Why is HE leaving?

Ravali Gampa, Under the guidance: Dr. Meiliu lu

***Abstract*— For a sustainable growth of any organization it needs to have consistency source of the resources. These resources can be human resources or non-human resources. Most important of all these resources would be the man power. No organization wishes to lose its talented group. We train our model with human resource analysis data set and predict if the employee is leaving the company. The deciding factors include Satisfaction level, Number of Projects undertaken, Last Evaluation result, Time Spent in the company and many more. Our aim is to build a model using Logistic Regression, SVM, Decision Tree, Random Forests for Prediction about the employee stay in the company. We find the accuracy of each model using confusion matrix. At the end, we compare the classification performance of the models using ROC curves. Our analysis show that random forest is the most suitable technique for the dataset.**

***Index Terms*—Employee, Satisfaction, Logistic Regression, SVM,ROC**

I. INTRODUCTION

Why is HE leaving? (Here He refers to the employee and the scenario refers to the instance when he is quitting his job). Before understanding the factors that lead the employee to quit the job, the reason or motivation to study this scenario is discussed below.

On a broad perspective, the resources needed for the growth of an organization is categorized as human resources and non-human resources. Human resources are in a way considered more important than non-human resources as these resources help in building the non-human resources. Human resources are important to organizations in various specific areas, ranging from strategic planning to company image.

There are few previous works which were discussed on the importance of the human resources to an organization. Like N. Costa [1] cited the importance of training and human resources management in the organization where he mentioned the need to increase awareness of the objectives set in professional training and development in his paper.

There are some metrics or measures to understand if an organization is losing out of human resources so that organization can take some safety precautions in sustaining their resources. Q. Qian [2] has written a paper which use qualitative and quantitative methods to study on early-warning management for human resource crisis systematic.

As these resources are of great importance to an organization. If the organization starts losing them it would be a great loss to the organization. So, this paper discusses on a tool which helps in prediction of the reasons why an employee is leaving the company. This prediction helps the organization to take safety measures to protect them.

We predict and analyze the factors such include Satisfaction level, Number of Projects undertaken, Last Evaluation Result, Time Spent in the company and many more of an employee in an organization. We analyze the human resources dataset and predict if the employee is leaving the company by building the models using machine learning algorithms on this data.

The paper is structured in the following way: Section 2 presents Literature Survey Section 3 emphasis on the tool used to model the data Section 4 introduces the data collection, followed by data preprocessing in Section 5. Section 6 discuss on the data visualization and the feature selection. Section 7 presents different machine learning algorithms applied for the model. Section 8 presents the comparison of the 4 models. Section 9 will conclude with some future scope.

II. LITERATURE SURVEY

Effective human resource management practices can promote the building, forming, innovating and developing of enterprise culture. Enterprises culture is a technique mentioned in [3] which helps in effective resource management. A research from Y.Li[4] was based on previous studies, combined with their own university Case Analysis, Analysis of the best domestic and international dimensions of human resource management literature and many human resource management in Chinese Universities based on the actual situation, make the most of University Human Resource Practice good dimension for the college to improve organizational performance to provide some references and ideas.

There were few works which were concerned about the human resources on a specific domain. Like X. Bing and J. Yingbo [5] compared the status of human resources in construction industry between China and America from the

.

labor productivity, the ratio of labor cost in total cost, the workers' average wage. Even J. Yang and Y. Sun [6] talked about the Human Resource Management in Enterprise Based on the Innovative Incentive.

III. TOOL USED

Analyzation of the dataset, building the model using the dataset and prediction of the results is done with help of tool called R-studio. R-Studio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

IV. DATA COLLECTION

The Dataset used for this work is collected from Kaggle [7]. It is a platform for predictive modelling and analysis. This dataset has rich source about the possible reasons that lead the employee to leave their job. Each cause is described as a column attribute in the dataset. It contains around 10 columns and 14980 rows.

V. DATA PREPROSSING

A. *Checking for Missing values*

Initially the dataset is tested to check if it contains any null or missing values using a function in R. But the results turned to be negative as the dataset did not have any missing values.

