International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-06, Issue-07, OCT 2020

A Study of Employee Attrition and prediction using Logistic Regression in R programming

Pramit Banerjee, Systems Analyst (IT MNC), Kolkata, India, pramiteejanai@gmail.com

Anchita Mishra, MBA Student (Full Time), IISWBM Kolkata, India, anchitamishra10@gmail.com

Abstract: Employee attrition is a major concern for the employer and most of the organizations are trying to finding ways to deal with the situation. It is very difficult to ascertain which set of employees will stay with the organization and who will leave the organization in the near future without understanding the pattern of employee behavior. This work aims to analyze the employee behavior based on the sample data and generate relationships among the data so as to build a model which may help any organization to predict the attrition of employees within the organization.

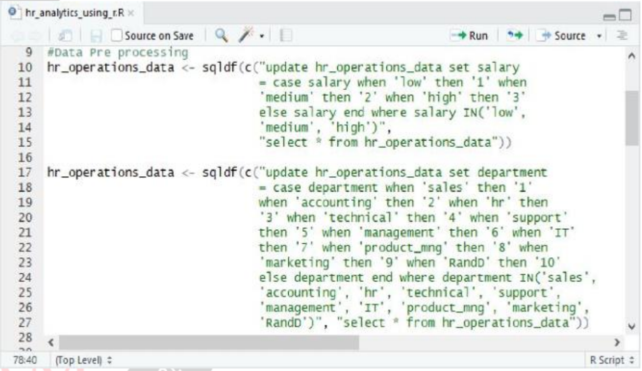
Keywords — Employee Attrition, Logistic Regression, Predictive Modelling, Machine Learning, R programming

I. Introduction

Employee attrition is a major concern for most of the organizations and it is tracked by organizations so as to maintain the optimal level of workforce for their day to day operations. If employees start leaving the organization, then without proper planning or a back up, it becomes difficult for the organization to maintain the daily operational activities, hampering the profitability in turn. In order to track the employee attrition, the different signals are being observed in the form of data which if properly analyzed, would lead to conclusions as to how many people and which set of people might exit the organization in the near future. One such attempt has been made through this research where historical data catering to the employees within an organization has been taken, such as the satisfaction level, last evaluation, project number, average monthly hours, time spent in company, work accident, promotion in last five years, department, salary and based on this the output has been predicted whether the employee is going to leave the organization or not. In this process, logistic regression model has been applied and the tool used is R programming. The data used here is open source data from Kaggle and github.

II. Data Preparation

The data consists of ten columns and out of these parameters, we consider the attribute employee left, as the output variable (dependent) which is a function of the other nine independent variables like satisfaction level, last evaluation, project number, average monthly hours, time spent in company, work accident, promotion in last five years, department, salary. At first, we have carried out the data preprocessing stage where the data has been cleaned and modified. For example, in the department column, there are ten distinct values and we have assigned categories from one to ten as numbers instead of text and also for the column salary, we have three distinct values and we have assigned categories from one to three as numbers instead of text. The libraries used in R programming for operations here are sqldf, tidyverse, caret. The code snippet is attached herewith.



```r
#Data Pre processing
hr_operations_data <- sqldf(c("update hr_operations_data set salary
                              = case salary when 'low' then '1' when
                              'medium' then '2' when 'high' then '3'
                              else salary end where salary IN('low',
                              'medium', 'high')",
                              "select * from hr_operations_data"))

hr_operations_data <- sqldf(c("update hr_operations_data set department
                              = case department when 'sales' then '1'
                              when 'accounting' then '2' when 'hr' then
                              '3' when 'technical' then '4' when 'support'
                              then '5' when 'management' then '6' when 'IT'
                              then '7' when 'product_mng' then '8' when
                              'marketing' then '9' when 'RandD' then '10'
                              else department end where department IN('sales',
                              'accounting', 'hr', 'technical', 'support',
                              'management', 'IT', 'product_mng', 'marketing',
                              'RandD')", "select * from hr_operations_data"))

```

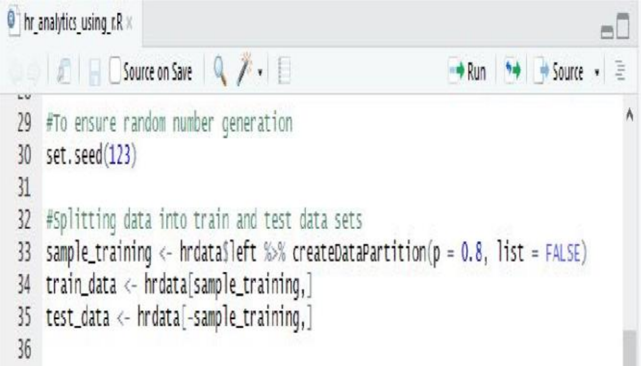
Figure - 1

In the above figure, the data set is first imported and then two update operations are performed using the sqldf package in R where the salary and department columns are revised by replacing the character values by numerical distinct values.

