Leveraging the Accuracy of Employee Attrition Model: A Machine Learning Approach

TANMAY DHAMDHERE[1], DR. VIPUL DALAL[2]
[1] *Student, Vidyalankar Institute of Technology*
[2] *Associate Professor, Vidyalankar Institute of Technology*

Abstract- Employee attrition is a process in which the employees working in a company quits his/her job due to various reasons. For instance, due to retirement, sacking from the organization or personal reasons. The task of churning the employee is a hectic one, as there are no fixed pattern or formula which can give an accurate prediction of whom to churn. But when implemented, it can certainly lead to a winning situation for the companies. The companies will be largely benefitted by the attrition of the employees as it can significantly reduce the cost of the labours, also it can bring an overall change which can positively affect the company's growth. Employees are the backbone for any company to bloom. There are certain factors such as age, increment, pay and many more which comes in picture for attrition of the employee. Using proper methodology and planning, it will be easy for a company to churn out the wrong employee and to proliferate their progress. In this paper, we have proposed a suitable method, which uses the techniques related to Machine Learning to yield an accuracy of 81.31%. Using this strategy, an overall accuracy of 96% can be achieved which can potentially help the companies towards its goal.

Indexed Terms- Machine Learning, attrition, churn, goal

I. INTRODUCTION

Attrition is a global scenario which has affected all the industries across the world [1]. It can happen because of various reasons, however ultimately it led to the reduction of workforce. In today's modern world, the Human resource team in any organization is very concerned about the attrition rate. The traditional method is mostly very costly and ineffective [2]. Accurately analyzing the data of the churn employees can yield better results and productivity. Attrition is good for unproductive employees and bad if the talented resource is churn of the organization. Prediction of the possible attrition can be known by historical and forecasting data analysis. Organization are sometime spending huge amount of money in providing unnecessary benefits to the employees hoping that this can help them retain in the workforce, however that is not very much true. Unscientific way of churning out employees can have disastrous effects in both short and long term within and outside the organization. In this paper, we have implemented an efficient method which makes use of Regression, a Machine Learning algorithm that calculates at a very high accuracy level.

For any Machine Learning model to implement effectively there are certain steps which need to be followed

Steps:
There are mainly 5 steps involved in Machine Learning:

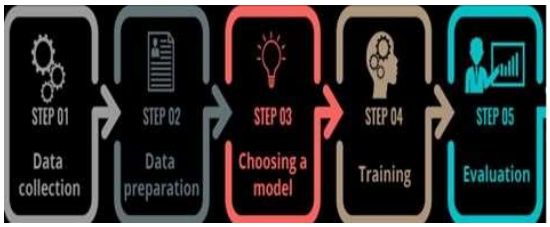
1. Collection of Data
   This is the first step towards finding how accurate the data is, i.e. the quality of the data is directly proportional to the accuracy of the model. If the quality and quantity of the data is good, it would help the model to yield precise accuracy.
2. Construction of Data
   This is the most important step among the step involved in machine learning. It is a process of cleaning, converting the raw data into useful and functional data.

a) Data cleaning, which is the most crucial step involved in this process. It is basically cleaning of the unwanted, redundant and corrupted data, which hampers the accuracy of the model.[3]

3. Selection of Algorithms
   Selection of an appropriate algorithm always gives an ideal solution of the model.[4]

There are different types of Machine Learning algorithms available. Such as Support Vector Machine (SVM), Random Forest, KNN, Logistic Regression and many more. But before choosing any of the algorithm, it is crucial to understand the data present in the set.

4. Data Pre-processing
   This is an important step in Machine Learning. This step involves splitting the data into test and train set.
5. Evaluation of the model
   This is the last step performed for evaluating the model's accuracy.
   This step can also involve utilization of various techniques such as hyperparameter tuning, feature engineering, ensemble methods which are used to increase the accuracy of the model.



The model in this paper uses Multiple Linear Regression algorithm, which is an extension of Simple Linear Regression. MLR allows us to model dependent variable 'y' as a Linear function of more than 1 predictor variables (x1, x2, x3…xk) [5].

For n number of know samples regression is given by:
$Y_i = \alpha + \beta_1 x1 + \beta_2 x2 + \beta_3 x3 + \beta_4 x4 + \ldots + \beta_k xk + \varepsilon_i$
For i=1 to n
α, β1, β2, β3, β4 regression coefficients
ε_i= represent error for i^{th} point.

II. IMPLEMENTATION

The dataset were given by the Hackerearth organization.

Data Set:

Train Set:
This is the data used to fit the model. It is a type of set which is used to build up a model, the model basically sees and learns from this dataset.[6]

Validation Set:
This is the sample data used to provide an unbiased evaluation of a model fit on the training dataset using hyperparameter tuning.

Test Set:
This is the actual set, where the model is trained i.e a subset to test the trained model. Using this dataset, we get the accuracy of the model

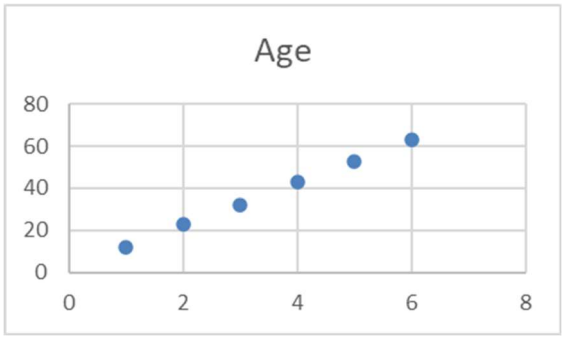
Variable Used in Data Set:

1. Employee_ID: It is the ID for each employee
2. Age: Age of a person
3. Unit: It has the department name for which an employee work
4. Gender: It indicates the gender of the employee (Numeric value)
5. Decision_skill_possess: The skill possessed by the employee
6. Post Level: Designation of the employee
7. Relationship Status: Status Married or Single
8. Pay Scale: Payment of the employee
9. Time of service: Time (Years) the employee was in the company
10. Growth rate: of an employee
11. Time since promotion: Time since last promotion received
12. Work life balance: work life balance of employee
13. Travel Rate: It is the history of travel rate
14. Hometown: City where the employee lives
15. Compensation and Benefits: Benefits for the employee
16. VAR 1 to 7: Unknown Variables (Anominised Variables)
17. Attrition Rate: The variables whose solution needs to be find out

Algorithm Used:
The algorithm used in, for this model is Multiple Linear Regression, due to the declaration of the variable description, as the problem needed to foresee the 'attrition rate' which is the 'target variable' ie the variable 'attrition rate' is dependent on other variables for its prediction. So, it's an ideal choice for selection of Multiple Linear Regression algorithm.

Let's take an example if the data present in the data contains a straight line i.e if the point creates a straight line then it's a linear curve.

Eg. For age variable, we take 6 values from it and then plot a curve using it.



We can clearly observe that the values create a linear curve. So, it's a regression problem.

Feature Selection:

There are in all 18 variables which were given in the dataset. However not all are useful for the prediction of the attrition rate of the model.

The art of choosing the correct variables depends on the nature of the dataset.[7]

In this dataset, we needed to find the 'attrition rate' for the employees. So basically, the attrition of any employees depends on certain criterias. For instance, their age, qualification etc.

The variables selected are 'Age', 'Pay_Scale', 'Time_of_service','growth_rate', 'V6','V7','Time_since_promotion'

The reason was choosing 'Anominised Variables' was because of its unknown values. These values may sometimes increase the accuracy, while sometimes it may decrease if. In this case, it increased the overall accuracy of the model so we had to keep it.

Feature Engineering:
There are various techniques of feature engineering namely one-hot encoding, binning, log transform, scaling and many more.[8]

But the most suitable technique for handling null values is Numerical Imputation. In numerical imputation, there are 3 different types. [9]

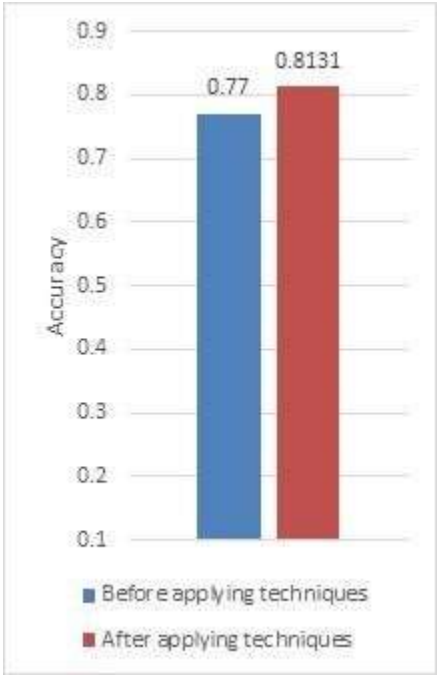
1. Mean Imputation
2. Median Imputation
3. Mode Imputation

For this particular dataset, we have choosen median imputation,i.e replacing all the null values by the median of that particular column.

The reason was selection of median imputation because the dataset has comparatively smaller numerical values ,its easy and can help to yield a better accuracy.

III. RESULTS

The employee attrition model was able to yield an accuracy of 77% before applying feature selection and feature engineering. However, after applying both the techniques, the model was able to yield an accuracy of 81.31%.So approximately 4 percent was enhanced by using these both techniques.



CONCLUSION

In modern world where so many advancements have taken place, but the problem of attrition of the employees is still on the run where the organizations are using their huge sum of money for employee attrition.[10] This is why, many potential employees ae getting churn out of the companies. To overcome this problem, the system presentated in this paper which works on Machine Learning concepts can yield a better accuracy and can potentially help the companies towards its growth.

REFERENCES

[1] N. Shilpa, A Study on Reasons of Attrition and Strategies for Employee Retention, IJERA, Dec 2015

[2] Saurabh Khanolkara, Mayuresh Gaitondea, Vishal Dabgotra, Leveraging the Efficiency of the Customer Retention Process: A Deep Learning Approach, IRJET, May 2020

[3] Tara Rawat, Dr. Vineeta Khemchandani,Feature Engineering (FE) Tools and Techniques for Better Classification Performance, International Journal of Innovations in Engineering and Technology (IJIET), May 2019

[4] Sebastian Raschka, Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, arXiv,2018

[5] Khushbu Kumari, Suniti Yadav, Linear regression analysis study, Journal of the Practice of Cardiovascular Sciences, January 2018

[6] Xue Ying, An Overview of Overfitting and its Solutions, IOP Conf. Series: Journal of Physics,2019

[7] Jianyu Miao, Lingfeng Niub, A Survey on Feature Selection, Information Technology and Quantitative Management (ITQM 2016)

[8] Ratnadeep R. Deshmukh, Vaishali Wangikar, Data Cleaning: Current Approaches and Issues, IEEE International Conference on Knowledge Engineering,Jan 2011

[9] M. Mostafizur Rahman and Darryl N. Davis, Machine Learning Based Missing Value Imputation Method for Clinical Datasets, Springer,2013

[10] Dr. K.Sunanda, AN EMPIRICAL STUDY ON EMPLOYEE ATTRITION IN IT INDUSTRIES- WITH SPECIFIC REFERENCE TO WIPRO TECHNOLOGIES, ISSN (Online)2231-2528,Sep 2017