B. *Handling of Categorical Values*

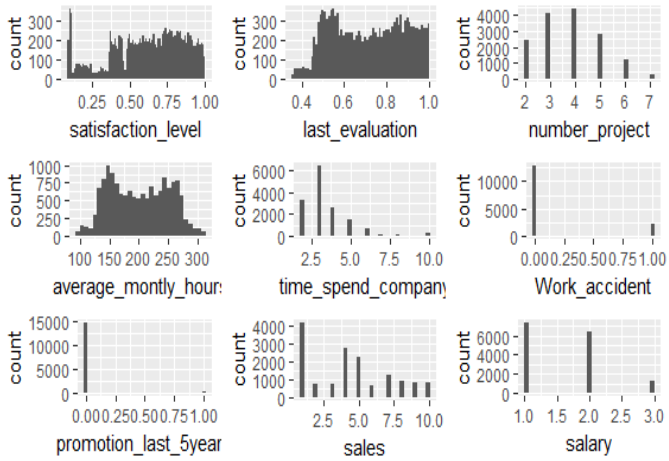
For the easy handling of data and application of the data to build the model, some of the categorical values needs to be converted to numerical values. The Dataset has eight attributes which have numerical data and two attributes which have categorical data Then attributes which have textual data namely, salary attribute and sales attribute are converted to numerical data. There are about 10 different types of values in the sales attribute, these ten types are mapped to numbers from 1 to 10. Salary attribute is mapped to numbers from 1 to 3.

VI. DATA VISULAIZATION AND FEATURE SELECTION

A. *DATA VISUALIZATION*

The Dataset is visualized as set of histograms for a better understanding of the data. A total of nine histograms are displayed. In each histogram, the count of number of employees leaving their job is represented on Y-axis and other attributes are represented on X-axis. This histogram helps in finding the correlation among the count of the employees left and reasons or attributes that lead to this scenario. The histograms obtained can be visualized in figure1.

FIGURE1: SET OF HISTOGRAMS



B. *FEATURE SELECTION*

Feature selection helps in choosing the most important features or attributes responsible in prediction of the values. Through the visualization of the features it is clearly understood that attributes ‘Promotion in the last five years’ and ‘Work accident’ do not contribute significantly. Hence these attributes are assigned NULL values for a higher prediction accuracy. Finally, this step produces a feature set containing important eight attributes. These attributes are summarized in the table1.

TABLE 1: FEATURE SET

FEATURE ID	FEATURE NAME
1	Satisfaction level
2	Last evaluation
3	Average monthly hours
4	Time spent at the company
5	Sales
6	Salary
7	Whether the employee has left or not
8	Number of projects

VII. IMPLEMENTATION

Initially, the dataset is divided into training data and testing data. Then the implementation is carried on by building the models using the Training data and prediction is done by testing the model using the Test data. The implementation is done with techniques like logistic regression, SVM, decision trees, random forests.

A. *Logistic Regression:*

It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function [8]. For this dataset, 'True' represents the case when the employees have left the job and 'False' represents the case when the employees are continuing with their job. On applying this technique, it can be figured which employees left their job.

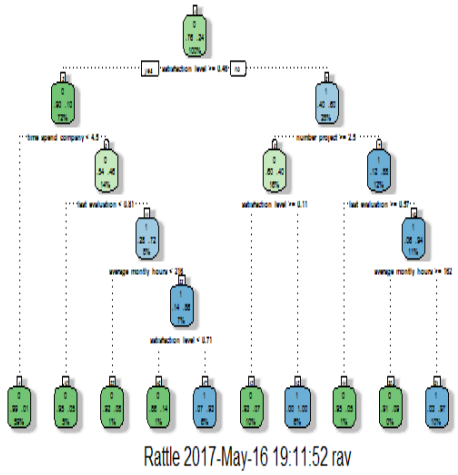
B. *Support Vector Machines*

Support Vector Machines (SVMs) are supervised learning methods used for classification and regression tasks that originated from statistical learning theory [9]. In this work, SVM is considered as a classification method, it constructs a hyperplane that separates two classes of data with the largest possible distance to data points. For this dataset, one side of hyper plane represents all the cases when the employees have left the job and the other represents the cases when the employees are continuing with their job. On applying this technique, it can be figured which employees left their job.

C. *Decision Trees*

Decision tree is a classification technique where it is graph to represent choices and their results in form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or condition [10]. On applying this technique to the data set it was seen that Satisfaction level was the most important factor in deciding if the employee wants to continue the job or not. The decision tree obtained is shown in figure 2.

FIGURE 2: PLOT OF DECISION TREE



D. *Random forests*

Random forests [11][12] are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. In most cases, Random Forests obtain the higher performance and accuracy in comparison to decision trees [8]. For the very same reason Random forests was used in this work

.

