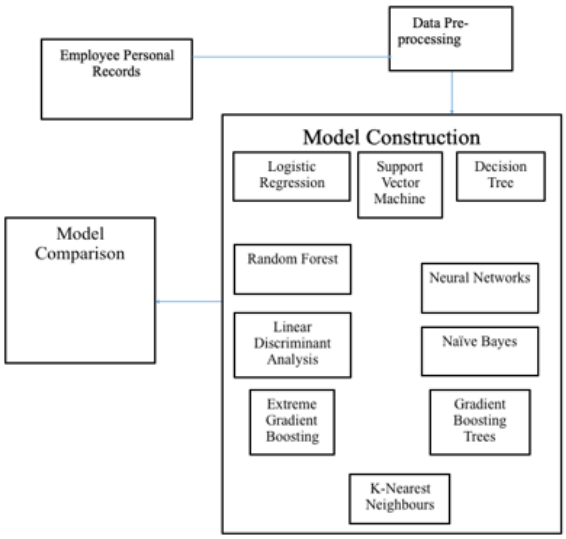

Figure 3.2: Research Stages

The attributes are decided by obtaining the background information of the employees. Data Preprocessing is mandatory as the information needs to be properly processed so that the noises present prominently in the HR data can be reduced. The data is compared and merged into one dataset. The task of removal of non-significant data is also executed. A select few attributes are selected for the final analysis and prediction. The prediction model requires a set of information for the classification algorithm that has been trained and arranged into some classes. In this research, we construct a classification model using ten supervised machine learning algorithms. After dividing the dataset into two sets of data, training data and testing data, the building of prediction model begins with 70% being reserved for training data and 30% for testing data. We make distinctions between the churners and non-churners. Churning is a very expensive

problem for the enterprise. The expenses for creating a replacement for an employee are often huge. Research shows that companies ordinarily pay one-fifth of an employee's salary to replace that employee, and there can be significant boost in the expenses if a senior executive is to be replaced.

This project is a standard supervised classification problem where the destination variable is the probability of an employee leaving the company. A strategic retention plan is created for all the risks associated with the employee churn rate and helps to classify whether the employee is at medium risk level or high risk level of leaving the company. The plan helps to conduct meetings more properly so that the work environment can be improved based on the risks which give a high probability of employee attrition and turnover.

IV. DATASETS

In this case study, a Human Resource dataset was obtained from IBM HR Analytics Employee Attrition & Performance which includes employee data for 1,470 employees with various data about the employees. We are going to use various datasets to predict when employees are going to quit or get fired by going through the main drivers of employee turnover and attrition. The HR dataset was created with the purpose of demonstrating the IBM Watson Analytics tool. Most of the experiments are conducted on public and standard dataset. Some of the attributes of the datasets are given below-

Education
1 Below College
2 College
3 Bachelor
4 Master
5 Doctor

EnvironmentSatisfaction
1 Low
2 Medium
3 High
4 Very High

JobInvolvement
1 Low
2 Medium
3 High
4 Very High

JobSatisfaction
1 Low
2 Medium
3 High
4 Very High

PerformanceRating
1 Low
2 Good
3 Excellent
4 Outstanding

RelationshipSatisfaction
1 Low
2 Medium
3 High
4 Very High

WorkLifeBalance
1 Bad
2 Good
3 Better
4 Best

V. CONCLUSION

Employee attrition and turnover has been classified as pivotal factors that curb the growth of company. In this research, the abilities of ten supervised machine learning methods were tested on various HR datasets. In addition to machine learning analysis, various data mining techniques have been featured and utilised including cross verification, information scaling and feature searching. Suggestions on how to use feature importance ranking and classifier visualisation examples for updating the interpretability of employee churn model.

Optimistic results can be obtained by using different algorithms on multiple datasets. When there are huge datasets, extreme gradient boosting is recommended for use because it has less information processing, modes predictive strength and ranks importance of parameters reliably and automatically. However, the user should attempt to find the classifier that best fits the data before deciding on a model.

VI. REFERENCES

1. Alao, D., Adeyemo, A.B.: Analyzing employee attrition using decision tree algorithms. (2013)
2. Zhao, Y., Fu, B.: Employee Turnover Prediction with Machine Learning: A Reliable Approach. (2019)
3. Harish, R., Tarun, N.: Predicting Employee Sustainability Using Machine Learning. (2018)
4. Sukhadiya, J., Kapadia, H.: Employee Attrition Prediction using Data Mining Techniques. (2018)
5. Yedida, R., Vahi, R.: Employee Attrition Prediction.
6. Barvey, A., Kapila, J.: Proactive Intervention

to Downtrend Attrition using Artificial Intelligence Techniques.
7. Saranya, S., Devi,J.: Predicting Employee Attrition using machine learning techniques and analysing reasons for attrition. (2018)
8. Gao, X., Zhang, C.: An improved random forest algorithm for predicting employee turnover. (2019)
9. Sonia, S., Rajakumar, S.: Churn prediction using MapReduce. (2014)
10. Richard, J., David L.: Collective Churn Prediction in Social Network.
11. Chourey, A., Mishra, S.: A survey paper on employee attrition prediction using Machine Learning Techniques. (2019)
12. Negi, G.: Employee Attrition: Inevitable yet Manageable. (2013)
13. Karande, S., Shelake, A.: Prediction of Employee Retention using Cassandra and Ensemble Learning. (2019)
14. Bhuva, K., Srivistava, K.: Comparative Study of the machine learning techniques for predicting employee attrition. (2018)
15. Palafox, L.: Prediction of student attrition using Machine Learning. (2019)
16. Punnoose, R., Ajit, P.: Prediction of employee turnover using machine learning methods. A case for extreme gradient boosting. (2016)
17. Dilip S., Pujahari A.: Evaluation of Machine Learning Models for Employee Churn Prediction. (2017)
18. Rajpoot, K.: Predicting employee attrition using Machine Learning. (2018)
19. Singh, M., Wang, J.: An analytics approach for proactively combating voluntary attrition of employees. (2012)
20. Devi, P., Umadevi, B.: A novel approach to control the employee's attrition rate of an organization. (2018)