HR Analytics

**CAPSTONE PROJECT**

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE EXECUTIVE MANAGEMENT DEVELOPMENT PROGRAM IN IN ADVANCED MARKETING STRATEGY AND ANALYTICS  
BY  
**GROUP<x>\_AMA01**INDIAN INSTITUTE OF MANAGEMENT KOZHIKODE

Date:- 20th July 2022

**ACKNOWLEDGMENTS**

**EXECUTIVE SUMMARY**

**TABLE OF CONTENT**

INTRODUCTION

BACKGROUND/MOTIVATION

OTHER RELEVANT SECTIONS ON RESEARCH PROBLEM IDENTIFICATION, INDUSTRY AND COMPANY BACKGROUND

DATA DESCRIPTION AND SOURCE

METHODOLOGY

RESULTS/ANALYSIS I

CLASSIFICATION MODELS

INTERPRETATIONS/DISCUSSION

CONCLUSION

REFERENCES

**INTRODUCTION**

A Machine learning project with the main aim of doing some Descriptive and Exploratory Data Analysis and then applying predictive modelling for predicting whether the employee will leave or not from the organisation.

**BACKGROUND/MOTIVATION**

As the number of employees leaving the company has increased , HR people need to identify the reasons behind employees leaving the company.

I this project using the provided data we do an analysis to identify the possible reasons behind employees leaving the company.

This analysis will help the HR to identify the reasons for employees attrition, thereby reducing the cost to hire new employees for the organisation.

It will also help to identify to the pain points which employees are going through where in they can have new policies in favour of employees. This good boost a happy environment in the organisation for the employees.

**OTHER RELEVANT SECTIONS ON RESEARCH PROBLEM IDENTIFICATION, INDUSTRY AND COMPANY BACKGROUND**

No such information is available , but analysis could be used by any organisation which employs employees to work.

**DATA DESCRIPTION AND SOURCE**

Context

Hr Data Analytics  
This dataset contains information about employees who worked in a company.

Content

This dataset contains columns: Satisfactory Level, Number of Project, Average Monthly Hours, Time Spend Company, Promotion Last 5 Years, Department, Salary

Source:-

Kaggle (<https://www.kaggle.com/datasets/mfaisalqureshi/hr-analytics-and-job-prediction>)

**METHODOLOGY**

We have made used of R in order to do both exploratory analysis as well as Prediction.

Have made used of bar plots, heatmaps and boxplots to do the Exploratory Analysis.

Logistic Regression is used for the Prediction.

**RESULTS/ANALYSIS I**

**## Exploratory Analysis**

> dim(data\_set)

[1] 14999 10

**Analysis:-**

* The dataset contains 10 columns and 14999 observations

> str(data\_set)

'data.frame': 14999 obs. of 10 variables:

$ satisfaction\_level : num 0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...

$ last\_evaluation : num 0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...

$ number\_project : int 2 5 7 5 2 2 6 5 5 2 ...

$ average\_montly\_hours : int 157 262 272 223 159 153 247 259 224 142 ...

$ time\_spend\_company : int 3 6 4 5 3 3 4 5 5 3 ...

$ Work\_accident : int 0 0 0 0 0 0 0 0 0 0 ...

$ left : int 1 1 1 1 1 1 1 1 1 1 ...

$ promotion\_last\_5years: int 0 0 0 0 0 0 0 0 0 0 ...

$ Department : chr "sales" "sales" "sales" "sales" ...

$ salary : chr "low" "medium" "medium" "low" ...

**Analysis:-**

* Most of the features\columns are in numeric or integer type except for Department and salary which are in char format

**# Let us find the attrition percentage of the employees**

> attrition<-as.factor(data\_set$left)

> summary(attrition)

0 1

11428 3571

> perc\_attrition\_rate<-sum(data\_set$left/length(data\_set$left))\*100

> print(perc\_attrition\_rate)

[1] 23.80825

**Analysis:-**

* Looks like approx.. 76% of employees stayed and 24% of employees left.

