BUAN 6356
BUSINESS ANALYTICS WITH R

PROJECT:
EMPLOYEE ATTRITION ANALYSIS

Hothal Vimalbhai Patel
(hxp172030)

Table of Contents

Executive Summary

It is very crucial to determine who is leaving the organization, when they are leaving, and why they are leaving. Enlisting, recruiting, onboarding, and preparing new workers costs organizations billions every year. Organizations additionally endure efficiency misfortunes and lost benefits when there is a lot of nonstop stir in the workforce. Top ability, specifically, can be exceptionally troublesome and costly to replace. Money aside, organizations are in an ideal situation when they can hold great employees and the skills they have.

This is where the power of analytics comes into the picture. Employee Attrition Analysis is explicitly centred around recognizing why workers deliberately leave, what may have kept them from leaving, and how we can utilize information to foresee steady loss hazard. Above all, this kind of analytics can be utilized to assist companies to recognize and structure the issues that will be best in reducing attrition.

How does Attrition influence organizations? also, how does Analytics help in breaking down steady loss? I will conduct detailed analysis in this project to understand what factors affect attrition and what steps we can take to prevent it.

Problem Statement

ABC is a company that has approximately 4400 employees and they are facing about 16% attrition and thus want to identify what factors are leading to it. Any kind of attrition causes harm to the company as the expense to hire new employees is more than retaining them.

The main goal is to identify the factors impacting attrition to reduce the costs and improve work efficiency.

- H0: The factors won't have any significant impact on the attrition rate of employees.
- H1: The factors will have a significant impact on the attrition rate of employees.

Data Description

The Employee Attrition database is used for this project. It is obtained from Kaggle. It has information about 4410 employees with 30 variable factors.

The data is available in five datasets as follows:

Employee Survey Data	Manager Survey Data
EmployeeID	EmployeeID
Employee Satisfaction	JobInvolvement
JobSatisfaction	PerformanceRating
WorkLifeBalance	

In Time	Out Time
Daywise Intime of employee	Daywise outtime of employee

General Data	
Age	Attrition
BusinessTravel	Department
DistanceFromHome	Education
EducationField	EmployeeCount
EmployeeID	Gender
JobLevel	JobRole
MaritalStatus	MonthlyIncome
NumCompaniesWorked	Over18
PercentSalaryHike	StandardHours
StockOptionLevel	TotalWorkingYears
TrainingTimeLastYear	YearsatCompany
YearsSinceLastPromotion	YearsWithCurrManager

Data Processing

a) Removing unnecessary columns:

- Names of columns having only 1 unique value.
- Remove *AvgWorkHours* as we have derived *Overtime* variable from it
- Removing *EmployeeID* from other datasets as files are merged

b) Converting the below variables into its factors as given in data dictionary:

Level	Education	EnvironmentSatisfaction	JobInvolvement	JobSatisfaction	PerformanceRating	WorkLifeBalance
1	Below College	Low	Low	Low	Low	Bad
2	College	Medium	Medium	Medium	Good	Good
3	Bachelor	High	High	High	Excellent	Better
4	Master	Very High	Very High	Very High	Outstanding	Best
5	Doctor					

c) Replaced NA values of *NumCompaniesWorked* and *TotalWorkingYears* with the median.

d) Converted time data into *AverageWorkingHours* for each employee and created a new data frame for the same and merged with employee ID into *workingtime*.

e) Derived three new variables:

- *TenurePerJob*: It defines the time period for which an employee remains in one organization.
- *YearsWithoutChange 1 & 2*: We create two variables to see how many years (at the company and total working years) it has been for an employee without any sort of change using Promotion and Job Change.

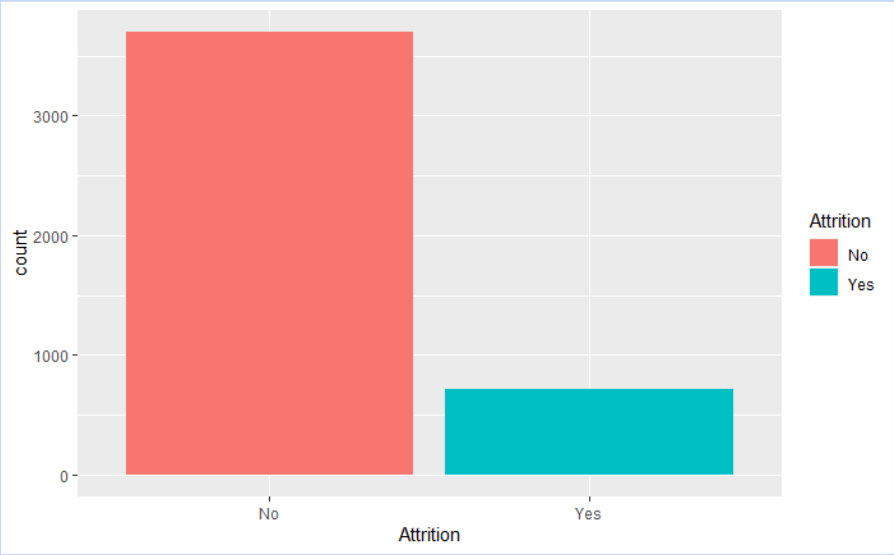
f) Merged all five datasets, Employee Survey Data, Manager Survey Data, General Data, Intime and Outtime data into one master dataset employee

g) Dropping columns *EmployeeCount*, *Over18*, and *StandardHours* as they have only one unique value and therefore it cannot be important

