iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5481

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Employee Attrition Prediction using Logistic Regression

Sri Ranjitha Ponnuru[1], Gopi Krishna Merugumala[2], Srinivasulu Padigala[3], Ramya Vanga[4], Bhaskar Kantapalli[5]
[1, 2, 3, 4]*UG Scholar,* [5]*Assistant Professor, Department of Computer Science and Engineering, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India*

Abstract: ***Worker wearing down is a circumstance looked by an association when the representative leaves the organization to join other association when he shows signs of improvement offer. It can likewise be named as Employee Defection. For the most part representative whittling down will be high when there is a squeezing need of workers in a specific industry because of mass retirements or development of association. At a certain point of time programming industry has confronted high whittling down rate by businesses because of enormous openings comprehensively in the product business because of the interest for programming items by all enterprises. Lessening the worker steady loss rate is a difficult issue looked by HR supervisors. This paper gives a definite perspective on foreseeing the representative turnover utilizing the Machine Learning algorithms. The forecast is finished utilizing the information sourced by IBM HR investigation. We utilized the Logistic Regression for the expectation and we got 85% exactness rate.***

Keywords: Employee Defection, HR supervisors, Logistic Regression, Machine Learning algorithm, Programming Industry.

I. INTRODUCTION

Employee Attrition can likewise be named as Employee Turnover. Wearing down is a basic issue and really high in these industry days. It is one of the serious issue in the vast majority of the associations. The progressive decrease in the quantity of representatives through retirement, renunciation or demise can be named as the Employee Defection. With regards to wearing down rates regular extent changes from industry to industry as far as its own principles and these rates can likewise vary among talented and incompetent positions. Organizations face overwhelming test of enrollment and holding gifts and simultaneously they have to deal with ability misfortune through steady loss be that because of industry midtowns of through intentional individual turnover. At whatever point a very much prepared and all around adjusted worker leaves the association, it makes a vacuum. So the association loses key abilities, information and bussiness connections. Current chiefs and individual executives are extraordinarily keen on lessening wearing down in the association, in such a path in, that it will contibute to the most extreme viability development and progress of the association. Anyway the representative acquiescences are organization for any bussiness. In the event that the circumstance has not dealt with appropriately, key staff individuals takeoffs can prompt a high misfortunes in profitability. Representative turnover brings about execution loses which can have long haul harmful impact on organizations. With whittling down rate being a genuine worry of each industry, organizations attempt to utilize inventive bussiness procedures to lessen the maintenance. There is no method to control steady loss totally, yet we can decrease it by arranging proper stratergies. Also, it would possible be able to be when supervisor predicts Employee trunover rate ahead of time. This paper can help to predict the voluntary attrition of the employee in a company by considering some of the factors like Age, Job Satisfaction, Monthly Income, Years At Company. The data of the employee is sourced from Kaggle by IBM HR analytics.

II. PROBLEM STATEMENT

Employee Attrition is an extremely large issue comprehensively. Whittling down rate is expanding step by step and particularly the product business is influenced the most in the current period. Why an Employee leaves an organization is the inquiry posed by the vast majority of the directors. Organizations even recruit private HR experts to contemplate the organization's work and discover why a representative is dissatisfied. HR division does the enrolling of new representatives and afterward send them for preparing with the goal that they can get work, work culture and become better experts. Every single organization faces representative turnover issue whether large or little. A worker leaves his current employment for another activity to show signs of improvement pay bundle and great working conditions. This makes an incredible misfortune to the organization. So the HR managers need to know number of employees need to be recruited and also the reasons for the high attrition rates. A statistical prediction need to be done.

©IJRASET: All Rights are Reserved

III.METHODOLOGY

The methodology we used is a Machine Learning Algorithm. As the term to be predicted is whether a particular employee leaves company or not. This problem comes under the classification techniques and we can use the binary classification techniques. Classification techniques are essential part of machine learning and data mining applications. Approximately 70% of the Data Science problems comes under classification problems. There are other category of classification that is Multinomial classification which handles the isseues where the multiple classes are present in the target variable.

