PREDICTING EMPLOYEE ATTRITION USING MACHINE LEARNING ALGORITHMS AND ANALYZING REASONS FOR ATTRITION

S. Saranya *, J. Sharmila Devi**
*(Department of Computer Science and Engineering, Loyola-ICAM College of Engineering and Technology, Chennai, India.
Email: saranya.19cs@licet.ac.in)
** (Department of Computer Science and Engineering, Loyola-ICAM College of Engineering and Technology, Chennai, India.
Email: sharmiladevi.19cs@licet.ac.in)

ABSTRACT

Whenever an employee leaves an organization, there is a source of advantage for the business competitor because of the invaluable tacit knowledge that the employee carry with them. Therefore, to be continuously competitive in the business, the organization should minimize the employee attrition. "Predicting the employee attrition and the reason for an employee leaving an organization" was performed to perceive the reasons, why the best and most experienced employees quit the company prematurely and try to predict which valuable employees are probable to leave the organization subsequently so as to find the areas where the organization is lagging behind. This model can be used by the Human Resource departments of the organizations to form efficient strategies to retain the valuable employees before they start looking for new jobs like by providing a hike in their salary, offering promotions if necessary travel and stay abroad or start the hiring process.

Keywords *- Employee Attrition, Classification, Logistic Regression, Naïves Bayes, Retention.*

I. INTRODUCTION

As in [1], The Barron's Business dictionary defined attrition as the normal and uncontrollable reduction of a work force because of retirement, death, sickness, and relocation. There are two types of employee turnover which are voluntary (turnover initiated by the employee) and involuntary (turnover initiated by the organization). This paper only focuses on the voluntary turnover. Involuntary turnover is not being concerned much because the decision is indeed made by the organization due to reasons like recession, retirement, death or in rare cases can be misconduct of the employee. Most companies face a formidable challenge of recruiting and retaining talents while at the same time having to manage talent loss through voluntary employee attrition.

Attrition becomes a problem to the company, considerably when an IT professional leaves an organization as this reduces the number of employees working for assignment. The professionals also take with themselves the tacit knowledge and the understanding of the specific business operations. Losing talented employees result in performance losses if the departing talent leaves gaps in its execution capability and human resource functioning which not only includes lost productivity but also possibly loss of work, team harmony and social goodwill. Nowadays the Human Resource department is greatly interested in reducing Attrition in the organization. One of the biggest problems that plague companies in the competitive marketplace is "How to retain the valuable employees before they start looking for new jobs?". The major aim of the paper is to analyze the employee dataset of an organization and find the reasons, why the best and most experienced employees leave the company prematurely and also try to predict which valuable employees are probable to leave the organization subsequently. This analysis was performed by using the graphs and algorithms available in R, a software environment for statistical computing.

II. RELATED WORK

II.I Data Collection

Data collection refers to the collection of relevant data from all the relevant sources to perform the analysis. The data used for this employee attrition analysis was gathered from sources like peer group of an employee, HR manager and self-assessment of an employee. The data set consists of the employee details of 14999 records.

The below mentioned attributes were considered from the employee database for building the target model.

a) Number_of_Projects
b) Satisfaction_level

c) Promotion_last_5years
d) Distance_Form_home
e) Average_Monthly_Hours
f) Years_at_Company

II.II Data Preparation

The data was prepared, pre-processed and cleaned with the help of data cleaning concepts in R software. The missing data in the dataset were identified and replaced with the global mean of the dataset. As we are concerned only about the efficient employees, the data set is further filtered by setting constraints to three of the attributes. The constrained attributes are

a) Last_Evaluation
b) Number_of_Projects
c) Time_Spent_in_Company

II.III Data and Quality report

Data quality refers to the condition of a set of values of qualitative or quantitative variables. Data quality is reported to be high if it is fit for intended uses in operations, decision making and planning. The data quality report of the dataset used gives the average for each of the parameter possessing numerical values and the class and mode for the parameters with categorical values. Selecting the part of the data by finding the efficient employees based on the three parameters.

