HR Analytics
Case Study

Group Name:

1. Member name: Hariharan S
2. Member name: Thripthi Raj
3. Member name: Sivaiah Aelasrolu
4. Member name: Abhishek Pandey

Objectives

To model probability of attrition using logistic regression to help the client understand:

#A. What factors they should focus on, in order to curb attrition.

#B. Which of these variables is most important and needs to be addressed right away.

Employee Data

Data

Employee general data

4410 records and 27 columns

Employee Survey data

4410 records and 4 columns

Manager Survey data

4410 records and 2 columns

In time and Out time

4410 records and 250 columns (in time and out time for 1st Jan 2015 and 31st Dec 2015)

Data Observation and Cleaning

Data

Identified the NA values for “No of companies worked” and “ Total working years” in general data

Analysed the monthly income, job role, years at company and total working years and replaced NA values with appropriate values.

Identified NA values in Employee survey data and replaces the same with appropriate values.

No duplicate records identified

No case sensitivity issues

Converted the date column as date type in in time and out time

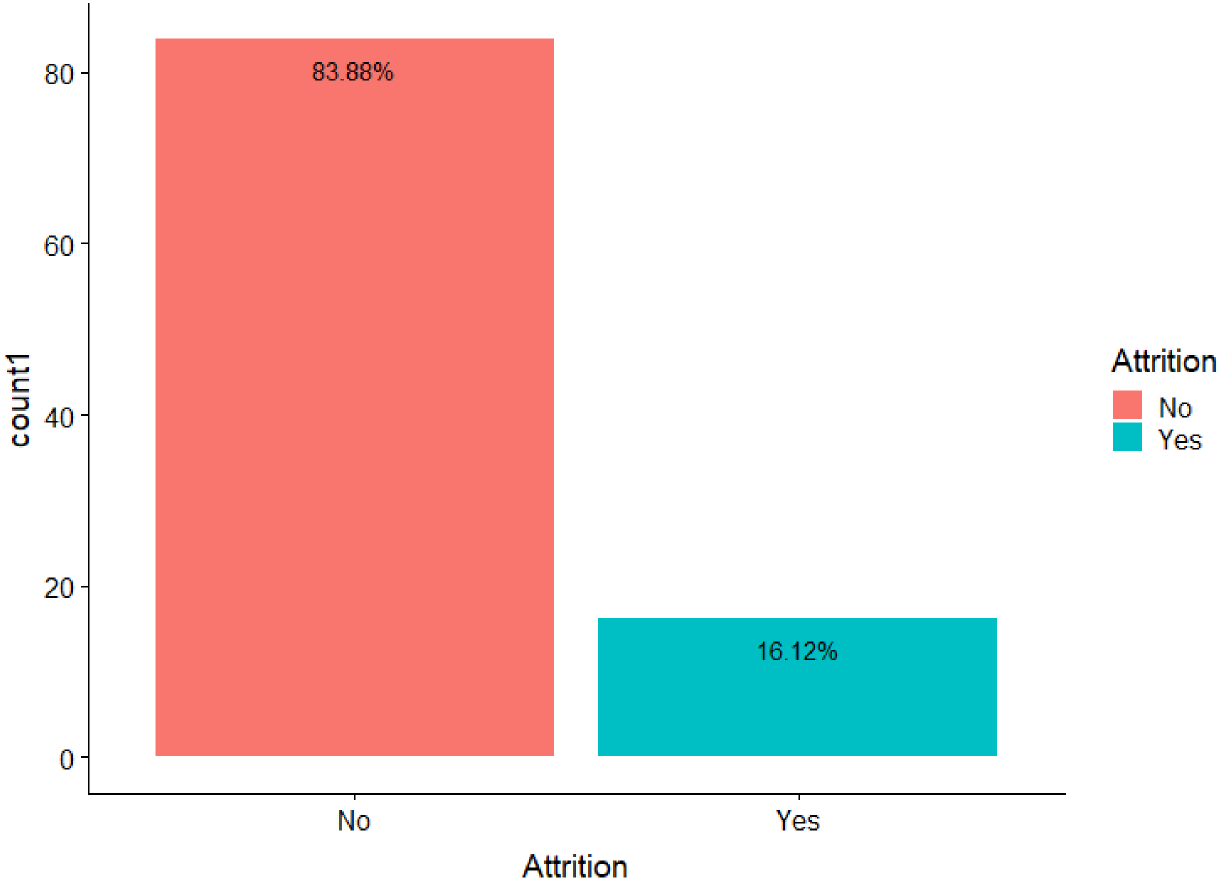
Derived Metrics

Calculated derived metrics for employees

- Average hours
- Leaves

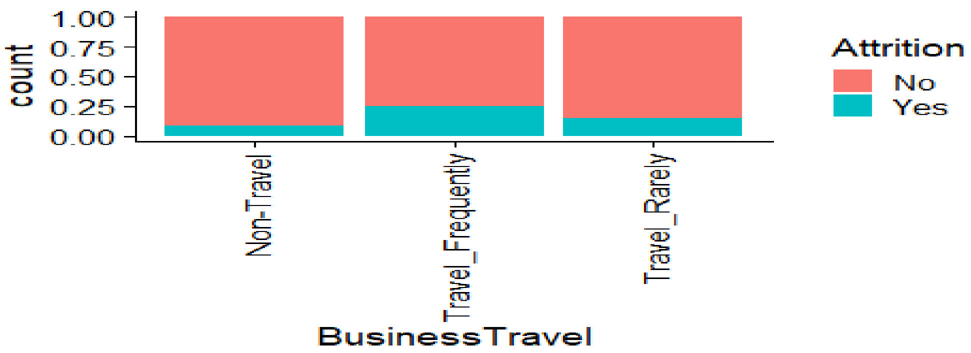
Created dummy variables for categorical variables for model design