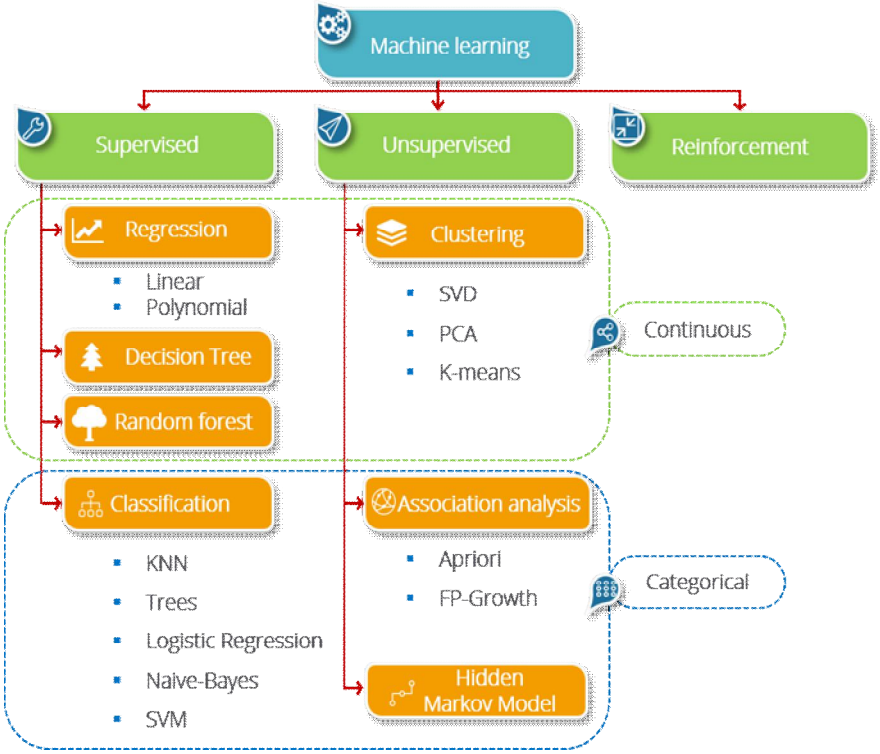


Fig. 1 Hierarchy of Machine Learning Algorithms.

The Logistic regression is one of the arrangement procedure. It is basic and most ordinarily utilized Machine learning calculations for two-class classifications. It is a measurable technique for predicitng the binary classes. It is anything but difficult to execute and can be utilized as the pattern for any parallel arrangement issues. Its essential thing ideas are likewise helpful in profound learning. Calculated regression depicts and assesses the connection between one ward twofold factor and free factors. Calculated Regression is the extraordinary instance of Linear Regression where the objective variable is straight out in nature. It utilizes a log of chances as the needy variable. Straight relapse gives you a ceaseless output, but calculated relapse gives a steady yield. Strategic Regression predicts the likelihood of event of a paired occasion using a logit work. The linear regression equation is given below in which y is dependent variable and X1, X2 …. and Xn are exploratory varibles.

$$y = \beta0 + \beta1X1 + \beta2X2 + …. + \beta nXn$$

A. Sigmoid Function

The sigmoid fucntion also called logistic function given an 'S' shaped curve that can take any real-valued number and map it into a value between 0 and 1. In the event that the bend goes positive boundlessness, y anticipated will get 0. On the off chance that the yield of the sigmoid capacity is more than 0.5 it is named the 1 or YES and in the event that it is under 0.5 it is named 0 or NO. for instance if the yield is 0.75, we can say as far as likelihood as there is a 75 percent of winning likelihood. On applying the sigmoid function we get

$$p = 1 / 1 + e^{-y}$$
$$e^{-y} = (p / p – 1)$$
$$y = \log(p / p – 1)$$
$$\log(p / p – 1) = \beta0 + \beta1X1 + \beta2X2 + … + \beta nXn$$

©IJRASET: All Rights are Reserved

The sigmoid curve for this is shown below :

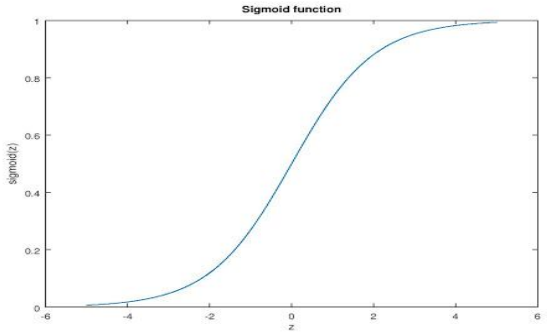


Fig. 2 S-Curve for Logistic Regression.