1. last_evaluation if it is above 0.70
2. time_spent_company if it is more then 4 hours
3. number_project if it is more than 5

```
satisfaction_level last_evaluation  number_project  average_montly_hours
Min.   :0.0900     Min.   :0.3600   Min.   : 2.000  Min.   : 96.0
1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0
Median: 0.6400     Median: 0.7200   Median: 4.000 Median: 200.0
Mean   : 0.6128    Mean   : 0.7161  Mean   :3.803   Mean   :201.1
3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000   3rd Qu.:245.0
Max.   : 1.0000     Max.   : 1.0000   Max.   : 7.000   Max.   : 310.0
time_spend_company Work_accident    left             promotion_last_5years
Min.   : 2.000     Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.00000
1st Qu.: 3.000     1st Qu.:0.0000    1st Qu.:0.0000    1st Qu.:0.00000
Median: 3.000      Median: 0.0000    Median: 0.0000    Median: 0.00000
Mean   : 3.498     Mean   : 0.1446   Mean   : 0.2381   Mean   : 0.02127
3rd Qu.: 4.000     3rd Qu.:0.0000    3rd Qu.:0.0000    3rd Qu.:0.00000
Max.   : 10.000    Max.   : 1.0000   Max.   : 1.0000   Max.   : 1.00000
sales              salary
Length: 14999      Length: 14999
Class: character   Class: character
Mode: character    Mode: character
DistanceFromHome NumCompaniesWorked   OverTime        YearsWithCurrManager
 Min.   : 1.00     Min.   : 0.000     Length: 14999     Min.   : 0.000
 1st Qu.: 2.00     1st Qu.:1.000      Class: character  1st Qu.: 2.000
 Median: 7.00      Median: 2.000      Mode: character   Median: 3.000
 Mean   : 9.13     Mean   : 2.686                       Mean   : 4.164
 3rd Qu.:14.00     3rd Qu.:4.000                        3rd Qu.: 7.000
 Max.   : 29.00    Max.   : 9.000                       Max.   : 17.000
YearsAtCompany
Min.   : 0.000
1st Qu.: 3.000
Median: 5.000
Mean   : 7.071
3rd Qu.:10.000
Max.   : 40.000
```

Fig. 1 Data quality Report

III. IMPLEMENTATION

To predict the reason for the employee attrition, three algorithms were used which are Tree Learning Algorithm, Naïve's Bayes Algorithm and Logistic Algorithm. The data mining techniques for the prediction are as follows:

III.I Tree Learning Algorithm

Decision tree is a conventional algorithm used for performing classifications based on the decisions made in one stage. This provide tree structured representation of the decision sets. This classification based on the decision tree let us predict the qualitative response without creating the dummy variables also the class proportion among the training region which fall into particular region can be predicted using this algorithm. This algorithm works fast and the tree structure of the algorithm in easy understandable. Each node in the tree is the representation of an attribute which is being tested for making a decision, every branch is the representation of the output of that test, the leaf nodes are the distributed sub-classes. The part of the data set was trained to create a decision tree model and the trained model was used for prediction in the other part of the data set. The r part function in the R tool was used to perform the classification. The accuracy of the algorithm was established. As [2], The Tree Learning algorithm out performs Naïve Bayes on large dataset.

III.I.I Code Snippet:

```
library("rpart")
library("rpart.plot")
install.packages('rpart.plot')

# train the model
rpartmodel<- train(left~., data=model_db,
trControl=train_control, method="rpart")

# make predictions
predictions<- predict(rpartmodel,model_db)
model_db__tree<- cbind(model_db,predictions)

# summarize results
confusionMatrix<-
confusionMatrix(model_db__tree$predictions,model_db
__tree$left)
confusionMatrix
```

III.II Naïve's Bayes Algorithm

Naïve's Bayes is a probability-based classification theorem which is based on the Bayes Theorem. It can be used to predict the outcome of an occurring event with independent conditions. The presence of one feature in a class will not affect the presence of any other feature but even if sometime the features depend on each other, these properties individually contribute to the probability hence called as Naive. In this algorithm, The probability of end result is encoded in the model along with the probability of the evidence variables occurring given that the end result occurs, As in [3]. Naïve Bayes is eager classifier and fast executing algorithm used for real time predictions and also has higher success rate compared to other algorithms.

III.II.I Code Snippet:

```
#prediction
predictions<- predict(model2,model_db)
model2bind <- cbind(model_db,predictions)

#summarize
 confusionMatrix2<-
confusionMatrix(model2bind$predictions,model2bind$l
eft)
 confusionMatrix2
```

III.III Logistic Regression

This is a regression analysis that is used to set a statistical process for estimating the relationship between dependent variable and one or more independent variable. As in [4], We can analyze how the dependent variable is affected when one of the independent variable is changed and by fixing the other independent variables. This technique is used to predict the qualitative response.

III.III.I Code Snippet:

```
# train the model
install.packages('caTools')
model3 <- train(left~., data=model_db,
trControl=train_control, method="LogitBoost")

# make predictions
predictions<- predict(model3,model_db)
model3bind <- cbind(model_db,predictions)

# summarize results
confusionMatrix3<-
confusionMatrix(model3bind$predictions,model3bind$l
eft)
confusionMatrix3
```

IV. RESULTS

The diagram on the right is called correlation representation, A dot representation was used where blue represents positive correlation and red negative correlation. The larger dot represents the larger correlation and smaller dot represents the small correlation. If the matrix is symmetrical and that the diagonal are perfectly positively correlated because it shows the correlation of each variable with itself. The parameters used for constructing the correlation graph is derived by plotting histogram for all the parameters in the dataset against the parameter left. The parameters majorly influencing to move forward towards the result are chosen.