Exploratory Data Analysis

a) Barcharts for Categorical Variables

- Graph below shows approximately 16% attrition rate
- Majority employees are not leaving the company (84%)
- Imbalanced Dataset



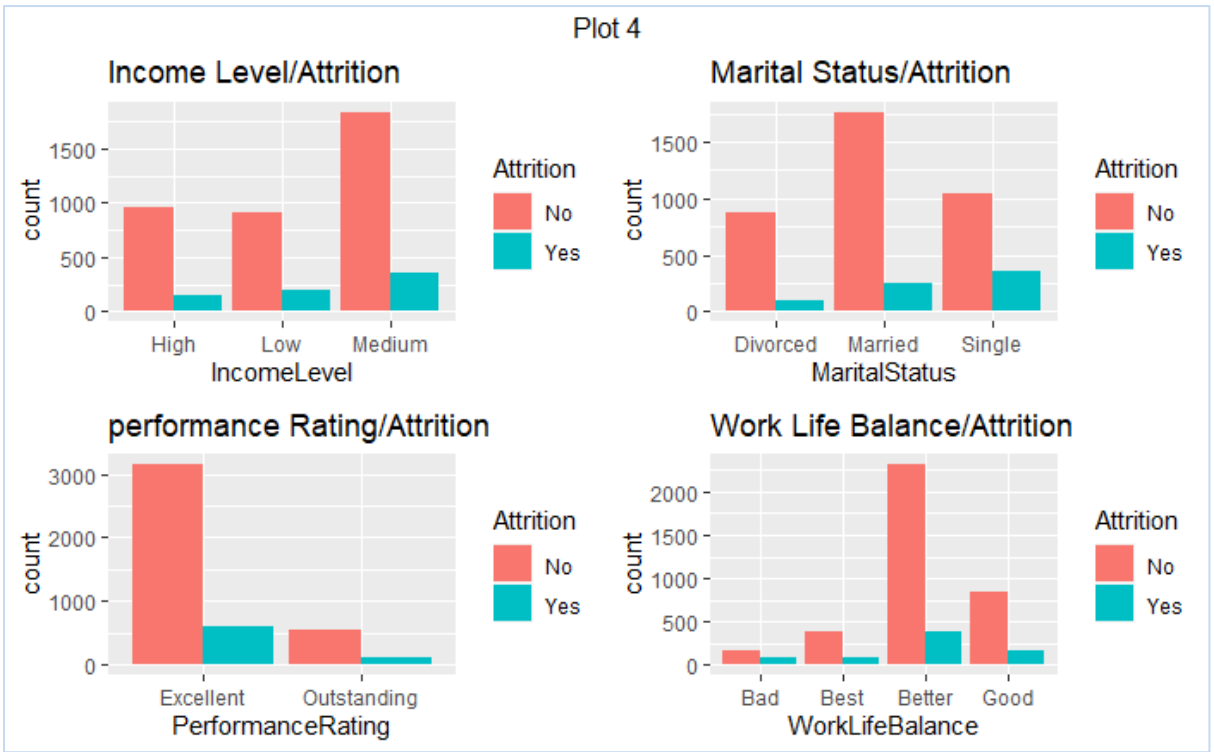

```
prop.table(table(employee$Attrition))


##
##     No    Yes
## 0.839 0.161
```

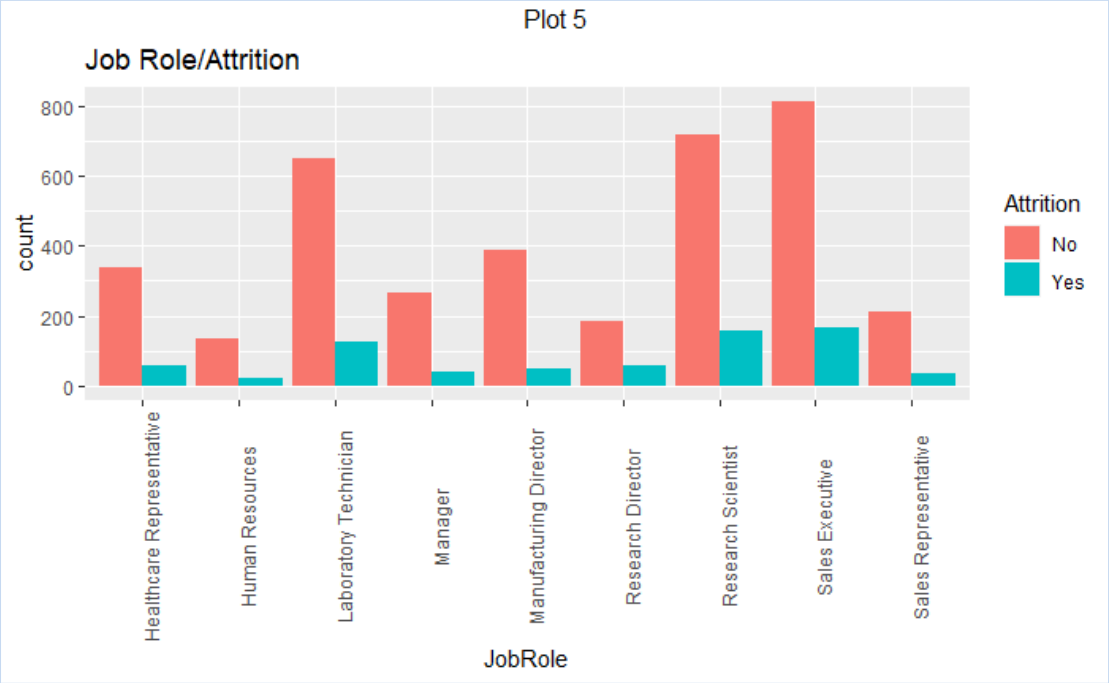
The graphs below show relationship between Attrition with respect to each important variable in the dataset.