IV. AN OVERVIEW OF PROPOSED SYSTEM

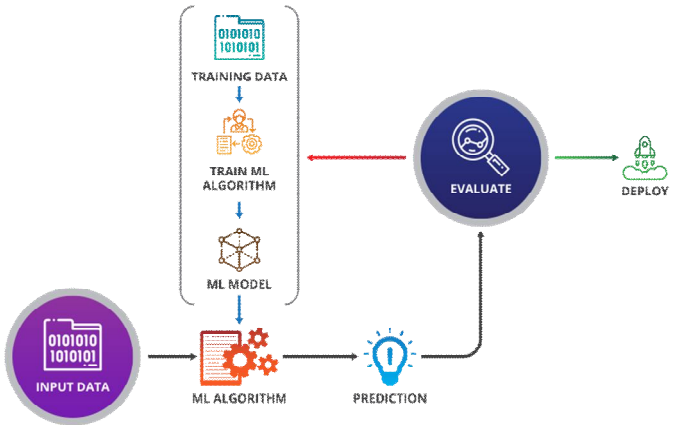


Fig. 3 Architecture Model.

The implementation of the proposed system is done by using Jupyter Notebook, IBM cloud for prediction and deploying it. To predict the result first we need to build the model using the previous data. To build a models there are mainly 4 steps.

1) Data gathering
2) Data Preprocessing
3) Model Training
4) Prediction

A. Data Gathering and Preprocessing

The data was sourced from IBM HR Analytics Employee Attrition and Performance which contains employee data for 1470 employees with various information about the employees[1]. It is collected from Kaggle. We used this dataset to predict whether the particular the leaves the company or not. The Data preprocessing is one of most step need to be followed to build a model. This includes different steps like

1) Import the Numpy, Pandas, Matplotlib libraries. The numpy is stands for Numerical Python which supports to store the large data and this supports to operate a highlevel mathematical functions. Pandas is used to analyze the data and manipulate it. Matplotlib is used as a visualization tool.
2) Checking the missing values and use the corelation heatmap to find varibles that doesn't have impact on the target variable and remove them. The data we considered doesn't have any missing values
3) Separating the independent and dependent variables
4) Converted the into numpy arrays, label encoding is performed on the categorical data and one hot encoding is done as the data has large variation in values like age and salary.
5) Spliting the data for training and testing. We used train_test_split from sklearn package to split the data. 70% of the dataset is used for training and 30% is used for testing the data.

©IJRASET: All Rights are Reserved

B. Model Training and Prediction

To train the model we need to import the model. As we are using the Logistic Regression we need to import the Logistic Regression class from sklearn.linear_model library. An object is created for the Logistic Regression class to implement its methods. The object declared is "classifier". In this is class we have fit() method whose parameters are the train values. The output we get is the model with the following details. After training the model its attribute values are set as class_weight = None, internet_scaling = 1, mutli_class = 'warn', tol = 0.0001.

Our model is trained using the training data and now we need to predicting by using the predict() method and test data. And the output is predicted values of test data.

V. RESULTS

A. Accuracy Score

By comparing the predicted values and actual test values we can get the accuracy of our model. For calculating the accuracy we have accuracy class from sklearn.metrics package. From that we got the accuracy 85% which is a good classification accuracy rate.

B. Confusion Matrix

The confusion matrix provides us a much more detailed representation of the accuracy score and the number of variables are correctly predicted and wrongly predicted. From the confusion matrix we can summarize the results of testing model as out of 30% of the data i.e 441 samples there are 367 true positive(TP), 8 true negative(TN), 4 false positive(FP), 62 false negative(FN).

C. ROC Curve

The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings.