**# Overview of summary (Turnover V.S. Non-turnover)**

> cor\_vars<-data\_set[,c("satisfaction\_level","last\_evaluation","number\_project","average\_montly\_hours","time\_spend\_company","Work\_accident","left","promotion\_last\_5years")]

>

> aggregate(cor\_vars[,c("satisfaction\_level","last\_evaluation","number\_project","average\_montly\_hours","time\_spend\_company","Work\_accident","promotion\_last\_5years")], by=list(Category=cor\_vars$left), FUN=mean)

Category satisfaction\_level last\_evaluation number\_project average\_montly\_hours time\_spend\_company Work\_accident

1 0 0.6668096 0.7154734 3.786664 199.0602 3.380032 0.17500875

2 1 0.4400980 0.7181126 3.855503 207.4192 3.876505 0.04732568

promotion\_last\_5years

1 0.026251313

2 0.005320638

>

> trans<-cor(cor\_vars)

> melted\_cormat <- melt(trans)

>

> ggplot(data = melted\_cormat, aes(x=Var1, y=Var2, fill=value)) +

+ geom\_tile() +theme(axis.text.x = element\_text(angle = 90, hjust = 1))

A picture containing square

Description automatically generated

**Analysis:-**

* From the heatmap, there is a **positive(+)** correlation between projectCount, averageMonthlyHours, and evaluation. Which could mean that the employees who spent more hours and did more projects were evaluated highly.
* For the **negative(-)** relationships, turnover and satisfaction are highly correlated. I'm assuming that people tend to leave a company more when they are less satisfied

**## Salary vs Turnover**

> vis\_1<-table(data\_set$salary,data\_set$left)

> d\_vis\_1<-as.data.frame(vis\_1)

> p<-ggplot(d\_vis\_1, aes(x=Var1,y=Freq,fill=Var2)) +

+ geom\_bar(position="dodge",stat='identity')+

+ labs(x = "Salary",y ="Frequency")

> p

Chart, bar chart

Description automatically generated

**Analysis:-**

* Most of employees who left were from low and medium salary range.
* Barely any employees left with **high** salary

**## Dept vs Turnover**

> vis\_1<-table(data\_set$Department,data\_set$left)

> #print(vis\_1)

> d\_vis\_1<-as.data.frame(vis\_1)

> library(ggplot2)

> p<-ggplot(d\_vis\_1, aes(x=Var1,y=Freq,fill=Var2)) +

+ geom\_bar(position="dodge",stat='identity')+labs(x = "Department",y ="Frequency")

> p

Chart, bar chart

Description automatically generated

**Analysis:-**

* Proportion of employees leaving across department seems to be same

**## Turnover vs projectcount**

> vis\_1<-table(data\_set$number\_project,data\_set$left)

> #print(vis\_1)

> d\_vis\_1<-as.data.frame(vis\_1)

> p<-ggplot(d\_vis\_1, aes(x=Var1,y=Freq,fill=Var2)) +

+ geom\_bar(position="dodge",stat='identity')+labs(x = "ProjectCount",y ="Frequency")

> p

Chart, bar chart

Description automatically generated

**Analysis:-**

* More than half of the employees left who were assigned 2 projects
* Majority of the employees who didn’t leave had 3,4 and 5 projects
* Proportion of employees leaving with 6 projects was higher than staying
* All of the employees with 7 projects left the company

**## Project\_Count vs satisfaction\_level**

>vis\_1<-table(data\_set$number\_project,data\_set$satisfaction\_level<0.1)

> #print(vis\_1)

> d\_vis\_1<-as.data.frame(vis\_1)

> p<-ggplot(d\_vis\_1, aes(x=Var1,y=Freq,fill=Var2)) +

+ geom\_bar(position="dodge",stat='identity')+labs(x = "ProjectCount",y ="FrequencyOFSatisfactionLevel<0.1",fill="SatisfactionLevel<0.1")