Conclusions:

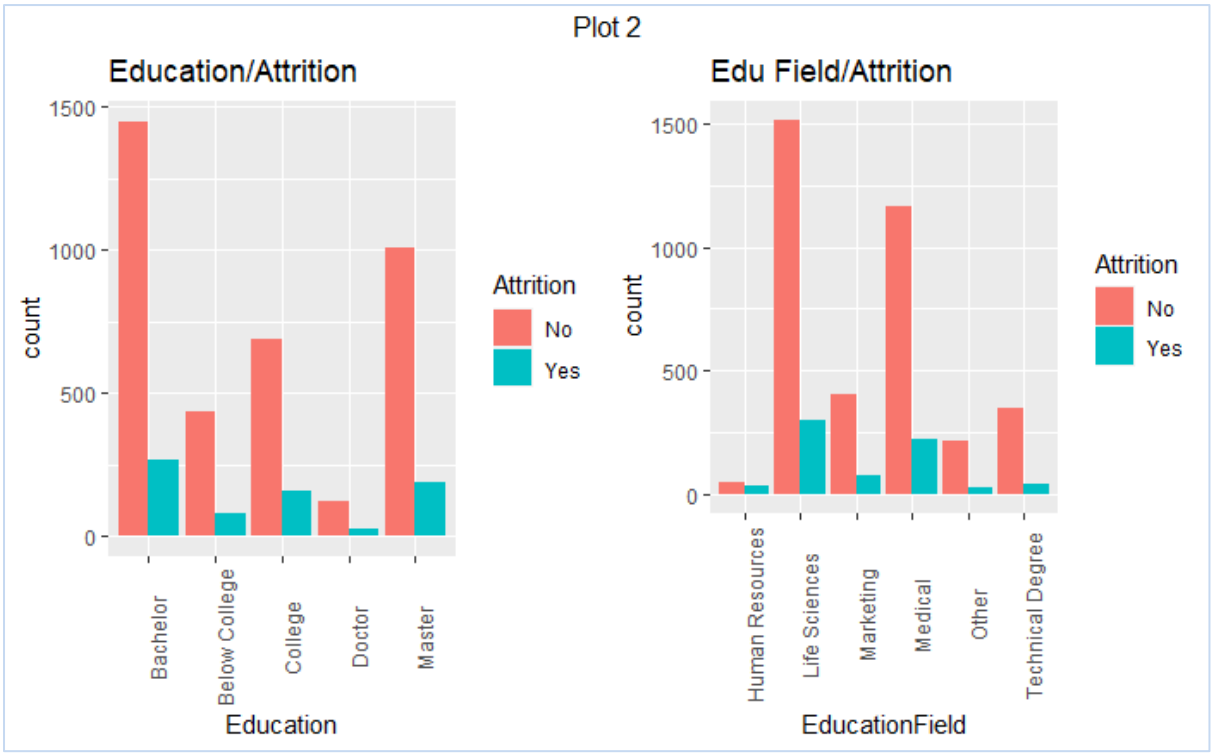
- Gender and Environment satisfaction does not have significance on Attrition
- Employees with High and Medium Job Involvement tend to leave the company more