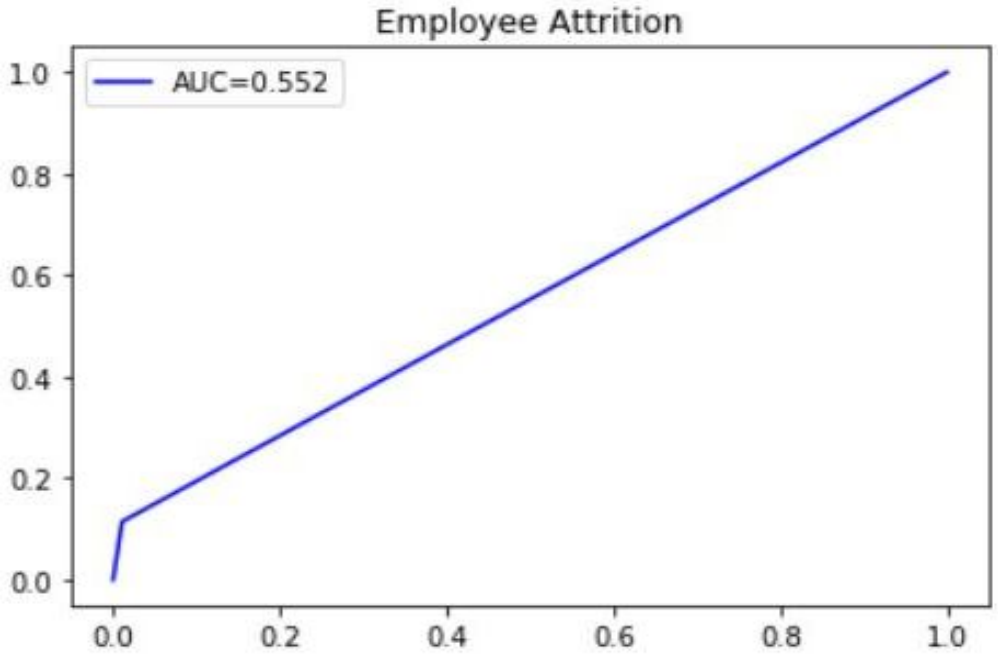


Fig. 4 ROC AUC Curve.

