International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. 05 Special Issue 01 | 2020

Predicting Employee Attrition using Logistic Regression

Nikita Bhansali[1], Shrutika Nakaskar[2], Madhu Mandhane[3], Parnika Pathode [4]

[1,2,3,4] *Student, Computer Science & Engineering, SSGMCE, Maharashtra, India*

ABSTRACT

Employee attrition is one of the major problems at workplaces. Employee attrition causes a significant cost to any organization which may later on affect its overall efficiency. For any organization, finding a well trained and experienced employee is a difficult task, but it's even more complicated to replace such well-trained employees. This not only increases the cost of significant Human Resource (HR), but also impacts the market value of the company. A successful prediction model to predict the reasons for employee attrition is needed in order to avert various negative impacts for the organization. Therefore, the aim of this paper is to provide an appropriate framework for predicting the employee attrition rate by analyzing the employee's precise behaviors and attributes over demographic data using regression techniques. We will model the probability of attrition. The result thus obtained will be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay and consequently help them to reduce their human resource costs.

Keyword : - Employee Attrition, Logistic regression, Exploratory Data Analysis, Predictive Analysis, Machine Learning.

1. INTRODUCTION

Today corporate and individual performance are major determinants of one another, therefore HR analytics has become a game changing phenomenon of the times. HR Analytics has the ability to connect the most senior leaders with farthest distance reaches to employee and trainee performance [1]. An employee decides to join or leave an organization based on several reasons and factor, for instance, working environment, work place, office location, gender equity, pay equity etc. Others may have personal reasons such as relocation due to family, maternity, health issues, conflict with the managers or colleagues in a team. Employee attrition is a big issue for the organizations specially when well trained, technically strong and key employees leave for a better opportunity in a competitor organization. It requires time, little efforts and this results in financial loss to replace a well-trained employee. Therefore, we use the current and past employee data to analyze the common causes for employee attrition. The employee attrition prediction helps in recognizing and solving the issues that results in attrition. This information is helpful in possible retention of the current employees.

2. RELATED WORKS

Preeti K. Dalvi proposed system which provides a statistical survival analysis tool to predict customer churn based on comparison between decision trees and logistic regression. The correct combination of attributes or variables and proper usage of values gives the more accurate results and predicting the important factors of customer churn. [1]

Employee churn is a big issue for organizations. For instance Ibrahim Onurlap Yigit demonstrated that for employee attrition, data mining algorithms can be used to build a precise predictive models. They predicted the churn probability for each employee by using exploratory data analysis and data mining techniques, [2]

There have frequent studies on churn prediction analysis in the literature. For instance, I.M.M. Mitkees proposed to solve a big problem of customer churn related to a business, especially telecommunications by building models with different techniques such as Classification for prediction, Clustering for detection and Association for detection [3].

In this paper, Shrisha Bharadwaj proposed two models i.e. Logistic Regression model and an Artificial Neural Network model to predict churn in the mobile telecommunications industry. These models predict customer churn by considering the client's behavior and are independent of other client's data. [4]

3. PROBLEM DEFINITION

A company, at any given point of time, has 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. This level of employee attrition is immoral for organization value, because of following reason-

1. The former employee’s projects got delayed, which makes it difficult to meet timelines, resulting in a reputation loss among consumers and partners.
2. A sizeable department has to be maintained, for the purpose of recruiting new talent.
3. More often than not, the new employees have to be trained for the job and/ or given time to acclimatize themselves to the company.

4. OUR APPROACH

We have divide our approach into 5 phases. The phases are as follows:

a) Data collection
b) Data cleaning
c) Exploratory Data Analysis
d) Model Building
e) Testing and Evaluation

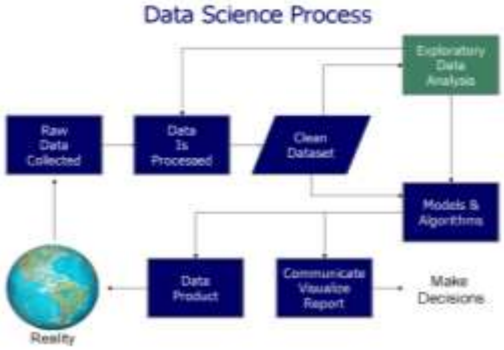

Figure 1. Data Visualization Process

4.1 Data Input

This involves data in the form of data sets. We have 5 data sets which has 30 attributes in total for each employee record. The data collected was divided two different samples of data. One data set sample was utilized for creating the model for the prediction of employee attrition, while the other sample was utilized for validating the model.

