Predictive Analysis of Employee Attrition

Sharan Devarshetty

15/06/2022

## Abstract

For any given association, employees are a significant resource. Therefore, when employees decide to quit jobs unexpectedly in any given institution, it brings in immense expenses that might not be accounted for. This is because hiring new employees always makes any institution incur a lot of time and finances. In addition, there needs to be some training for the newly hired employees to ensure they make the institution productive. This training is time-consuming, leading to the organization incurring losses. Therefore, this project has developed a machine learning model that will be essential to the organizations in predicting employee attrition rates dependent on the dataset obtained from HR analytics retrieved from the Kaggle.com website. The project predicts reasons for employees to leave an organization and employee attrition. The objective was to identify the various reasons why most and best-experienced employees are most likely to leave an institution to ensure that the institution determines the areas where they lag. The project used the Logistic regression model. The results of this project are helpful to the Human Resource Management of the organization as they can identify reasons for employee attrition and implement the best strategies to ensure they retain already existing employees before they start looking for new employees. The project's main objective will be to identify the departments and job titles with the highest employee attrition. In addition, the project will work on the research question of whether age is one of the contributing factors to employee attrition.

## Introduction

Employee attrition for any organization involves a manpower decrease when employees willingly decide to leave or resign by the Human Resource of the organization. There is always high employee turnover when there is a high employee attrition rate. This, in return, causes the organization to incur huge expenditures on human resources in recruiting new employees, training the newly hired employees, and maintaining the performance of the employees to the organization's standards. Therefore, for any organization to ensure that they reduce employee attrition, they have made efforts to improve employee morale and provide them with conducive working conditions. For any organization, there are different explanations for where an employee can leave their job. With the prediction of employee attrition, it will be essential for the organization to take quicker actions in their strategies and policies internally where the employees with a higher risk of leaving the company can be provided with alternative recommendations like proper training or increment in their pay to ensure the probability of leaving have been reduced.

## Literature Review

Mitkees et al. (2017) provided a proposal for solving the customer churn problem identified through developing models using procedures like the association for detection, classification for prediction, and clustering for detection. In another project by Khare et al. (2011), the researchers created a logistic regression method dependent on employee data in building a system that was essential in predicting employee attrition. They were able to identify the reasons for high employee attrition and an action plan taken to minimize the employee churn risk. In their research, Coussement & Van den Poel (2008) implemented a support vector machine method to predict employee churn. From their study, it was identified that with noisy marketing data, there is a good generalization performance for the supporting vector machines. In their study, Omar Ali (2017) identified that there was a higher likelihood for the employees to leave an institution as a result of disagreement with the senior manager. Several major factors were identified that affect the employees' attrition.

## Theory

H1: There is higher employee attrition for older employees than for younger ones.

## Data

The dataset has been downloaded from <https://www.kaggle.com/HRAnalyticRepository/employee-attrition-data/discussion>.

## EmployeeID recorddate\_key birthdate\_key orighiredate\_key terminationdate\_key

## 1 1318 12/31/2006 0:00 1/3/1954 8/28/1989 1/1/1900

## 2 1318 12/31/2007 0:00 1/3/1954 8/28/1989 1/1/1900

## 3 1318 12/31/2008 0:00 1/3/1954 8/28/1989 1/1/1900

## 4 1318 12/31/2009 0:00 1/3/1954 8/28/1989 1/1/1900

## 5 1318 12/31/2010 0:00 1/3/1954 8/28/1989 1/1/1900

## 6 1318 12/31/2011 0:00 1/3/1954 8/28/1989 1/1/1900

## age length\_of\_service city\_name department\_name job\_title store\_name

## 1 52 17 Vancouver Executive CEO 35

## 2 53 18 Vancouver Executive CEO 35

## 3 54 19 Vancouver Executive CEO 35

## 4 55 20 Vancouver Executive CEO 35

## 5 56 21 Vancouver Executive CEO 35

## 6 57 22 Vancouver Executive CEO 35

## gender\_short gender\_full termreason\_desc termtype\_desc STATUS\_YEAR STATUS

## 1 M Male Not Applicable Not Applicable 2006 ACTIVE

## 2 M Male Not Applicable Not Applicable 2007 ACTIVE

## 3 M Male Not Applicable Not Applicable 2008 ACTIVE

## 4 M Male Not Applicable Not Applicable 2009 ACTIVE

## 5 M Male Not Applicable Not Applicable 2010 ACTIVE

## 6 M Male Not Applicable Not Applicable 2011 ACTIVE

## BUSINESS\_UNIT

## 1 HEADOFFICE

## 2 HEADOFFICE

## 3 HEADOFFICE

## 4 HEADOFFICE

## 5 HEADOFFICE

## 6 HEADOFFICE

Check for any missing values in the data. The result is FALSE, meaning there are no missing values in the data.