VIII CONCLUSIONS AND RESULTS OBTAINED

A. *Metrix to find the accuracy*

Confusion matrix is used to predict the accuracy of the four models built during the implementation phase. A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. The confusion matrix was used for performance evaluation. The confusion matrix has the following four values namely:

1. True positive-TP (actually left the company and classified as left the company)
2. True negative- TN (actually present in the company and classified as present in the company)
3. False positive- FP (actually present in the company but classified as left the company)
4. False negative - FN (actually left the company but classified as present in the company)

The accuracy of the classifier is calculated using

[(TP + FP)/ (TP + FP + FN + FP)]*100

The accuracy results obtained are tabulated in table2.

TABLE 2: ACCURACY OF MODELS

MODEL NAME	**ACCURACY OBTAINED**	
	Split ratio=0.7	Split ratio=0.8
Logistic Regression	77.26%	77.16%
SVM	96.3%	96.53%
Decision Trees	96.6%	97.26%
Random Forests	98.4%	98.93%

B. *Decision tree rules*

These are the different rulses which are obtained from the

decision tree:

1. Satisfaction level appears to be the most important attribute. If it is above 0.46 employee is much more likely to stay (which is what observed above in figure2)

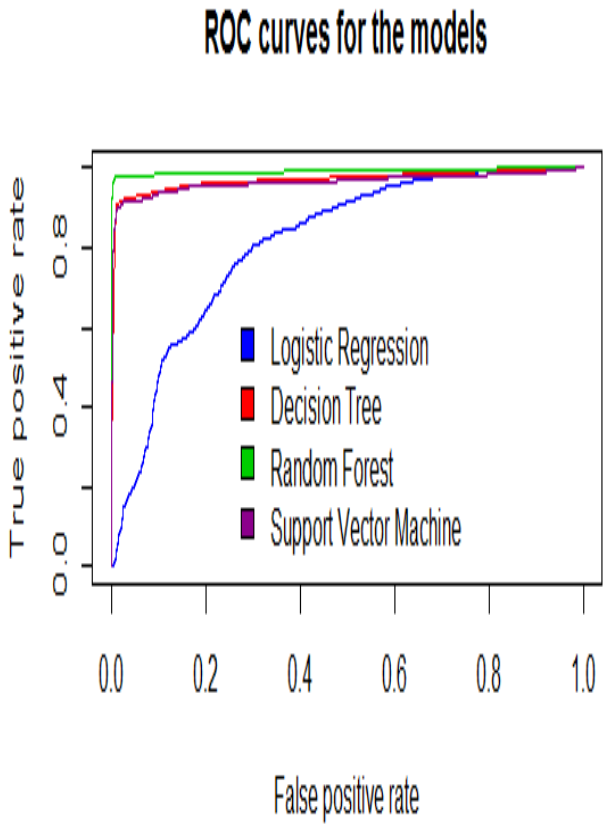
2. If employee have low satisfaction, the number of projects becomes important factor. If employee is working on more number of projects more chances to remain.

3. If the employee has been at the company for less than 4.5 years, and score over 81% on the last evaluation, employee is very likely to leave. And, it appears as if the "decider" is monthly hours over 216.

C. Comparison of Models

Models are compared with one another through the ROC curves. ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The ROC curve for the models is displayed in figure3.

FIGURE 3: ROC CURVES FOR THE MODELS



VIII FUTURE WORK

1. Correlation between the attributes can be studied to have a deeper understanding of the most importance features or attributes and even for a better prediction accuracy.

2. A more specific set of attributes can be collected depending upon the sector or industry the employee is working. This helps in obtaining the specific factors which lead to this scenario.

IX REFERENCES

[1] N. Costa, "Training, innovation and human resources - Why are they so important in business?," *2014 11th International Conference on Live Maintenance (ICOLIM)*, Budapest, 2014, pp. 1-5.doi: 10.1109/ICOLIM.2014.6934361

[2] Q. Qian, "Research on Early-warning Management for Crisis of Human Resource: Knowledge Frame and Evaluation Index System," *2010 International Conference on E-Business and E-Government*, Guangzhou, 2010, pp. 4374-4377. doi: 10.1109/ICEE.2010.1099