III. Methodology and Implementation

After the data pre processing stage, we have split the data into the test data set and the train data set according to the principles of machine learning. The data set contains around 15000 inputs and we have broken the dataset in such a way that we have trained around 80% of the dataset and based on the inputs and patterns of the train data set, we have applied it on the remaining 20% which is the test dataset. The code snippet is attached herewith.

DOI : 10.35291/2454-9150.2020.0633

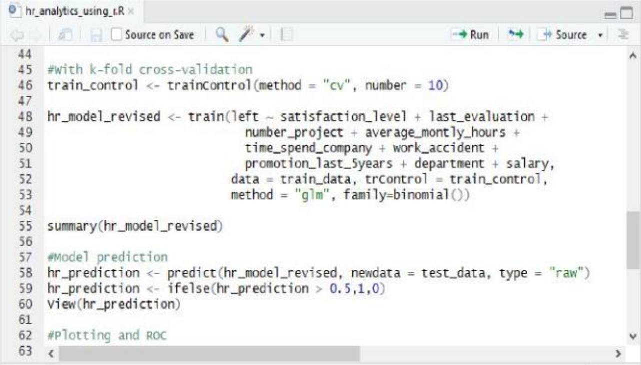


```r
#To ensure random number generation
set.seed(123)

#Splitting data into train and test data sets
sample_training <- hrdata$left %>% createDataPartition(p = 0.8, list = FALSE)
train_data <- hrdata[sample_training,]
test_data <- hrdata[-sample_training,]
```

Figure - 2

In the above figure, using the method createDataPartition using the caret library in R, the data partition has been done and two data frames train and test data have been created subsequently.

In order to ascertain better accuracy, we have done 10 times cross validation in this train and test data set before applying the model. This will in turn help the model to be more accurate. Finally, we have applied the model using logistic regression where employee left is taken as a function of the other parameters. The code snippet is attached herewith.



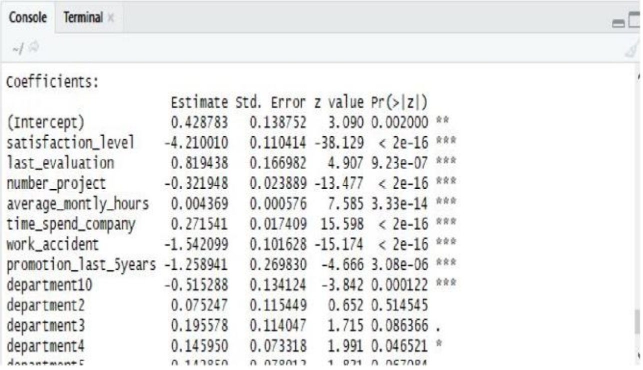
```r
#With k-fold cross-validation
train_control <- trainControl(method = "cv", number = 10)

hr_model_revised <- train(left ~ satisfaction_level + last_evaluation +
                            number_project + average_montly_hours +
                            time_spend_company + work_accident +
                            promotion_last_5years + department + salary,
                          data = train_data, trControl = train_control,
                          method = "glm", family=binomial())

summary(hr_model_revised)

#Model prediction
hr_prediction <- predict(hr_model_revised, newdata = test_data, type = "raw")
hr_prediction <- ifelse(hr_prediction > 0.5,1,0)
View(hr_prediction)

#Plotting and ROC
```

Figure - 3

In the above figure using the trainControl method in R, the ten-fold cross validation has been performed and afterwards, logistic regression is performed in train data set and the test data set is applied in the model.