Chart, bar chart

Description automatically generated

**Analysis:-**

* It can be seen that proportion of employees with lesser satisfaction level is more in number of projects 6 (13%) and 7 (23%)

**#KDEPlot: Kernel Density Estimate Plot**

**Density plot for Left\_data(in red) and stay\_data (in green) vs avg\_monthly\_hrs**

> left\_data<-subset(data\_set,left==1)

> stay\_data<-subset(data\_set,left==0)

> ggplot() + geom\_density(aes(x=average\_montly\_hours), colour="red", data=left\_data) +

+ geom\_density(aes(x=average\_montly\_hours), colour="blue", data=stay\_data)

Chart, line chart

Description automatically generated

**Analysis:-**

* Employees who had less hours of work **(~150hours or less)** left the company more
* Employees who had too many hours of work **(~250 or more)** left the company
* Employees who left generally were **underworked** or **overworked**.

**#Turnover V.S. Evaluation**

ggplot() + geom\_density(aes(x=last\_evaluation), colour="red", data=left\_data) +

geom\_density(aes(x=last\_evaluation), colour="blue", data=stay\_data)

Chart, line chart

Description automatically generated

**Analysis:-**

* Employees with **low** performance tend to leave the company more
* Employees with **high** performance tend to leave the company more
* The **sweet spot** for employees that stayed is within **0.6-0.8** evaluation

**#ProjectCount VS AverageMonthlyHours [BOXPLOT]**

<-ggplot(data\_set, aes(x = factor(number\_project), y = average\_montly\_hours, fill = factor(left))) +

geom\_boxplot() + scale\_fill\_manual(values = c("yellow", "orange"))

<-print(p)

Chart, box and whisker chart

Description automatically generated

**Analysis:-**

* Looks like the average employees who stayed worked about 200hours/month.
* Those that had a turnover worked about 250hours/month or 150hours/month

**#ProjectCount VS Evaluation**

<-ggplot(data\_set, aes(x = factor(number\_project), y = last\_evaluation, fill = factor(left))) +

geom\_boxplot() + scale\_fill\_manual(values = c("yellow", "orange"))

print(p)

Chart, box and whisker chart

Description automatically generated

**Analysis:-**

* Looks like employees who did not leave the company had an average evaluation of around 70% even with different projectCounts
* There is a huge skew in employees who had a turnover though. It drastically changes after 3 projectCounts.
* Employees that had two projects and a horrible evaluation left. Employees with more than 3 projects and super high evaluations left

## **Satisfaction VS Evaluation**

library(ggplot2)

ggplot(data\_set, aes(satisfaction\_level, last\_evaluation, color = left)) +

geom\_point(shape = 16, size = 5, show.legend = FALSE) +

theme\_minimal() +

scale\_color\_gradient(low = "#0091ff", high = "#f0650e")

Chart

Description automatically generated

**Analysis:-**

* #Cluster 1 (Hard-working and Sad Employee): Satisfaction was below 0.2 and evaluations were greater than 0.75. Which could be a good indication that employees who left the company were good workers but felt horrible at their job.
* # Cluster 2 (Bad and Sad Employee): Satisfaction between about 0.35~0.45 and evaluations below ~0.58. This could be seen as employees who were badly evaluated and felt bad at work.
* # Cluster 3 (Hard-working and Happy Employee): Satisfaction between 0.7~1.0 and evaluations were greater than 0.8. Which could mean that employees in this cluster were "ideal". They loved their work and were evaluated highly for their performance.

**CLASSIFICATION:-**

We also wanted to classify the employee will leave or not based on the data collected, so as to help the HR to identify them before , where in they can come with various measures which can help them to retain the employees who are on the verge to leave.