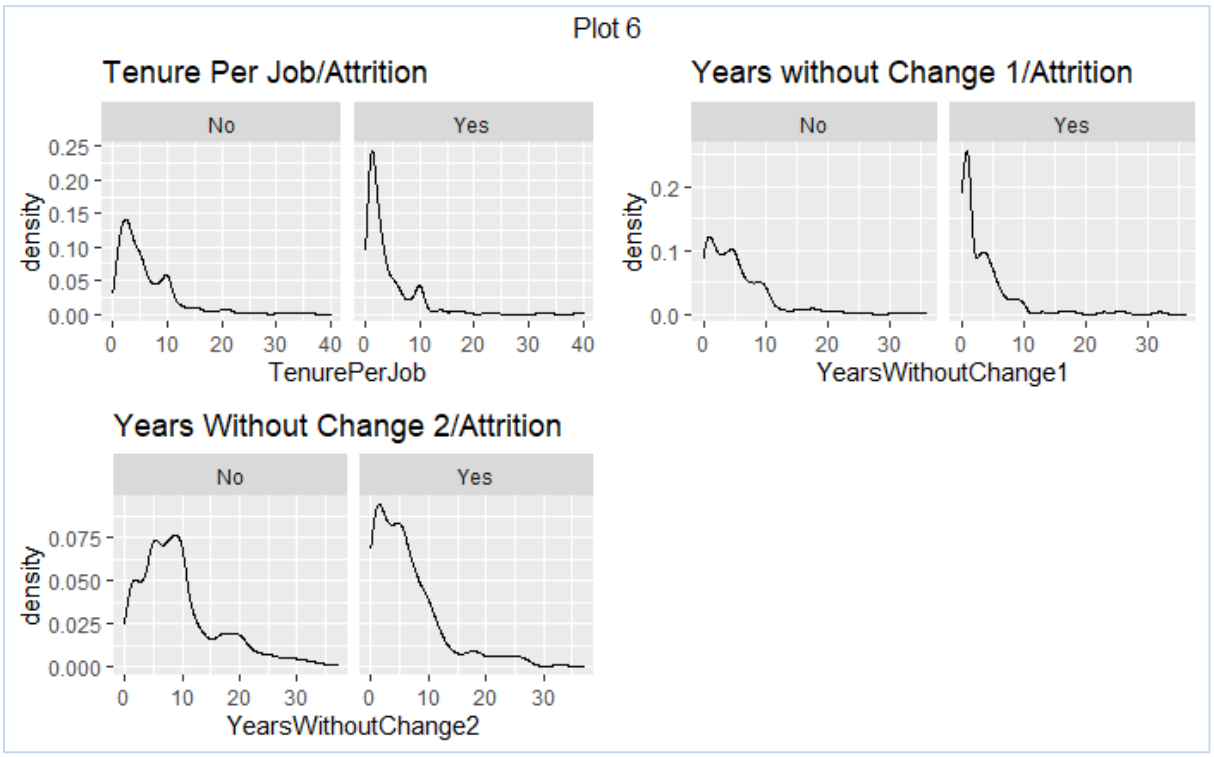


- *IncomeLevel:* We see higher levels of attrition among the lower segment of monthly income.
- *MaritalStatus:* The attrition rate is more in the case of Single and less among divorcees.
- *JobRole:* More attrition in the case of Research Scientists and Sales Executives and less in the case of HR



- Bachelor's and Master's Level of education is more prominent
- Employees from Life Sciences and Medical field have more attrition rate
- R&D and Sales Department has more amount of attrition as well as employees
- Employees who travel tend to quit job more than non-travellers