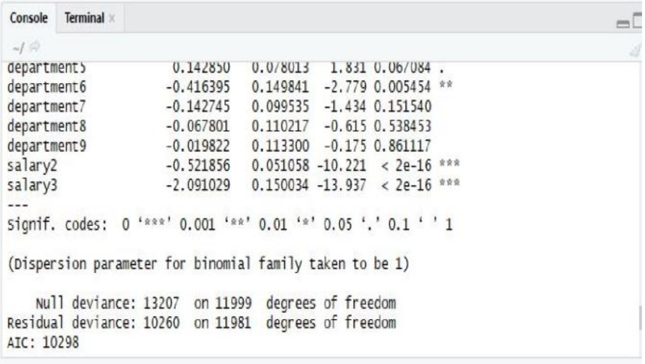
IV. RESULTS AND CONCLUSION

The output of the model is found in the R console and it is given as below:



```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.428783   0.138752   3.090 0.002000 **
satisfaction_level    -4.210010   0.110414 -38.129  < 2e-16 ***
last_evaluation        0.819438   0.166982   4.907 9.23e-07 ***
number_project        -0.321948   0.023889 -13.477  < 2e-16 ***
average_montly_hours   0.004369   0.000576   7.585 3.33e-14 ***
time_spend_company     0.271541   0.017409  15.598  < 2e-16 ***
work_accident         -1.542099   0.101628 -15.174  < 2e-16 ***
promotion_last_5years -1.258941   0.269830  -4.666 3.08e-06 ***
department10          -0.515288   0.134124  -3.842 0.000122 ***
department2            0.075247   0.115449   0.652 0.514545
department3            0.195578   0.114047   1.715 0.086366 .
department4            0.145950   0.073318   1.991 0.046521 *
```

Figure - 4

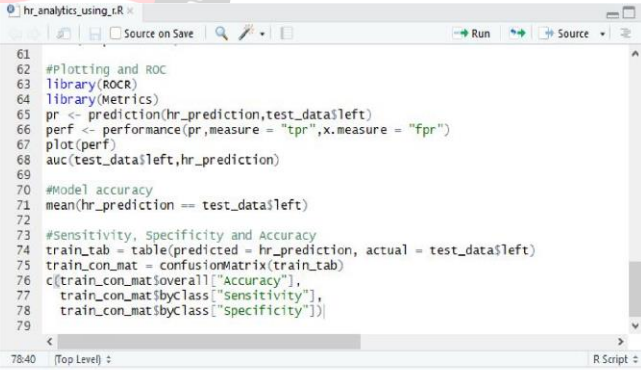


```
department5            0.142850   0.078013   1.831 0.067084 .
department6           -0.416395   0.149841  -2.779 0.005454 **
department7           -0.142745   0.099535  -1.434 0.151540
department8           -0.067801   0.110217  -0.615 0.538453
department9           -0.019822   0.113300  -0.175 0.861117
salary2               -0.521856   0.051058 -10.221  < 2e-16 ***
salary3               -2.091029   0.150034 -13.937  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13207  on 11999  degrees of freedom
Residual deviance: 10260  on 11981  degrees of freedom
AIC: 10298
```

Figure - 5

From the above output, we see that there is strong negative correlation between employee left and satisfaction of the employee. This denotes that those employees whose satisfaction level is less, is more likely to leave the organization. Likewise, we see that those employees who have had a work-related accident, with less promotions in last five years are also more likely to leave the organization. From the above output, we see the p-value of the parameters. We find that for the parameters, satisfaction level, last evaluation, number project, average monthly hours, times pend in company, work accident, promotion in last five years, salary, the p value is less than 0.05 which means that they are statistically significant and hence these parameters play an important role in the employee attrition.

In order to identify the other metrics, we find out the accuracy, sensitivity and specificity of the model using the libraries ROCR and Metrics. The code snippet is given below:



```r
#Plotting and ROC
library(ROCR)
library(Metrics)
pr <- prediction(hr_prediction,test_data$left)
perf <- performance(pr,measure = "tpr",x.measure = "fpr")
plot(perf)
auc(test_data$left,hr_prediction)

#Model accuracy
mean(hr_prediction == test_data$left)

#Sensitivity, Specificity and Accuracy
train_tab = table(predicted = hr_prediction, actual = test_data$left)
train_con_mat = confusionMatrix(train_tab)
c(train_con_mat$overall["Accuracy"],
  train_con_mat$byClass["Sensitivity"],
  train_con_mat$byClass["Specificity"])
```

Figure - 6

In the above figure, using the confusionMatrix function in R, the test data has been compared against the predicted data and thereby the true positive rate, i.e. the sensitivity, the true negative rate i.e. the specificity and the accuracy which is defined as ratio of sum of true positive and true negative to the sum of true positive, true negative, false positive and false negative data.