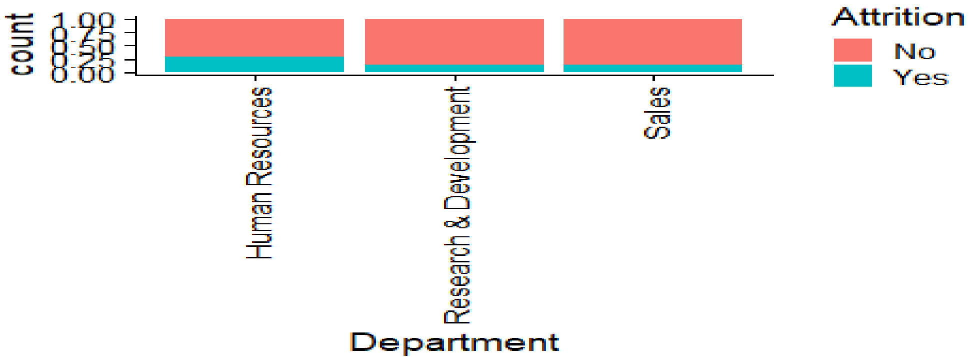
EDA

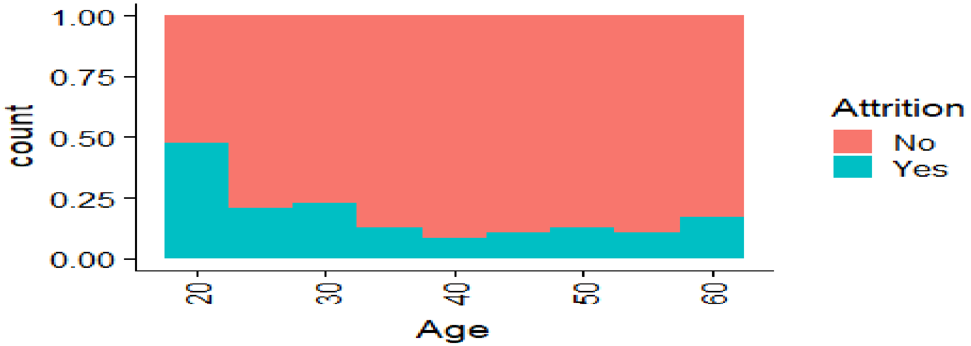


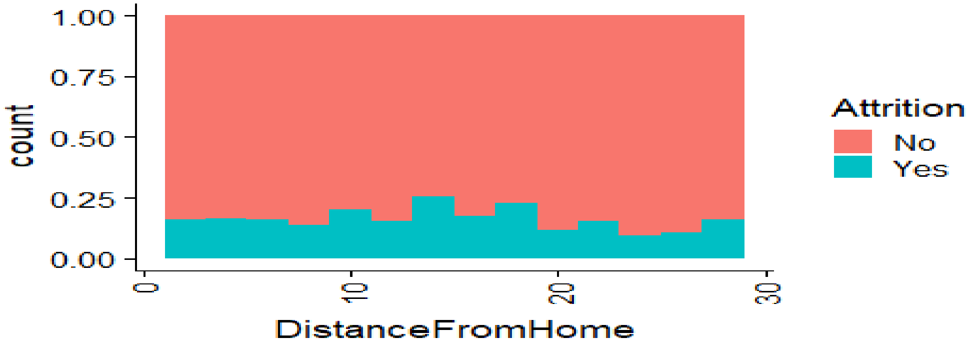
- Attrition rate is 16.12%

EDA



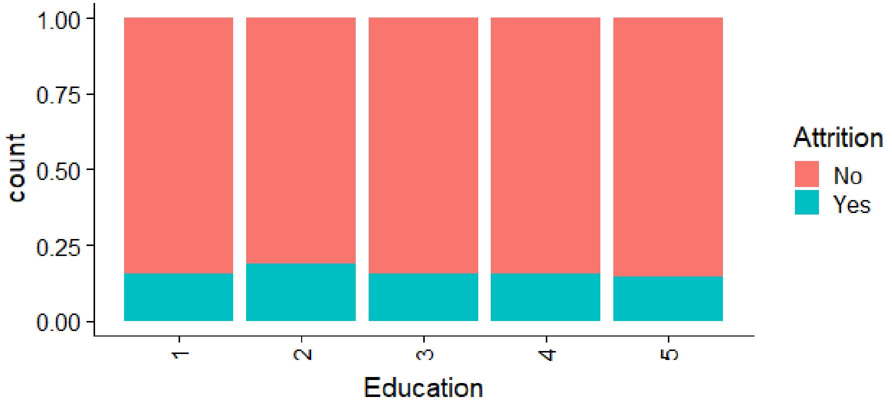


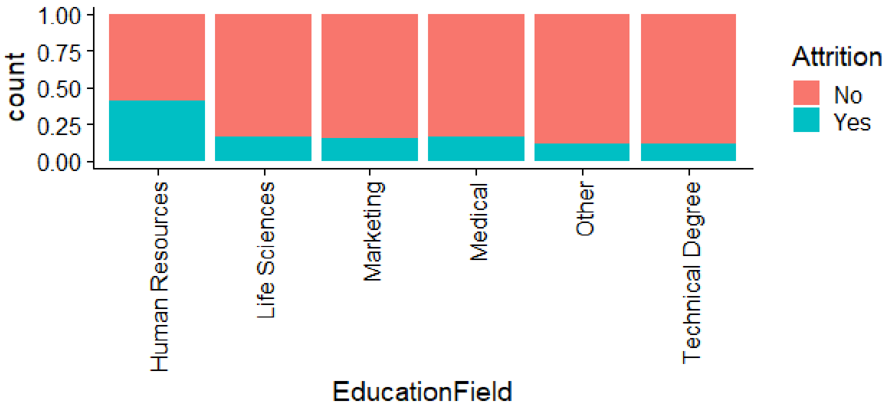


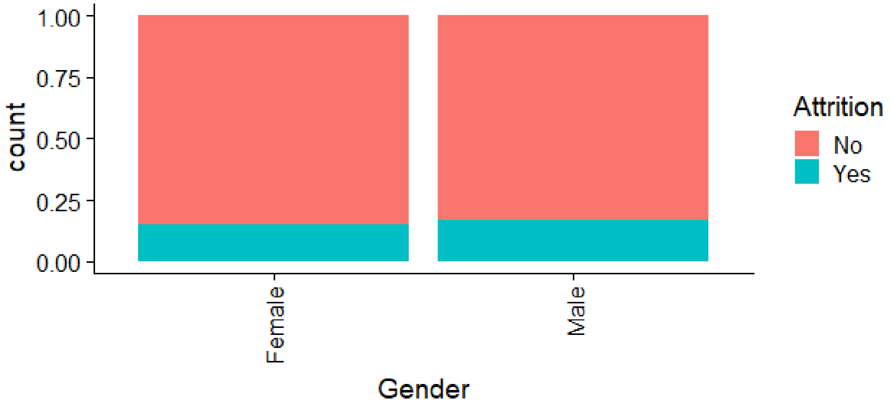


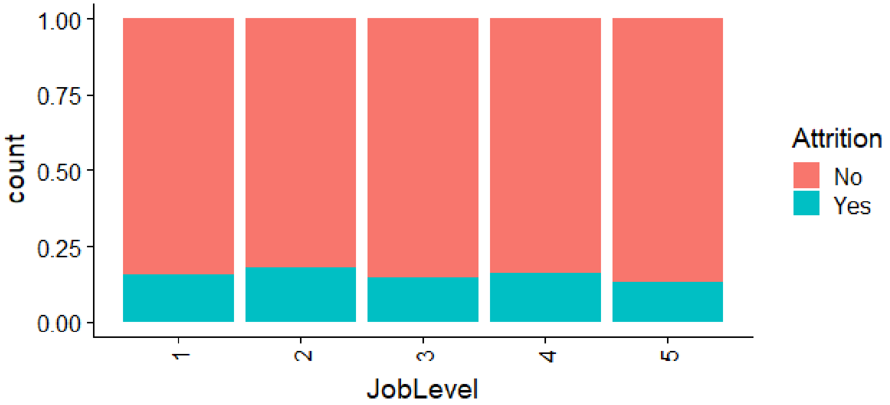
- Employees who travel rarely have the highest count of attrition. However when looking at the proportion 24.9% employees who travel frequently and 15% of employees who travel rarely are likely to attrition.
- Research & Development has the highest count of attrition employees while Human Resource has the highest proportion of Attrition members in a department
- Younger Employees are more likely to attrition than older ones
- No clear distinction between distance from home and Attrition