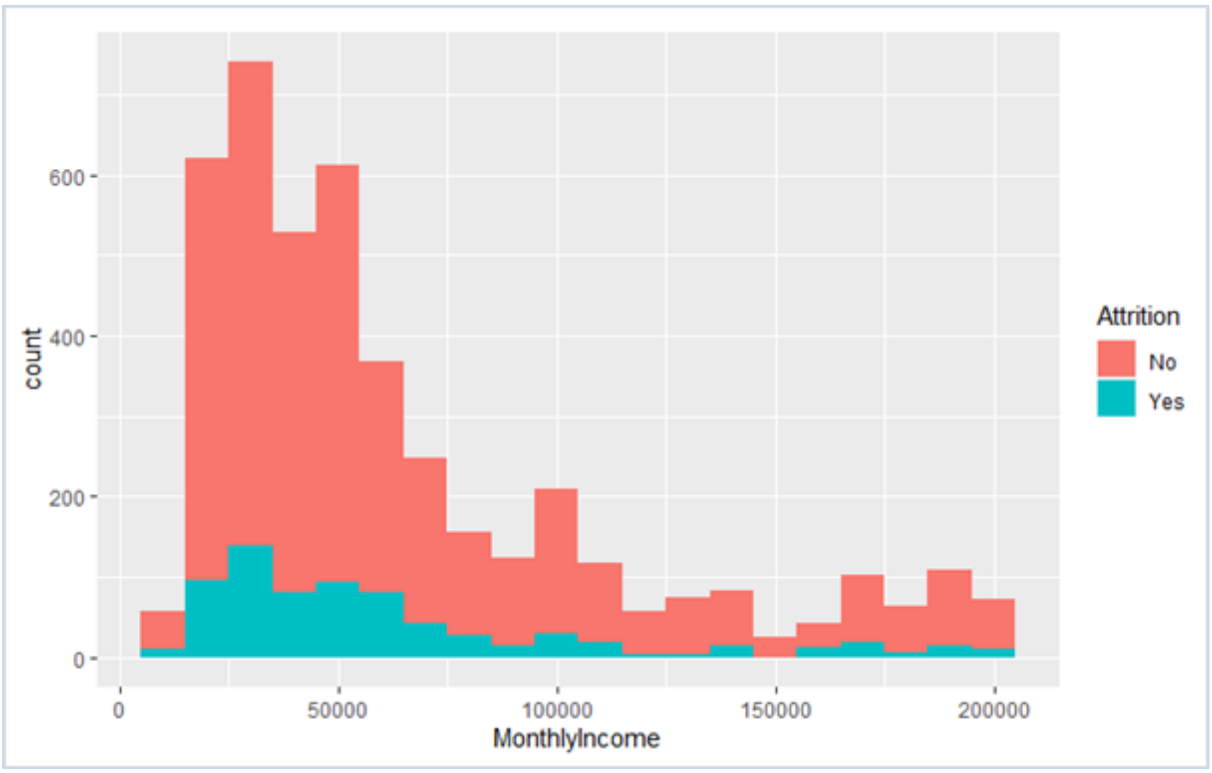




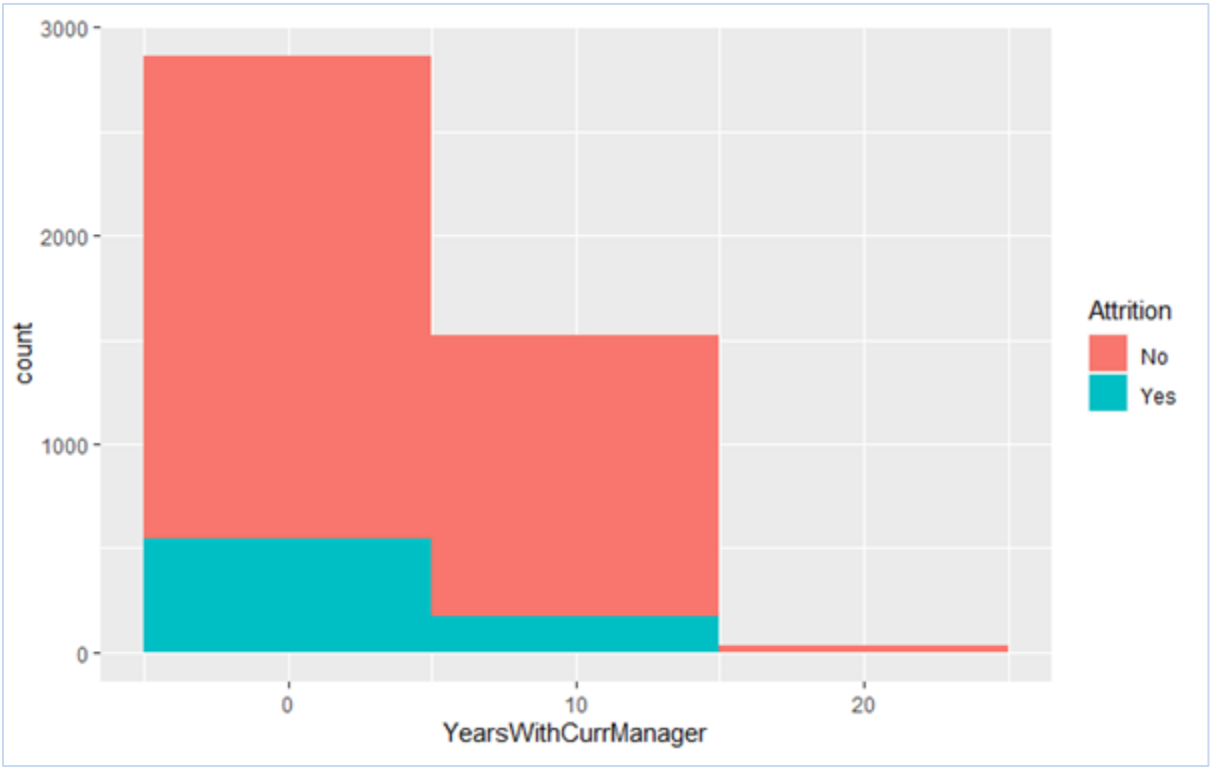
- *TenurePerJob:* Usually, people who have worked with many companies for small periods tend to leave early.
- For the newly derived metrics, *YearsWithoutChange1*, and *YearsWithoutChange2*, we can see with low tenure and low years without change tend to leave the organization.

b) Histograms for numeric variables:

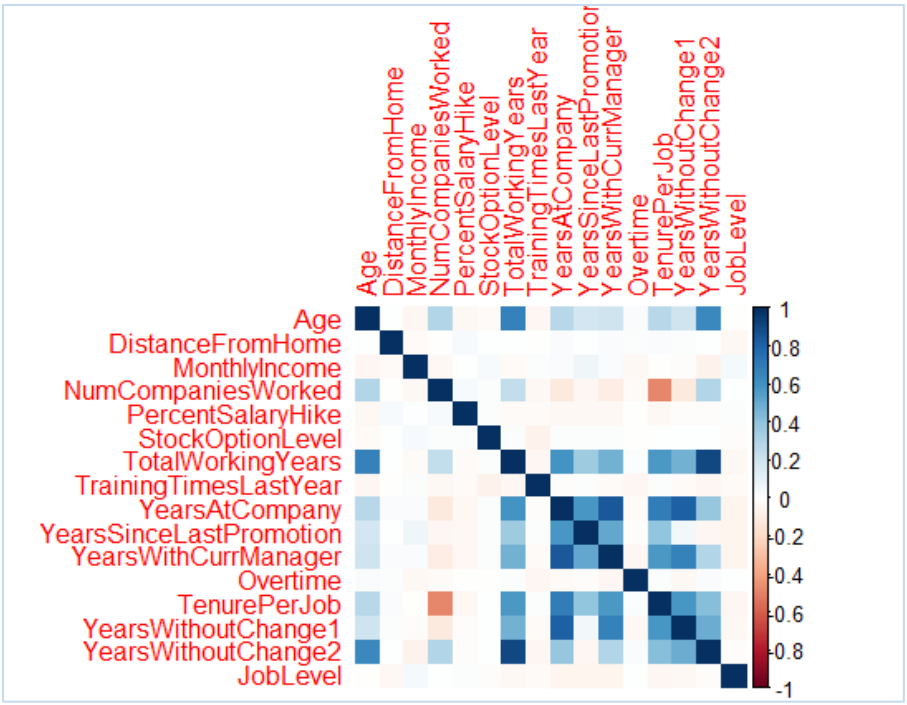
Higher levels of attrition among the lower segment of monthly income:



As expected, a new Manager is a big cause for quitting:



c) Correlation between Numerical Variables



A strong correlation exists between:

- Tenure per job and number of companies worked
- Years without change 2 and Total Working years
- Years without change 1 and years at the company

Outliers Treatment and Standardization

Outliers Treatment for numeric variables:

- Used Boxplots to identify the outliers
- Columns *Age*, *DistancefromHome*, *PercentSalaryHike* have no outliers
- Columns *MonthlyIncome*, *NumCompaniesWorked, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear,YearsAtCompany, YearsSinceLastPromotion, YearsWithCurrManager, TenurePerJob, YearsWithoutChange1, YearsWithoutChange2* have outliers and are treated.

Normalization:

a) Converted target variable Attrition from No/Yes character to factor with levels 0/1

b) Scaled the following variables:

- *Age*
- *DistancefromHome*

- *PercentSalaryHike*
- *MonthlyIncome*
- *NumCompaniesWorked*
- *TotalWorkingYears*
- *TrainingTimesLastYear*
- *YearsAtCompany*
- *YearsSinceLastPromotion*
- *YearsWithCurrManager*
- *TenurePerJob*
- *YearsWithoutChange1*
- *YearsWithoutChange2*

c) Created dummy variables for categorical variables

Predictive Modelling

I have used **Logistic Regression** to build our predictive model. The initial model contains all the variables and uses here **stepwise variable selection method** due to a large number of variables. We use a VIF check to remove multicollinearity. We find our 10 significant factors in the 29th model.

Steps taken to build the model:

- Split dataset into train (70%) and test (30%) data
- Model building starts with all variables
- Use StepAIC to remove insignificant variables with high multicollinearity
- Eliminate variables with high VIF value
- The final model contains variables that are highly significant and low multicollinearity

Below are the 10 final variables:

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.91306    0.13699 -21.264  < 2e-16 ***
NumCompaniesWorked               0.32364    0.05565   5.816 6.04e-09 ***
YearsWithCurrManager            -0.33127    0.06792  -4.878 1.07e-06 ***
YearsWithoutChange2             -0.70848    0.07415  -9.555  < 2e-16 ***
BusinessTravel.xTravel_Frequently  0.77436    0.12874   6.015 1.80e-09 ***
MaritalStatus.xSingle            1.02349    0.11111   9.212  < 2e-16 ***
EnvironmentSatisfaction.xLow     1.04796    0.12689   8.259  < 2e-16 ***
JobSatisfaction.xLow             0.55966    0.13538   4.134 3.56e-05 ***
JobSatisfaction.xVery.High      -0.63536    0.13714  -4.633 3.60e-06 ***
WorkLifeBalance.xBetter         -0.39448    0.11109  -3.551 0.000384 ***
Overtime                         1.31834    0.11311  11.656  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

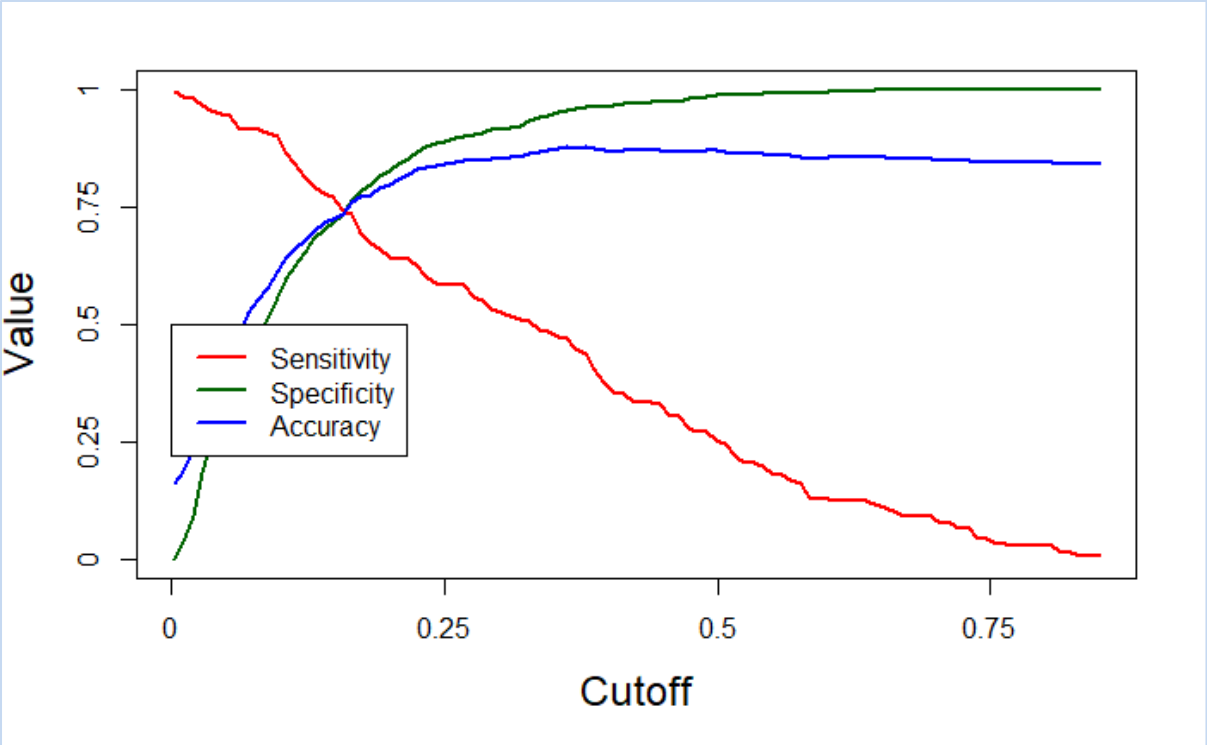
The factors that will lead to Employee Attrition: Number of Companies worked, Business Travel (Frequently), Marital Status (Single), Low Environment Satisfaction, Low Job Satisfaction, and Overtime

