**Predicting Employee Attrition: Human Resource Concern**

A picture containing drawing

Description automatically generated

**Course**: [ALY6015 20975 Intermediate Analytics](https://northeastern.blackboard.com/webapps/blackboard/execute/courseMain?course_id=_2605149_1)

**Instructor:** Dr. Roy Wada

**Submission date:** Feb 15, 2021

By,

Madhav Khanzode

Shruti Avinash Pawar

Suvarna Suresh Tambe

**Introduction:**

Employee attrition is the frightful thing for organization to face inevitably. Predicting an attrition and keeping attrition under limit is the important task of human resource partners. In attrition, an employee walks out with valuable business knowledge and productivity that has been built over the time into the organization. Employee attrition has negative impact on the organization profit and loss. Using machine learning algorithms and advancements it is possible to get accurate predictive performance and insights about employee attrition. ("RPubs - Predicting employee attrition - Who will quit &amp; when?", 2020) We will be using logistic regression model, hypothesis testing and decision tree to predict the employee attrition. In this report, our objective will be to investigate how the parameters in our employee data such as satisfaction level of employee, last evaluation rating of employee, employee promotion of last\_5 years, etc can impact on the attrition of employee and using these parameter building an accurate model that can predict if any employee will leave the organization or not. The tool that will be used for this analysis is R and R studio.

**Analysis:**

The dependent variable in our dataset is Attrition which is a binary variable with 0 being no and 1 being yes. There are 9 independent variables and below is the descriptive summary of those variables:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1:** Descriptive Summary Statistics of Employee Data | | | | | | | | | | |
| **Heads** | **Satisfaction Level** | **Last Evaluation Rating** | **Projects Worked On** | **Average Monthly Hours** | **Time Spent in Company** | **Work Accidents** | **Promotion Last 5 Years** | **Dept** | **Salary** | **Attrition** |
| Minimum | 0.90 | 3.60 | 2.00 | 96.00 | 2.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Quarter 1 | 4.40 | 5.60 | 3.00 | 160.00 | 3.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Median | 6.50 | 7.20 | 4.00 | 204.00 | 3.00 | 0.00 | 0.00 | 3.00 | 1.00 | 0.00 |
| Mean | 6.14 | 7.17 | 4.22 | 205.00 | 3.50 | 0.14 | 0.02 | 3.45 | 0.60 | 0.24 |
| Quarter 3 | 8.20 | 8.70 | 5.00 | 249.00 | 4.00 | 0.00 | 0.00 | 6.00 | 1.00 | 0.00 |
| Max | 10.00 | 10.00 | 8.00 | 320.00 | 10.00 | 1.00 | 1.00 | 9.00 | 2.00 | 1.00 |
| N | 25491 | 25491 | 25491 | 25491 | 25491 | 25491 | 25491 | 25491 | 25491 | 25491 |
| SD | 2.49 | 1.71 | 1.32 | 50.18 | 1.46 | 0.35 | 0.14 | 2.81 | 0.64 | 0.42 |

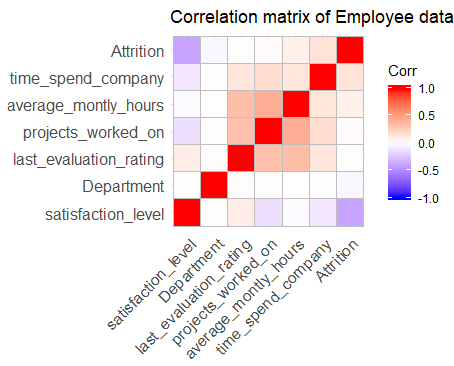
From table 1 we can see that there is a lot of deviation in satisfaction level data and the mean satisfaction of employees is 6.14 which is quite lower than 10, the maximum satisfaction at workplace. Also, we will not be referring the descriptive statistics for variables work accident and promotion last 5 years as these are binary variables and Department and Salary as these are categorical variables. Salary 0 is low , 1 is medium and 2 is high and for Department, 0 is sales, 1 accounting, 2 is hr, 3 is technical, 4 is support, 5 is management, 6 is IT, 7 is product management, 8 is marketing and 9 is R&D. There is not much of a difference between the mean and the median of the first three variables in table 1 thus we can say that there are no outliers in this data. The next step will be to check the correlation between the variables. We have used the below code for a correlation matrix to check correlation between certain variables from the data:

corr1 <- cor(mydata[c("satisfaction\_level","last\_evaluation\_rating","projects\_worked\_on","average\_montly\_hours","time\_spend\_company","Department","Attrition")])

corr1

ggcorrplot(corr1, title = "Correlation matrix of Employee data",hc.order=TRUE)

Following is the output of the code:

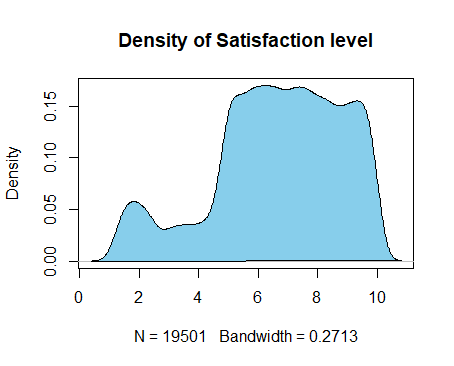


**Figure 1:** Correlation Matrix of Employee Data

From figure 1 it is certain that there is a significant negative correlation between attrition and satisfaction level of the employee. The correlation is valid as higher the satisfaction level, the employee will stay in the organization and lower the satisfaction level the employee will leave for a better opportunity. A substantial positive correlation can also be seen between average monthly hours and projects worked on, last evaluation rating and projects worked on and average monthly hours and projects worked on. This correlation is also valid as the evaluation rating will be higher if the number of projects worked on is high. Department does not have to seem any good correlation with any of the variables. We have used the below code to plot the density plot for Satisfaction level of employees whose attrition is 0:

plot(density(att0$satisfaction\_level),main="Density of Satisfaction level")

polygon(density(att0$satisfaction\_level),col="skyblue")



**Figure 2:** Density plot of satisfaction level of employees with Attrition ‘0’

In figure 2 we can see that greater number of employees with attrition zero are above the satisfaction level 6 and as per the correlation matrix, satisfaction level is negatively correlated to attrition. Thus, our claim is that employee with satisfaction level above 6 will not leave the organization. We will use T test to check for the validity of our claim and the sample data will be the data where attrition is 1. The sample will be picked from subset of the data where attrition is 1 and the sample size will be 1000. Below is the summary table of the sample selected:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2**: Summary statistics of T test Sample | | | | | | | |
| Heads | Minimum | Quarter 1 | Median | Mean | Quarter 3 | Maximum | N |
| Value | 0.90 | 1.10 | 4.10 | 4.47 | 7.40 | 9.20 | 1000 |

Null Hypothesis: Attrition happens when satisfaction level is greater than or equal to 6

Ho: µ > = 6

Alternative Hypothesis: Attrition happens when satisfaction level is less than 6

Ha: µ < 6

Below code was used to run the t test

t.test(mysample$satisfaction\_level[200:1200], alternative = "less", mu = 6)

Below is the output of the code:

## One Sample t-test

## data: mysample$satisfaction\_level[200:1200]

## t = -18.496, df = 1000, p-value < 2.2e-16

## alternative hypothesis: true mean is less than 6

## 95 percent confidence interval:

## -Inf 4.579003

## sample estimates:

## mean of x

## 4.44016

The p-value is less than 0.05 and 6 (hypothesized mean) does not fall in the confidence interval which is 4.27 to 4.60. Thus, there is significant evidence to reject the null hypothesis which states that attrition happens even when satisfaction level is greater than or equal to 6.

Now we will evaluate which model from logistic regression and decision tree is the best one to predict attrition. To do so we have split the Employee data into train and test with 70% of the data in train and 30% in test. On the train data a model will be fit and on the test data, the prediction will be made. The predict code in R gives an output for predicted y which will be calculated with the actual y in the test data to determine model accuracy. Using the below code, we will now run a logistic regression model and see which variables have a significant impact on Attrition:

logitTmodel<-glm(Attrition~satisfaction\_level+last\_evaluation\_rating+projects\_worked\_on+average\_montly\_hours+time\_spend\_company+as.factor(Work\_accident)+as.factor(promotion\_last\_5years)+as.factor(Department)+as.factor(salary),family="binomial", data = trainSet)

Note that we have used as.factor for certain variables in the code as these are categorical variables. Below table is the output of the code:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 3:** Logistic Regression Summary on Training data set | | | | |
| **Heads** | **Coefficients** | **Standard Error** | **P Value** |  |
| Intercept | 0.467 | 0.116 | 0.0000555 | \*\*\* |
| Satisfaction Level | -0.410 | 0.009 | < 2e-16 | \*\*\* |
| Last Evaluation Rating | 0.053 | 0.014 | 0.0000046 | \*\*\* |
| Projects Worked on | -0.287 | 0.018 | < 2e-16 | \*\*\* |
| Average Monthly Hours | 0.005 | 0.001 | < 2e-16 | \*\*\* |
| Time Spent in Company | 0.268 | 0.014 | < 2e-16 | \*\*\* |
| Work Accidents 1 | -1.485 | 0.080 | < 2e-16 | \*\*\* |
| Promotion Last Year 1 | -1.768 | 0.266 | 3.24E-11 | \*\*\* |
| Department 1 | 0.123 | 0.092 | 0.17906 |  |
| Department 2 | 0.260 | 0.094 | 0.00548 | \*\* |
| Department 3 | 0.144 | 0.060 | 0.01627 | \* |
| Department 4 | 0.113 | 0.065 | 0.08109 | . |
| Department 5 | -0.499 | 0.127 | 7.88E-05 | \*\*\* |
| Department 6 | -0.194 | 0.082 | 0.01832 | \* |
| Department 7 | -0.076 | 0.092 | 0.4066 |  |
| Department 8 | 0.018 | 0.093 | 0.84783 |  |
| Department 9 | -0.483 | 0.107 | 6.70E-06 | \*\*\* |
| Salary 1 | -0.488 | 0.042 | < 2e-16 | \*\*\* |
| Salary 2 | -1.86 | 0.115 | < 2e-16 | \*\*\* |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’