[3] W. Hao and Z. Feng, "Study on Structural Features of University Teacher Need and Teacher Professional Development," 2009 First International Workshop on Education Technology and Computer Science, Wuhan, Hubei, 2009,pp.399-403.doi:10.1109/ETCS.2009.351

[4] Y. Li, "Research on analysis of human resource management," *2010 2nd IEEE International Conference on Information and Financial Engineering*, Chongqing, 2010, pp. 894-897.doi:10.1109/ICIFE.2010.5609497

[5] X. Bing and J. Yingbo, "The Comparison of Human Resources in Construction Industry between China and America," *2010 International Conference on E-Business and E-Government*, Guangzhou, 2010, pp. 1033-1036. doi:10.1109/ICEE.2010.267

[6] J. Yang and Y. Sun, "Research of the Human Resource Management in Enterprise Based on the Innovative Incentive," *2008 International Seminar on Business and Information Management*, Wuhan, 2008, pp. 445-448. doi:10.1109/ISBIM.2008.160

[7] Ludovic Benistant, "human activity recongization", gathered dataset and deposited in Kaggle,2017

[8] T. Haifley, "Linear logistic regression: an introduction," IEEE International Integrated Reliability Workshop Final Report, 2002., 2002, pp. 184-187. doi: 10.1109/IRWS.2002.1194264Basic format for books

[9] V. Vapnik, Statistical learning theory. Wiley, New Tork (1998)

[10] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660-674, May/Jun 1991. doi: 10.1109/21.97458

[11] Ho, Tin Kam (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **20** (8): 832–844. doi:10.1109/34.709601

[12] M. V. Datla, "Bench marking of classification algorithms: Decision Trees and Random Forests - a case study using R," *2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15)*, Bangalore, 2015, pp. 1-7.doi: 10.1109/ITACT.2015.7492647

[13] Huan Liu and Lei Yu, "Toward integrating feature selection algorithms for classification and clustering," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no.4, pp.491-502,April2005.doi: 10.1109/TKDE.2005.66

[14] V. U. B. Challagulla, F. B. Bastani, I-Ling Yen and R. A. Paul, "Empirical assessment of machine learning based software defect prediction techniques," *10th IEEE International Workshop on Object-Oriented Real-Time Dependable Systems*, 2005, pp. 263-270. doi: 10.1109/WORDS.2005.32