The factors that will lead to Employee Retention: Years with current manager, Years without change, High Job Satisfaction and Better Work life Balance

Thus, through our analysis we see that the factors have a significant impact on the attrition rate hence, **we reject the null hypothesis.**

Model Evaluation

The optimal probability cut off from the graph below is approximately 0.16



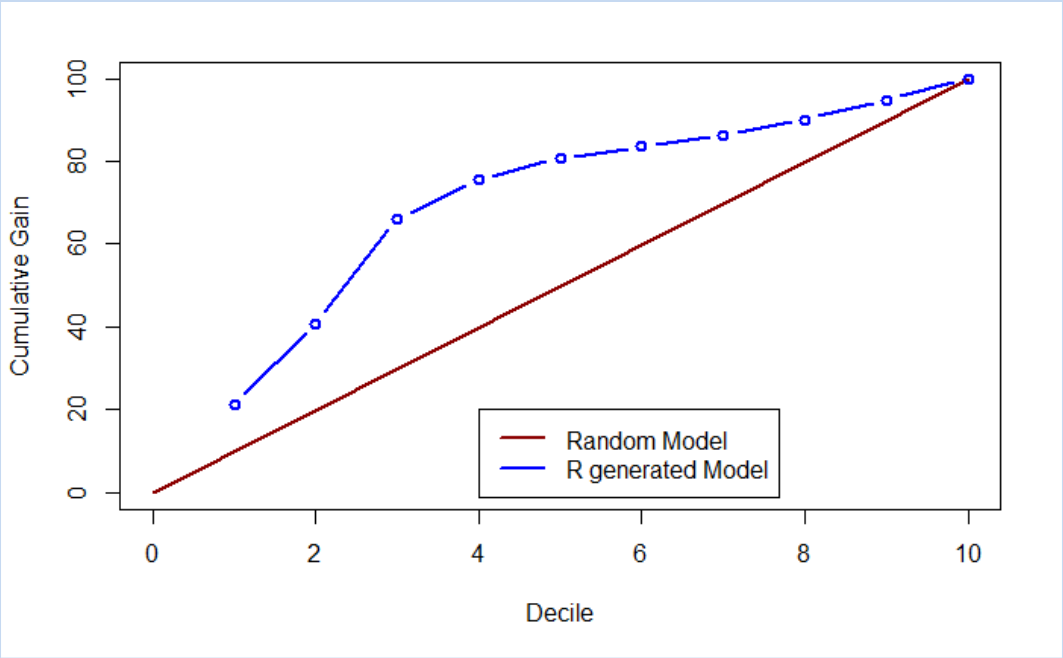
- As cutoff increases sensitivity decreases.
- As cutoff increases specificity and Accuracy increases.

From the model, the following values are obtained:

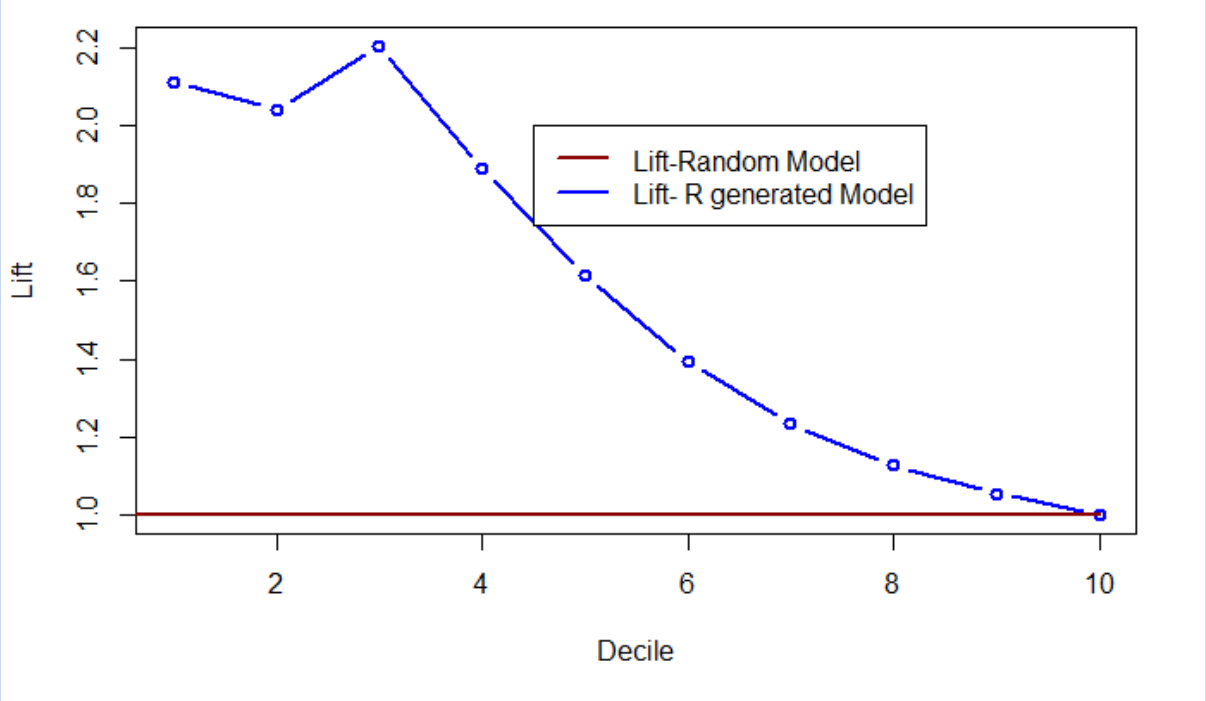
Accuracy	Sensitivity	Specificity
0.735	0.742	0.733