Reason for the employee to leave the organization from the correlation graph: From the several results obtained from the graph, the reason for an employee to leave the organization can be predicted. From the correlation graph it can be inferred that, since the dot representing the relation between left and the satisfaction level is orange in colour it shows that the value of left increases with the decrease in the value of satisfaction level and the dot representing the relation between left and number of projects and number of monthly hours respectively are in blue colour, it shows the value of left increases with the increase in the same. The employee who leaves the organisation on an average has a lesser value of satisfaction level, work on many projects, spend many hours in the company each month and are not promoted.

The confusion matrix and the accuracy figures of the three predictive models show that the predictive power is similar and robust. The logistic regression model performs prediction with an accuracy of about ninety-five percentage and a kappa of eighty-four percentage. Therefore, we choose this model to lay the actionable insights. A scatter plot is constructed by plotting the probability to leave of an employee against their performance from the dataset. Therefore, our area of interest is the top-right corner of the graph, since it consists of the employees whose performances are high and are highly probable to leave the organization.

Fig. 2 Correlation Graph

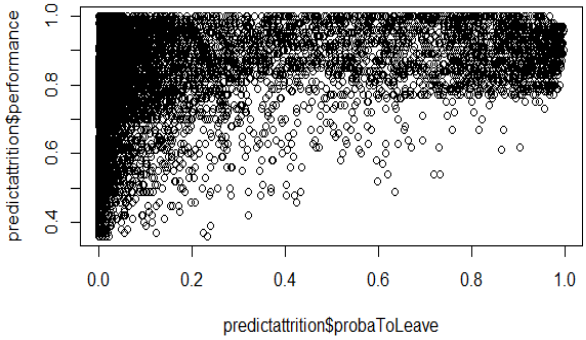

Fig. 3 Scatter Plot

V. CONCLUSION

In this analysis an employee data set consisting of 14999 Fourteen thousand nine hundred ninety-nine records of the employees were used who worked in an organization and left the organization for several reasons. The attributes which included job related

information were used majorly for performing classifications and predicting an analysis based on the classification. By using the R tool three predictive models and two rule-sets of the data set were generated. The predictive model with best performance was identified based on the accuracy rate of the intermediate result produced by the three models. The best predictive model was used to predict new cases of employee attrition.

VI. FUTURE WORK

This project has been executed based on the small dataset, but the real company dataset can be expected to be huge and with a greater number of attributes. This project can be tweaked to predict the result given the real and huge dataset.

1. The project can be extended by developing GUI, the prediction result can be presented in the human readable format with means of GUI which present the end user of this application with the information on "when and why the employee is predicted to leave the organization".

2. The application can be extended so that it provides the user with the retaining strategy that the organization has to use to minimize the employee attrition.

REFERENCES

[1] Alao D. & Adeyemo A. B, Analyzing Employee Attrition Using Decision Tree Algorithm, *Computing, Information Systems & Development Informatics, 4(1),* 2013, 17-28.

[2] Jiang Su and Harry Zhang, Fast decision tree learning algorithm, *American Association of Artificial Intelligence,* 2006.

[3] Mrutyunjaya Panda and Manas Ranjan Patra, Network Intrusion Detection Using Naïve Bayes, *International Journal of Computer Science and Network Security, 7(12),* 2007, 258-263.

[4] Miftar Ramosacaj1, Vjollca Hasani and Alba Dumi, Application of Logistic Regression in the Study of students' Performance Level, *Journal of Educational and Social Research, 5(3),* 2015, 239-244.

[5] Amir Mohammad Esmaieeli Sikaroudi, RouzbehGhousi and Ali EsmaieeliSikaroudi, A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing), *Journal of Industrial Systems and Engineering, 8(4),* 2015.

[6] S.J.Russel, and Norvig, "Artificial Intelligence: A modern approach (International edition), *Pearson US imports & PHIPES*, Nov 2002.

[7] Miftar Ramosacaj1, Vjollca Hasani and Alba Dumi, Application of Logistic Regression in the Study of Students' Performance Level, *Journal of Educational and Social Research, 5(3),* 2015, 239-244.

[8] Prince Austin and RB. Mohanty, A Diagnostic Study of Employee Attrition in an Indian Automotive Company , *International Journal of Indian Culture and Business Management, 5(5),* 2012.