AIC: 15301

Logistic regression is used to fit the model as our dependent variable Attrition is a binary categorical variable which has values 0 – No attrition and 1 – Yes attrition. In R, glm() function is used to fit the model. As per the logistic regression model output in table 3 all variables apart from department are significantly related to Attrition. Now that the logistic model is ready on the training dataset, we have used the predict code to predict values for attrition. (Alice,2020)

Below is the confusion matrix of the prediction:

**Table 4:** Confusion Matrix table for Logistic regression Model

|  |  |  |
| --- | --- | --- |
| **Attrition** | **Actual No** | **Actual Yes** |
| **Predict No** | 5318 (TN) | 1204 (FP) |
| **Predict Yes** | 409 (FN) | 579 (TP) |

TN means True Negative, TP is true positive, FN is False negative, and FP is false positive. The classification accuracy is calculated by dividing the true predictions with the total predictions. (Brownlee, 2020) In the above case it will be TN+TP/TN+TP+FN+FP. (“Decision Tree in R with Example”,2020) The accuracy for the logistic model is 78.5% which is very low. As we can see in table 4, there are a lot of false positive values. There are 409 employees that were predicted as attrition 1 however the attrition is 0. Also, there are 1204 employees who left the organization, but the model predicted as attrition 0. Therefore, logistic regression model is not a perfect fit for this data set as there are a lot of type 1 and type 2 errors. Using the same train and test dataset, we fit a decision tree model and below is the code used in R to do so:

treeFit <- rpart(Attrition~., method="class", data=trainSet)

print(treeFit)

plot(treeFit)

rpart.plot(treeFit)

A screenshot of a cell phone

Description automatically generatedBelow was the output of the decision tree code:

**Figure 3:** Decision tree to predict Attrition (0-No, 1- Yes)

The decision tree plot in figure 3 Is the model output that was fit on the training data. The root node in the model is satisfaction level based on which attrition yes or no is predicted. Using the predict function in R, we found the predicted attrition and compared the same with the actual attrition. Below is the confusion matrix used to calculate accuracy of the decision tree model:

**Table 5:** Confusion Matrix table for Decision Tree Model

|  |  |  |
| --- | --- | --- |
| **Attrition** | **Actual No** | **Actual Yes** |
| **Predict No** | 5653 (TN) | 74 (FP) |
| **Predict Yes** | 162 (FN) | 1621 (TP) |

If we compare table 4 and table 5, we can see that in table 5 that is the decision tree confusion matrix table, type 1 and type 2 errors are much lesser than that in table 4. The accuracy of decision tree which is 96.85% is calculated using below code:

accuracy <- sum(diag(table\_mat)) / sum(table\_mat)

(“Decision Tree in R with Example”,2020)

**Conclusion:**

To recapitulate, in this report we have identified that satisfaction level is negatively correlated with Attrition and we have T test evidence that employee with satisfaction level above 6 will stay in the organization. It is evident that decision tree is the best classifier for this employee data as it gives an accuracy of 96.5% as opposed to 78.5% of logistic model. The decision tree accuracy can further be improved by bootstrapping or Random Forest. (Gupta, 2020). Once the organization identifies the chunk of employees that are classified as potential leavers using the decision tree model, the human resource department can then focus on the variables that will help to retain the employee. In case of our data satisfaction level is the most significant and correlated with Attrition. A linear regression or Hypothesis testing can be done to identify the variables that most affect the satisfaction level of the employee. These variables can be focused on to increase the satisfaction level and retain the employee.

**Reference:**

1. RPubs - Predicting employee attrition - Who will quit &amp; when?. (2020). Retrieved 2 February 2020, from <https://rpubs.com/heruwiryanto/PEA>
2. Alice, M. (2020). How to Perform a Logistic Regression in R. Retrieved 2 February 2020, from <https://datascienceplus.com/perform-logistic-regression-in-r/>
3. Brownlee, J. (2020). What is a Confusion Matrix in Machine Learning. Retrieved 15 February 2020, from <https://machinelearningmastery.com/confusion-matrix-machine-learning/>
4. Decision Tree in R with Example. (2020). Retrieved 11 February 2020, from <https://www.guru99.com/r-decision-trees.html>
5. Gupta, P. (2020). Boosting the accuracy of your Machine Learning models. Retrieved 15 February 2020, from <https://towardsdatascience.com/boosting-the-accuracy-of-your-machine-learning-models-f878d6a2d185>
6. Employee Attrition. (2020). Retrieved 2 February 2020, from <https://www.kaggle.com/HRAnalyticRepository/employee-attrition-data>.