anyNA(emp\_attrition)

## [1] FALSE

Convert the resignation to factor for the dataset.

emp\_attrition$termreason\_desc <- as.factor(gsub("Resignaton", "Resignation", emp\_attrition$termreason\_desc))

Check for rows and columns in the data.There are 49653 employees and 18 features in the dataset.

dim(emp\_attrition)

## [1] 49653 18

Give a summary of the data emp\_attrition

summary(emp\_attrition)

## EmployeeID recorddate\_key birthdate\_key orighiredate\_key

## Min. :1318 Length:49653 Length:49653 Length:49653

## 1st Qu.:3360 Class :character Class :character Class :character

## Median :5031 Mode :character Mode :character Mode :character

## Mean :4859

## 3rd Qu.:6335

## Max. :8336

## terminationdate\_key age length\_of\_service city\_name

## Length:49653 Min. :19.00 Min. : 0.00 Length:49653

## Class :character 1st Qu.:31.00 1st Qu.: 5.00 Class :character

## Mode :character Median :42.00 Median :10.00 Mode :character

## Mean :42.08 Mean :10.43

## 3rd Qu.:53.00 3rd Qu.:15.00

## Max. :65.00 Max. :26.00

## department\_name job\_title store\_name gender\_short

## Length:49653 Length:49653 Min. : 1.0 Length:49653

## Class :character Class :character 1st Qu.:16.0 Class :character

## Mode :character Mode :character Median :28.0 Mode :character

## Mean :27.3

## 3rd Qu.:42.0

## Max. :46.0

## gender\_full termreason\_desc termtype\_desc STATUS\_YEAR

## Length:49653 Layoff : 215 Length:49653 Min. :2006

## Class :character Not Applicable:48168 Class :character 1st Qu.:2008

## Mode :character Resignation : 385 Mode :character Median :2011

## Retirement : 885 Mean :2011

## 3rd Qu.:2013

## Max. :2015

## STATUS BUSINESS\_UNIT

## Length:49653 Length:49653

## Class :character Class :character

## Mode :character Mode :character

##

##

##

## Methodology

Having the clear picture of the dataset, then we start working on the dataset to identify various analysis related to termination.First, we find the number of employees who has left the company each year.

attrition\_status <- with(emp\_attrition, table(STATUS\_YEAR, STATUS))

attrition\_status <- spread(data.frame(attrition\_status), STATUS, Freq)

attrition\_status$previous\_active <- shift(attrition\_status$ACTIVE, 1L, type = "lag")

attrition\_status$percent\_terminated <- 100\*attrition\_status$TERMINATED / attrition\_status$previous\_active

attrition\_status

## STATUS\_YEAR ACTIVE TERMINATED previous\_active percent\_terminated

## 1 2006 4445 134 NA NA

## 2 2007 4521 162 4445 3.644544

## 3 2008 4603 164 4521 3.627516

## 4 2009 4710 142 4603 3.084945

## 5 2010 4840 123 4710 2.611465

## 6 2011 4972 110 4840 2.272727

## 7 2012 5101 130 4972 2.614642

## 8 2013 5215 105 5101 2.058420

## 9 2014 4962 253 5215 4.851390

## 10 2015 4799 162 4962 3.264813

Next, we depict and visualize in a barchart the employee attrition in each year and the reasons for attrition.

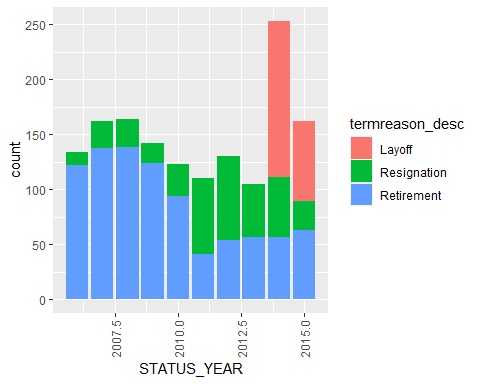
reasons\_attrition <- as.data.frame(emp\_attrition %>% filter(STATUS=="TERMINATED"))

# plot reason for termination

library(ggplot2)

ggplot() + geom\_bar(aes(y = ..count..,x = STATUS\_YEAR, fill = termreason\_desc), data=reasons\_attrition, position = position\_stack()) +

theme(axis.text.x = element\_text(angle = 90, hjust = 1, vjust = 0.5))



Next, we perform modelling to identify ways through which employee attrition can be predicted.