Table 1: Dataset Features

Sr no	Attributes	Datatypes	Sr no	Attributes	Datatypes
1	Age	Numeric	16	Marital Status	Categorical
2	Attrition	Categorical	17	Monthly Income	Numeric
3	Business Travel	Categorical	18	Number of Companies Worked	Numeric
4	Department	Categorical	19	Over 18	Categorical
5	Distance From House	Numeric	20	Percent Salary Hike	Numeric
6	Education	Numeric	21	Performance Rating	Numeric
7	Education Field	Categorical	22	Relationship Satisfaction	Numeric
8	Employee Count	Numeric	23	Standard Hours	Numeric
9	Employee Number	Numeric	24	Stock Option Level	Numeric
10	Environment Satisfaction	Numeric	25	Total Working Years	Numeric
11	Gender	Categorical	26	Training Time Last Year	Numeric
12	Job Involvement	Numeric	27	Work Life Balance	Numeric
13	Job Level	Numeric	28	Years At Company	Numeric
14	Job Role	Categorical	29	Years Since Last Promotion	Numeric
15	Job Satisfaction	Numeric	30	Years With Current Manager	Numeric

4.2 Data Cleaning / Exploratory Data Analysis

4.2.1Data cleaning and preparation:

Data cleaning and predation includes removal of the existing duplicates in the data by checking the data sets thoroughly. If required, merge some of the attributes or delete them. When our data is absolutely free from ambiguous and repeated values, this indicates that we can proceed further.

Steps of data cleaning and preparation are

1. Removed duplicates in the data by checking on employee id of employee.
2. Merged all the data file on the key employee id, created a master file.
3. Check whether predicted variable has any missing values.
4. As there are some records exists with predicted variable missing, which indicates that those applicants have been rejected.
5. We subsetted the data which has predicted variable missing and kept it aside for further usages in model evaluation.
6. After sub setting we removed those subsetted data from master data set.
7. Then we have checked for any missing values in predictor variables. We have found some variables which are having values missed.
8. As we have decided to do missing value imputation using WOE, we performed EDA as the next step to assess which are important predictor variables.

4.2.2 Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is the process of visualizing and analyzing to extract data insights from given dataset. Simply, EDA is the process of briefing important characteristics of data in order to enhance understanding of the dataset available.

1. We did analysis on each predictor variable present in the data with predicted variable to check the default rate effected
2. Conversion of metric data into categorical data from some demographical questions like Age, Year of service etc.
3. The missing data for performance rating was replaced with root mean squared value. Performance rating is an important attribute for analyzing attrition and as a result, the team could not afford to lose out on team members for whom the performance rating was missing.
4. We have used bar charts to check the default rate with each variable and found that some variables are stronger predictors of the predicted variable i.e., default rate is very high for some categories in the variable.
5. Data cleaning step involved cleaning of outliers, cleaning of invalid data points and removal of individuals whose information was missing.
6. A variable named ‘Attrition’ was created in the data set. This variable contain either ‘0’ (zero) or ‘1’ (one) depending on whether the employee was existing or separated respectively converting it to categorical data.
7. The model treated ‘Attrition’ as a dependent variable while demographic variables were treated as independent variables.
8. We stored this data file separately for further analysis.
9. As we can’t conclude with this minimal analysis on important variables, we decided to keep all the variables in the model building stage.
10. Next step is to build model using cleaned data.

4.3 Model building

As demographic data is merged in the main dataset we subsetted only demographic data from merged set and used for model building. Logistic regression modeling is built over the attrition data. Logistic Regression model includes all demographic variables and which is subsequently eliminated insignificant variables through an iterative process of model building and calculation of P-Value. Estimates are used to identify coefficients for significance for considering or removing a particular factor from the model.

As demographic data is merged in the main dataset we subsetted only demographic data from merged set and used for model building.

1. We did all the data preparation activities on demographic data. We removed data where variables having missing values as the number of records is less than 2%.
2. Checked for any outliers and replaced outliers with recent quantile values.
3. Then we go on building model by dividing data into train and test data sets.

4. After building model we have checked how good our model is performing on test data and also analyzed what are the important variables our model has given.
5. As we have different models to try out, we have chosen logistic regression model to start with.
6. We converted all categorical variables to dummy variables.
7. We divided data into train and test and built model using train data set.
8. We did check P-Value and VIF for variable importance and correlation factors and removed the variables which is of less importance and high correlation factor.
9. In the final model we are left with the variables which are of highly important in predicting the default of an applicant.

1.4 Testing & execution

From the models that we have built we conducted tests on the models which gives more default prediction rate. We divide the dataset into training and testing dataset in the ratio 70:30 .The 70% of the data is used for training or the learning of the model while the other 30% is needed to validate the model built. The model testing was done the remaining 30%of data.

The final attributes that predicted the reasons of attrition of employee's are listed in the table 2.

Table 2. Final Attributes

Attribute name	Description
Age	Age of the employee
Business Travel	How frequently the employees travelled for business purposes in the last year
Environment Satisfaction	Work Environment Satisfaction Level
Job Role	Name of job role in company
Job Satisfaction	Job Satisfaction Level
Marital Status	Marital status of the employee
Number Companies Worked	Total number of companies the employee has worked for
Total Working Years	Total number of years the employee has worked so far
Training Times Last Year	Number of times training was conducted for this employee last year
Work Life Balance	Work life balance level
Years Since Last Promotion	Number of years since last promotion
Years With Current Manager	Number of years under current manager

4.4 Model evaluation

From the models that we have built we conducted tests on the models which gives more default prediction rate. For this we followed different approaches respective to the models. We found that logistic regression model is predicting the likelihood of default. We have chosen that logistic regression model is best for our data.