The steps used in classification are as follows:-

**Feature Selection:-**

It is more important to identify the features which actually impact the dependent variable.

We used logistic Regression for the classification. As logistic Regression needs all the variables in numeric format. We created Dummy variables for sting features i.e. Department and salary.

Further the dataset was divided into train(70% of the data) and test data(30% of the data).

Inorder to get rid of the multicollinearity we used vif and removed any variables with vif <5 (No variables were actually found with vif<5) .

**Modeling:-**

Once the features are selected we perform modelling on the data. We are using over here the basic classification machine learning model Logistic Regression.

fit=glm(left~.,family=binomial,data=train)

Once the model is fitted we create a predicted variable which predicts the probability of employee leaving the organisation on train

train$score=predict(fit2,newdata=train,type="response")

Next we predict the scores on test data

test$score=predict(fit,newdata=test,type="response")

After finding the threshold\cut-off values for the score we compare the values with that of the actual ones.

table(test$left,as.numeric(test$score>cutoff\_KS))#

0 1

0 2726 710

1 320 744

**Accuracy**:-

(TP+TN)/(TP+TN+FP+FN) = (2726+744)/(2726+744+710+320)

= 0.7711111

Using the basic classification model we are able to classify whether employee will churn or not with an accuracy of 77.1 %.

Further Enhancements can be done to the classification by trying various other Machine learning techniques, like bagging , boosting , Random Forest etc.

**INTERPRETATIONS/DISCUSSION**

It is evidence from the data that the employees with low satisfaction were mostly to turn over. Further analysis\surveys can be taken up with the HR in order to find the base reason for lower satisfaction rate.

Also employees can be accessed with the number of projects they are assigned to; inorder to understand the workload across the projects. Employees with project number 2, 6, 7 can be accessed specifically to understand their workload.

Employees with low to medium salaries are at a greater possibility to leaving the organisation. Such employees should be evaluated further to find if they are underpaid as oppose to the work they are doing.

Employees generally seem to be leaving after spending 4-5 years in the organisation. Various measures should be kept in place like employee manager relationship over the period of time, like whether the employees are happy in their current work environment, if not what are the possible reasons for same.

Finding out reasons for over working hours of the employees. Taking actions on them like dividing the projects across the employees who are having low working hours. Getting employees cross trained so that they can work on multiple domains. Understanding the employee concerns with respect to working hours etc.

**CONCLUSION:-**

**Summary:** With all of this information, this could be the possible reasons because of which employees left:-

1. Employees generally left when they are **underworked** (less than 150hr/month or 6hr/day)
2. Employees generally left when they are **overworked** (more than 250hr/month or 10hr/day)
3. Employees with either **really high or low evaluations** should be taken into consideration for high turnover rate
4. Employees with **low to medium salaries** are the bulk of employee turnover
5. Employees that had **2,6, or 7 project count** was at risk of leaving the company
6. Employee **satisfaction** is the highest indicator for employee turnover.
7. Employee that had **4 and 5 yearsAtCompany** should be taken into consideration for high turnover rate

**REFERENCES:-**

[**https://machinelearningmastery.com/machine-learning-in-r-step-by-step/**](https://machinelearningmastery.com/machine-learning-in-r-step-by-step/)

[**https://www.geeksforgeeks.org/logistic-regression-in-r-programming/#:~:text=Logistic%20regression%20in%20R%20Programming%20is%20a%20classification%20algorithm%20used,function%20in%20a%20binomial%20distribution**](https://www.geeksforgeeks.org/logistic-regression-in-r-programming/#:~:text=Logistic%20regression%20in%20R%20Programming%20is%20a%20classification%20algorithm%20used,function%20in%20a%20binomial%20distribution)**.**

[**https://www.kaggle.com/datasets/mfaisalqureshi/hr-analytics-and-job-prediction**](https://www.kaggle.com/datasets/mfaisalqureshi/hr-analytics-and-job-prediction)