We can say that the model predicts with **73.5% accuracy**

Lift and Gain Charts:

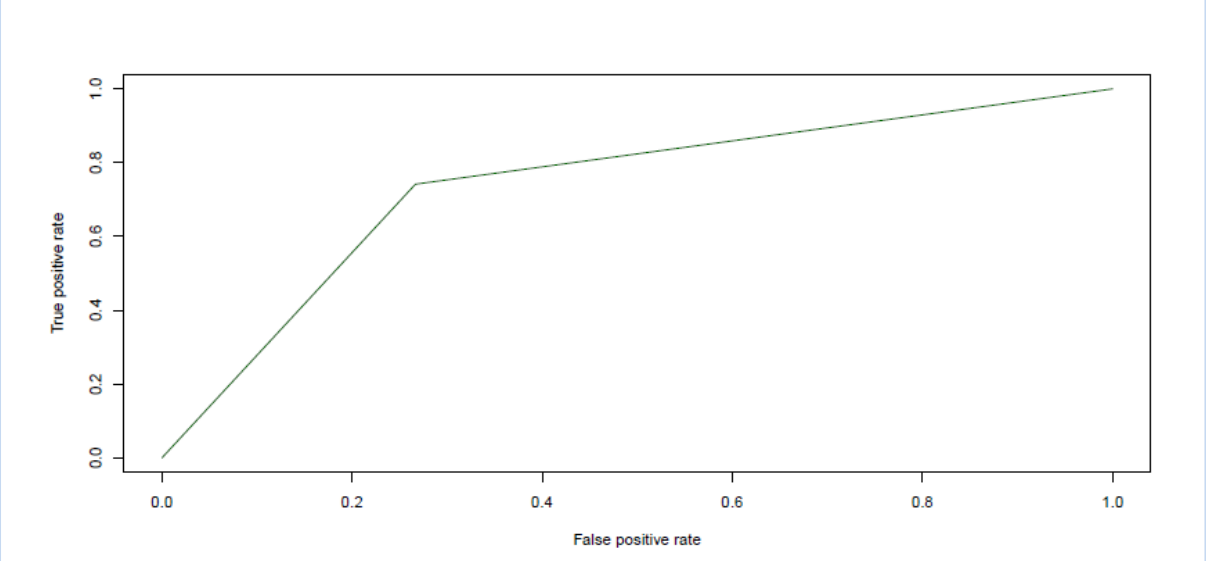


- We are able to capture 66% of the people who will leave in the 3rd decile.
- From the Gain chart, we can determine that the top 30% contains approximately 66% of employees that are going to leave the organization.
- At the 3rd decile, the lift is 2.21 and hence the model is 2 times better than a random model.



K-S Statistics:

The k-s statistic is 0.47 and hence 47%. The ideal k-s statistic should be > 40%.



Conclusion

- In this case Log Regression was the best model, as it predicted a higher area under the curve and a better accuracy (73%) compared to other model I tried like Random Forest.
- Because of the imbalance in dataset the model is more biased towards non-attrition class

Recommendations

- The company's management should not make employees travel too much.
- They should be kept under great leaders (managers).
- The management should keep a track of employees working for a greater number of hours and should be given breaks at regular intervals of time or their workload should be distributed equally along with other team members.
- While selecting a new candidate the HR should check on how many companies has the person worked in earlier.
- Environmental satisfaction should be addressed right away as it is the key to creating a better workforce.
- The company should ensure that project estimation is given correctly so that there is no pressure of delivery at the last moment.
- The motivation for employees in terms of vouchers and recognition.

References

[1]
https://www.researchgate.net/publication/322896996_Employee_Attrition_and_Employee_Retention-Challenges_Suggestions
[2] https://towardsdatascience.com/modelling-binary-logistic-regression-using-r-research-oriented-modelling-and-interpretation-f67b3a954101
[3]
https://www.r-graph-gallery.com/
[4]
Dataset: https://www.kaggle.com/vjchoudhary7/hr-analytics-case-study
[5]
https://towardsdatascience.com/class-imbalance-a-classification-headache-1939297ff4a4