# subset by selecting variables important in predicting employee attrition

reasons\_attrition\_vars <- c("age","length\_of\_service","city\_name", "department\_name","job\_title","store\_name","gender\_full","BUSINESS\_UNIT","STATUS")

# split data to training and testing sets

emp\_attrition\_reasons\_train <- subset(emp\_attrition, STATUS\_YEAR < 2015)

emp\_attrition\_reasons\_test <- subset(emp\_attrition, STATUS\_YEAR == 2015)

set.seed(99) # set a pre-defined value for the random seed so that results are repeatable

# Create the Decision tree model

tree <- rpart(STATUS ~.,

data = emp\_attrition\_reasons\_train[reasons\_attrition\_vars],

method = 'class',

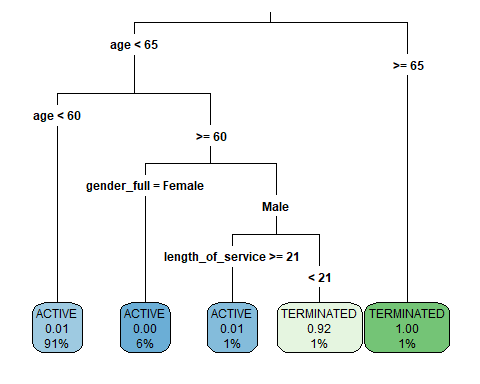
parms = list(split='information'),

control = rpart.control(usesurrogate = 0,

maxsurrogate = 0))

# Plot the decision tree

rpart.plot(tree, roundint = FALSE, type = 3)



Perform a prediction in future voluntary terminations by creating the resigned variable.

# create a voluntary\_terminations column

emp\_attrition$resigned <- ifelse(emp\_attrition$termreason\_desc == "Resignation", "Yes", "No")

# convert resigned column to factor (from character)

emp\_attrition$resigned <- as.factor(emp\_attrition$resigned)

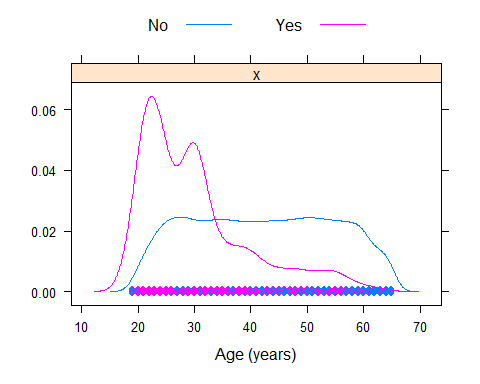
summary(emp\_attrition$resigned)

## No Yes

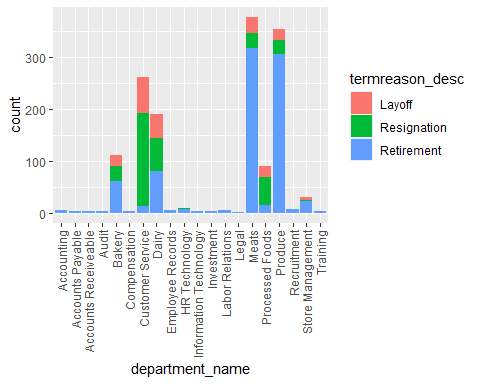
## 49268 385

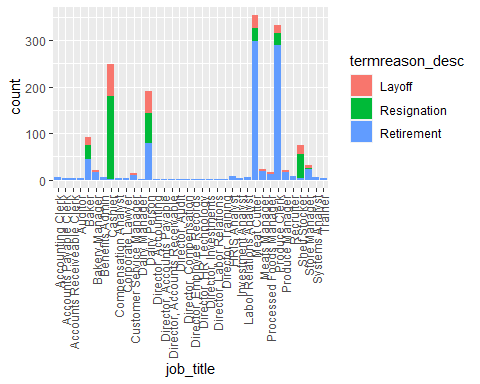
## Results

The plots illustrate that the employee attrition is more for younger employees are at aged between 20 years and 30 years.



It is also clear that employee attrition is high in particular job titles like shelf stocker and cashier. Additionally, departments like Processes Foods, Customer Service, and Diary have higher employee attrition.





## Implications

This project only applied the decision tree and some Exploratory Data Analysis. For future research, I would recommend implementing the Random Forest model to help identify more insights and at-risk employees for attrition.

## Conclusion

The analysis in this project has illustrated the various factors that majorly contribute to employee attrition. Therefore, with this, companies will be able to address these factors. In addition, with the prediction, the managers will be able to improve their employees' engagement and experiences to ensure they retain them.

## References

Mitkees, I. M., Badr, S. M., & ElSeddawy, A. I. B. (2017, December). Customer churn prediction model using data mining techniques. In 2017 13th International Computer Engineering Conference (ICENCO) (pp. 262-268). IEEE.

Omar Ali, N. Z. M. (2017). Factors affecting employee turnover in organization.

Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. Information & Management, 45(3), 164-174.

Khare, R., Kaloya, D., Choudhary, C. K., & Gupta, G. (2011, January). Employee attrition risk assessment using logistic regression analysis. In Int. Conf. Adv. Data Anal. Bus. Anal. Intell (pp. 1-33).