The other metrics are as follows:

Accuracy: 0.7902634

Sensitivity: 0.9243478

DOI : 10.35291/2454-9150.2020.0633

Specificity: 0.3490701

We find that the accuracy of the model is quite good and the sensitivity is also very high which denotes that the true positive rate is good and the correct prediction has been done on the test data.

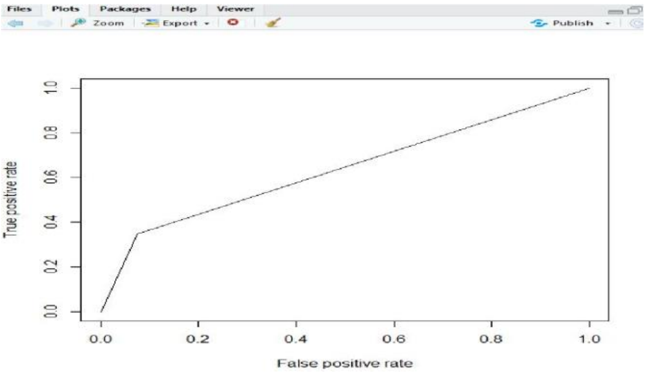


Figure - 7

In the above figure, we see the ROC (receiving operating characteristics) curve. Here, the X -axis is the false positive rate and the Y-axis is the true positive rate. The more the area under the curve, the better is the accuracy of the model.

We see that the area under the curve is: 0.636709, which indicates that the model is quite good.

V. FUTURE WORK

Through this model, it has been highlighted, how to predict the employee attrition of an organization using logistic regression. There is further scope of study using this model which can be carried out using higher volumes of data and across different data sets. We would also like to use different other algorithms in machine learning to optimize the results.

VI. REFERENCES

[1] Sung-Lin Chan, Xiang zhe Meng, S¨uha Kagan K¨ose, “EPFLMachineLearningCS-433-Project2Twitter Sentiment Analysis”, 2018

[2] A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses by Jessica L. Alzen, Laurie S. Langdon & Valerie K. Otero

[3] Logistic Regression Analysis of Graduate Student Retention by Sandra W. Pyke & Peter M. Sheridant, The Canadian Journal of Higher Education, volume XXIII-2, 1993

[4] Human Resource Analytics Kaggle Dataset by Ryan Nazareth and Hannes Draxl

[5] https://github.com/ryankarlos/Human-Resource-Analytics

[6] Twitter Sentiment Analysis using XG Boost and Random Forest Classification Algorithms: A Hybrid Approach by Ashwini M. Joshi, Sameer Prabhune, Nesara B R, IJREAM Vol – 05, 6th Sept 2019.

[7] Binary Logistic Regression Analysis in Assessment and Identifying Factors That Influence Students’ Academic Achievement: The Case of College of Natural and Computational Science, Wolaita Sodo University, Ethiopia by Bereket Tessema Zewude (MSc) 1* Kidus Meskele Ashine (Ass. Professor)2 1. Wolaita Sodo University, College of Natural and Computational Sciences, Wolaita Sodo, Ethiopia, P.O.Box 138 2. Wolaita Sodo University, School of Law, Wolaita Sodo, Ethiopia. P.O. Box 138 - Journal of Education and Practice www.iiste.org ISSN 2222-1735 (Paper) ISSN 2222-288X (Online) Vol.7, No.25, 2016

[8] Application of Logistic Regression in the Study of Students’ Performance Level (Case Study of Vlora University) by Miftar Ramosacaj Prof. Dr. Vjollca Hasani Prof. Dr. Alba Dumi - Journal of Educational and Social Research MCSER Publishing, Rome-Italy, Vol. 5 No.3 September 2015

[9] The importance of Logistic Regression implementation in the Turkish livestock sector and logistic regression implementations/fields by Murak Korkmaz, Selamy Guney, Sule Yuksel Yigiter – Journal Article, J.Agric. Fac. HR.U., 2012, 16(2): 25-36

[10] Noora Shrestha. Application of Binary Logistic Regression Model to Assess the Likelihood of Overweight. American Journal of Theoretical and Applied Statistics. Vol. 8, No. 1, 2019, pp. 18-25. doi: 10.11648/j.ajtas.20190801.13

[11] http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/

DOI : 10.35291/2454-9150.2020.0633