VI. CONCLUSIONS

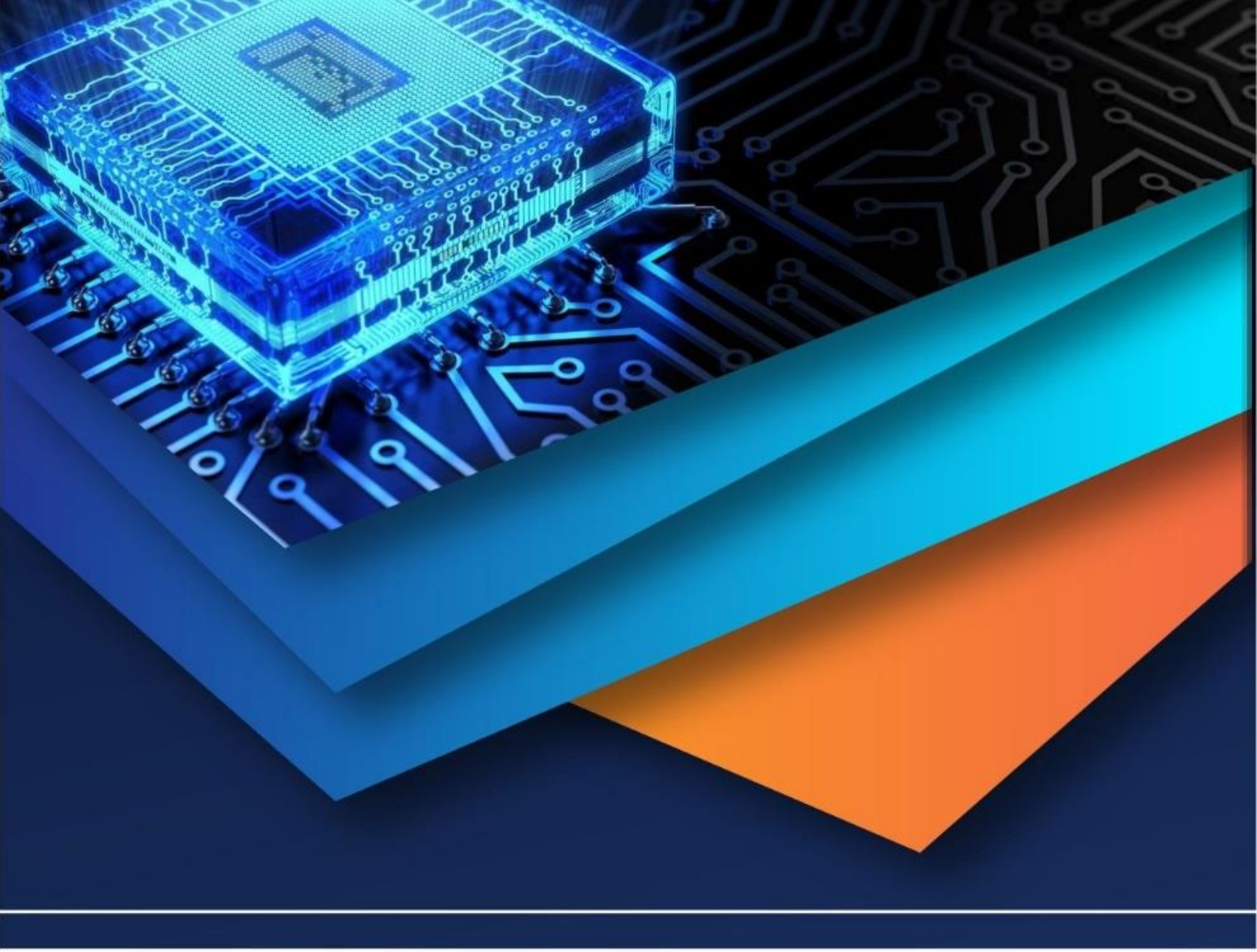
The principle point of the association is to acquire benefit. Yet, to gain most extreme benefit, the association should focus more on workers and the approaches to hold them for their since quite a while ago run. From the examination it is recognized that absence of development openings and pay are the main considerations which powers workers to change their occupations. This investigation infers that to diminish the steady loss, ventures ought to make a few open doors for the development of their representatives with in the association by receiving new imaginative advances and viable preparing programs.

©IJRASET: All Rights are Reserved

REFERENCES

[1] Cotton, J .L. and Tuttle, J .M., 1986. “Employee turnover: A meta-analysis and review with implications for research” Academy of management review, pp.55-70.
[2] B. Latha Lavanya, 2017. “A Study on Employee Attrition: Inevitable yet Manageable”. [Online]. Available: https://pdfs.semanticscholar.ord/d7f1/44238ce5e695e055af78fc0987023d3c2d0e.pdf
[3] Y. Rahul, V. Rakshit, K. Deepti, and Abhilash, “Employee Attrition Predition”. [Online]. Available: https://www.researchgate.net/publication/326059536_Employee_Attrition_Prediction
[4] https://patents.google.com/patent/US20090307025
[5] IBM HR Analytics attrition dataset, [Online]. Available: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
[6] Sarah S. Alduayj, Kashif Rajpoot. “Predicting Employee Attrition using Machine Learning”, Available: https://ieeexplore.ieee.org/document/8605976
[7] Hamza Bendemra. “Building an Employee Chrun Model in Python to Develop a Strategic Retention Plan. [Online]. Available: https://towardsdatascience.com/building-an-employee-chrun-model-in-python-to -develop-a-strategic-retention-plan-57d5d882c2d
[8] Saishruthi Swaminathan. “Logistic Regression – Detailed Overview”. [Online]. Available: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
[9] Bhasker Gupta, 2016. “Predictive Attrition Model: Using Analytics to Predict Employee Attrition”. [Online]. Available: https://analyticsindiamag.com/predictive-attrition-model
[10] Jason Brownlee, 2016. “Logistic Regression for Machine Learning”. [Online]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)