EDA



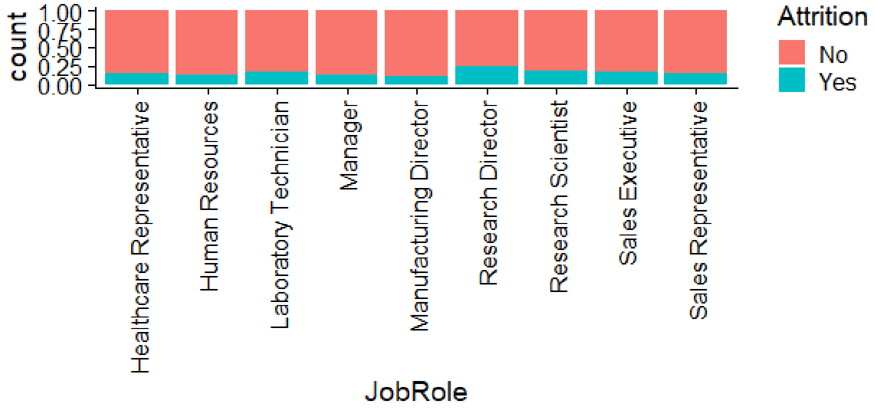


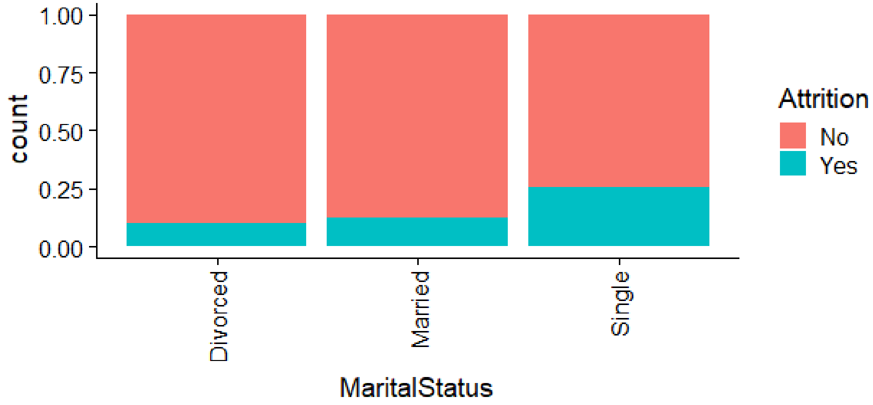


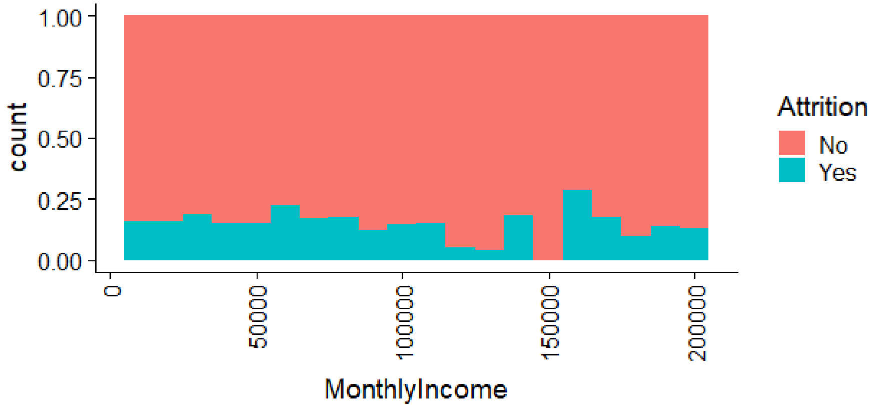


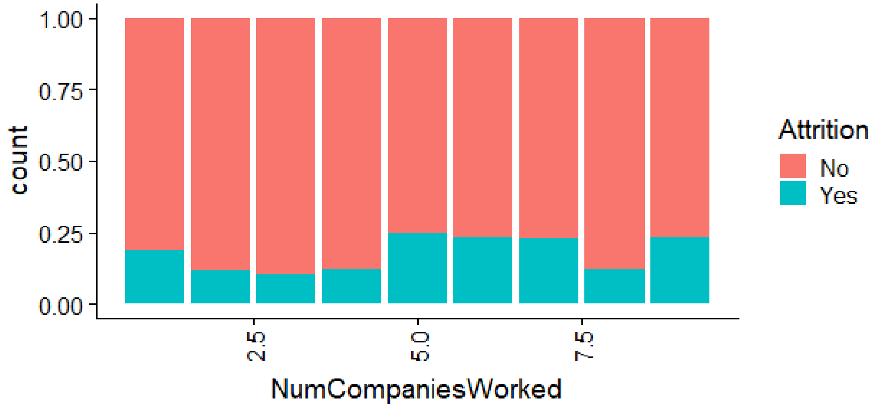
- Employees with a Life Sciences educational field account for the highest number of Attrition. However, employees with Human resources