The three main characteristics we considered in model evaluation are

1. Sensitivity
2. Specificity
3. Accuracy

5. FINAL OUTPUT

The model accuracy is checked by the test data (30% of remaining) data. The model created on training data (70% of data) gives the similar result thus validating the result. The values predicted by this model is as listed in the tables below. The model evaluation is performed by using the factors listed in the table 3 and table 4.

Table 3. Significant Values

Accuracy	0.772093
Specificity	0.7703349
Sensitivity	0.7703349

Table 4. Significant values

Positive Predicted Value	0.60748
Negative Predicted Value	0.87828
Balanced Accuracy	0.63608
Detection Rate	0.05039

6. VISUALIZATION

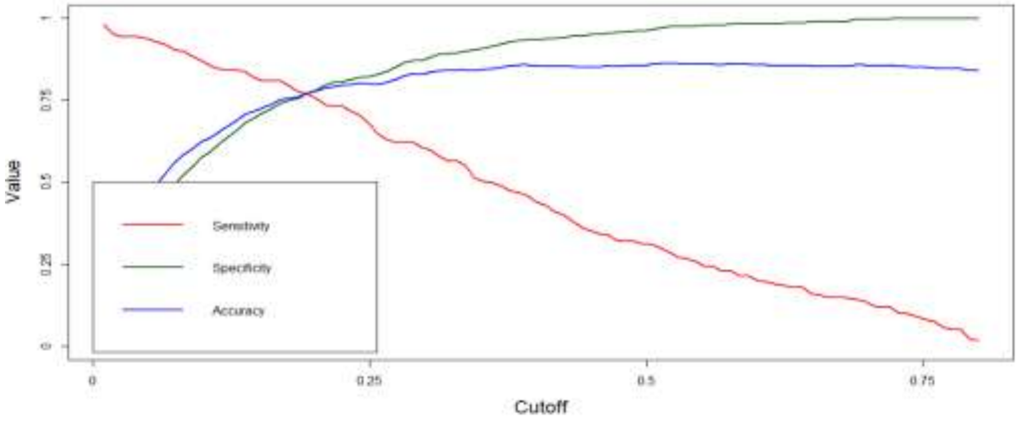

Figure 2. Graphical representation

7. CONCLUSION

The corporate world has changed drastically over the past few decades. These changes are due to globalization, huge amount of data, and the requirements of a high-tech technologies. One specific change that has been noticed correlates to the resources held by a company. Whereas in the 1970s, more than 95% of a company's assets could be attributed to tangible holdings, by the early 2000s that number had reduced to less than 30% (McClure, 2003). This means that more than 70% of a firm's total worth is due to intangible assets, including human capital. This highlights the importance of evaluating how to best use any available strategies, including using HR analytics to meet the demand for better decisions on how a firm may invest it's in limited resources.

In data analytics, frequently one sets out to solve a specific problem that leads to different direction by the data completely. Although the primary problem should not be forgotten, there is a need to acceptant to other possibilities for improvement organization as analytics majority of times highlights the hidden management aspects and issues to the focus. Efforts in data analytics help to not only inspire firm-wide improvement in terms of proper management of human resource, but, also by providing quantitative information about HR performance, can also put HR leaders on more equal footing with competing organization for limited resources (Schwarz & Murphy, 2008).These models can help us in listing the features with higher impact in attrition of an employee and the possible reasons behind it so that HR can take suitable decision for the retention process

8. REFERENCES

[1] Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, Prof. V. A. Kanade, "Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression", 2016 Symposium on Colossal Data Analysis and Networking (CDAN).

[2] Ibrahim Onuralp Yigit, Hamed Shourabizadeh, "An Approach for Predicting Employee Churn by Using Data Mining", IEEE, 2017.

[3] I. M. M. Mitkees, S. M. Badr and A. I. B. El Seddawy, "Customer churn prediction model using data mining techniques," 13th International Computer Engineering Conference (ICENCO), pp. 262-268, Cairo, 2017.

[4] Shrisha Bharadwaj,Anil B.S.,Abhiraj Pahargarh, Adhiraj Pahargarh, P S Gowra, Sharath Kumar, "Customer Churn Prediction in Mobile Networks using Logistic Regression and Multilayer Perceptron(MLP)" , 2018 IEEE.

[5] Andry Alamsyah, Nisrina Salma, "A Comparative Study of Employee Churn Prediction Model", 2018 4th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia.

[6] Rupesh Khare , Dimple Kaloya , Chandan Kumar Choudhary, Gauri Gupta," Employee Attrition Risk Assessment using Logistic Regression Analysis", 2nd IIMA International Conference on Advanced Data Analysis, Business Analytics and Intelligence,2011.

[7] Sandeep Yadav, Aman Jain, Deepti Singh," Early Prediction of Employee Attrition using Data Mining Techniques" ,IEEE,2018.