[15] “IMPLEMENTATION OF CODE” is attached as Appendix1

APPENDIX 1: IMPLEMENTATION OF THE CODE

```
################################################################################
        # LOADING THE REQUIRED PACKAGES #
 ################################################################################

library(ggplot2)
library(readr)
library(rpart)
library(rattle)
library(ROCR)
library(randomForest)
library(gridExtra)
library(reshape)
library(caTools)
library(e1071)
library(caret)
library(ROCR)


 ################################################################################
          # INPUTING THE DATA FILE #
 ################################################################################


df <- read.csv("HR_comma_sep.csv")
# Look at the data
sum(is.na(df))

################################################################################
        # PREPROSSING OF THE INPUT DATA#
################################################################################


sales <- unique(df$sales)
df$sales <- as.numeric(1:10)[match(df$sales, sales)]
df$salary <- as.numeric(1:3)[match(df$salary, c('low', 'medium', 'high'))]


################################################################################
        # VISUALIZATION OF INPUT DATA
################################################################################


p1 <- qplot(satisfaction_level, data=df, geom="histogram", binwidth=0.01)
p2 <- qplot(last_evaluation, data=df, geom="histogram", binwidth=0.01)
p3 <- qplot(number_project, data=df, geom="histogram")
p4 <- qplot(average_montly_hours, data=df, geom="histogram")
p5 <- qplot(time_spend_company, data=df, geom="histogram")
p6 <- qplot(Work_accident, data=df, geom="histogram")
p7 <- qplot(promotion_last_5years, data=df, geom="histogram")
p8 <- qplot(sales, data=df, geom="histogram")
p9 <- qplot(salary, data=df, geom="histogram")
library(gridExtra)
grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, p9, ncol = 3, nrow = 3)
```

```
#############################################################################
                # FEATURE SELECTION #
#############################################################################

df$Work_accident <- NULL
df$promotion_last_5years <- NULL


#############################################################################
                  #SPLITTING OF INPUT DATA#
#############################################################################

set.seed(300)
df$sp <- sample.split(df$left, SplitRatio=0.8)
train <- subset(df, df$sp==TRUE)
test <- subset(df, df$sp==FALSE)

#############################################################################
                   # LOGISTIC REGRESSION#
#############################################################################


# let us first start with logistic regression
# Train the model using the training sets and check score
   model_glm <- glm(left ~ ., data = train, family='binomial')

# Predict Output of test data
   predicted_glm <- predict(model_glm, test, type='response')
   predicted_glm <- ifelse(predicted_glm > 0.5,1,0)

# Confusion matrix of Logistic regression
   table(test$left, predicted_glm)

# Accuracy of model
   mean(predicted_glm==test$left)


#############################################################################
               # SUPPORT VECTOR MACHINE #
#############################################################################

# Train the model using the training sets and check score
   model_svm <- svm(left ~ ., data=train)

# Predict Output of test data
   predicted_svm <- predict(model_svm, test)
   predicted_svm <- ifelse(predicted_svm > 0.5,1,0)

# Confusion matrix of SVM
   table(test$left, predicted_svm)

# Accuracy of SVM
   mean(predicted_svm==test$left)


#############################################################################
                     # DECISION TREES #
#############################################################################
```

```
# Let us try decision trees
# Train the model using the training sets and check score
   model_dt <- rpart(left ~ ., data=train, method="class", minbucket=25)

# View decision tree plot
   fancyRpartPlot(model_dt)

# Predict Output of test data
   predicted_dt <- predict(model_dt, test, type="class")

# Confusion matrix of decision tree
   table(test$left, predicted_dt)

# Accuracy of decision tree
   mean(predicted_dt==test$left)


##############################################################################
                  #RANDOM FORESTS#
##############################################################################


# We shall do random forests with 200 trees
# Train the model using the training sets and check score
   library(randomForest)
   model_rf <- randomForest(as.factor(left) ~ ., data=train, nsize=20, ntree=200)

# Predict Output of test data
   predicted_rf <- predict(model_rf, test)

# Confusion matrix of random forest
   table(test$left, predicted_rf)

# Accuracy of random forest
   mean(predicted_rf==test$left)


##############################################################################
                  #COMPARSION OF THE METHODS#
##############################################################################


# Tuning the parameters increases the accuracy of SVM to 97.53%
# Let us plot the ROC curves for all the models
# Logistic regression
   library(ROCR)
   predict_glm_ROC <- predict(model_glm, test, type="response")
   pred_glm <- prediction(predict_glm_ROC, test$left)
   perf_glm <- performance(pred_glm, "tpr", "fpr")

# Decision tree
   predict_dt_ROC <- predict(model_dt, test)
   pred_dt <- prediction(predict_dt_ROC[,2], test$left)
   perf_dt <- performance(pred_dt, "tpr", "fpr")

# Random forest
   predict_rf_ROC <- predict(model_rf, test, type="prob")
   pred_rf <- prediction(predict_rf_ROC[,2], test$left)
```

```
    perf_rf <- performance(pred_rf, "tpr", "fpr")

# SVM
    predict_svm_ROC <- predict(model_svm, test, type="response")
    pred_svm <- prediction(predict_svm_ROC, test$left)
    perf_svm <- performance(pred_svm, "tpr", "fpr")

# Area under the ROC curves
    auc_glm <- performance(pred_glm,"auc")
    auc_glm <- round(as.numeric(auc_glm@y.values),3)
    auc_dt <- performance(pred_dt,"auc")
    auc_dt <- round(as.numeric(auc_dt@y.values),3)
    auc_rf <- performance(pred_rf,"auc")
    auc_rf <- round(as.numeric(auc_rf@y.values),3)
    auc_svm <- performance(pred_svm,"auc")
    auc_svm <- round(as.numeric(auc_svm@y.values),3)
    print(paste('AUC of Logistic Regression:',auc_glm))
    print(paste('AUC of Decision Tree:',auc_dt))
    print(paste('AUC of Random Forest:',auc_rf))
    print(paste('AUC of Support Vector Machine:',auc_svm))


# Plotting the three curves
    plot(perf_glm, main = "ROC curves for the models", col='blue')
    plot(perf_dt,add=TRUE, col='red')
    plot(perf_rf, add=TRUE, col='green3')
    plot(perf_svm, add=TRUE, col='darkmagenta')
    legend('bottom', c("Logistic Regression", "Decision Tree", "Random Forest", "Support Vector Machine"), fill =
c('blue','red','green3','darkmagenta'), bty='n')
```