background accounts for the highest proportion of Attrition [Almost 40% of all employees with HR background Attrition
- Seems to be no clear distinction of Attrition Rate between Male and Female
- No clear distinction between Attrition Rate between Job levels

EDA



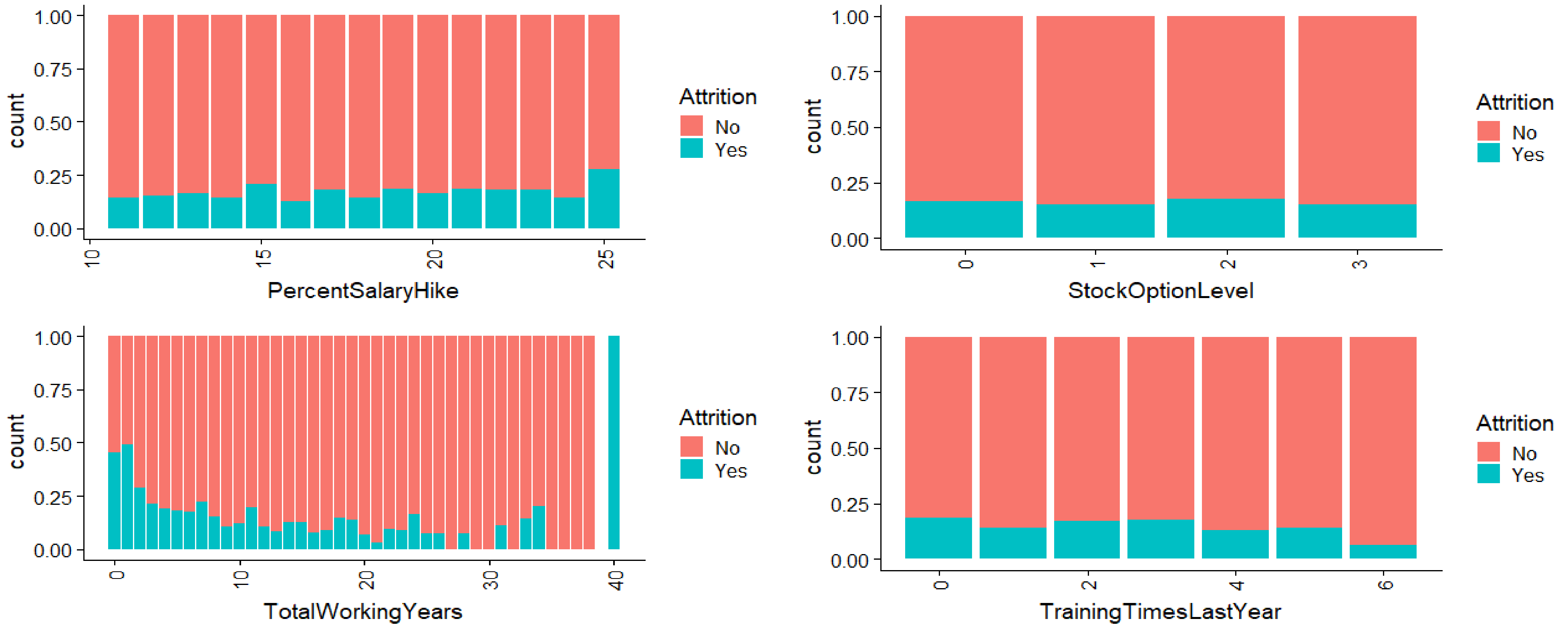






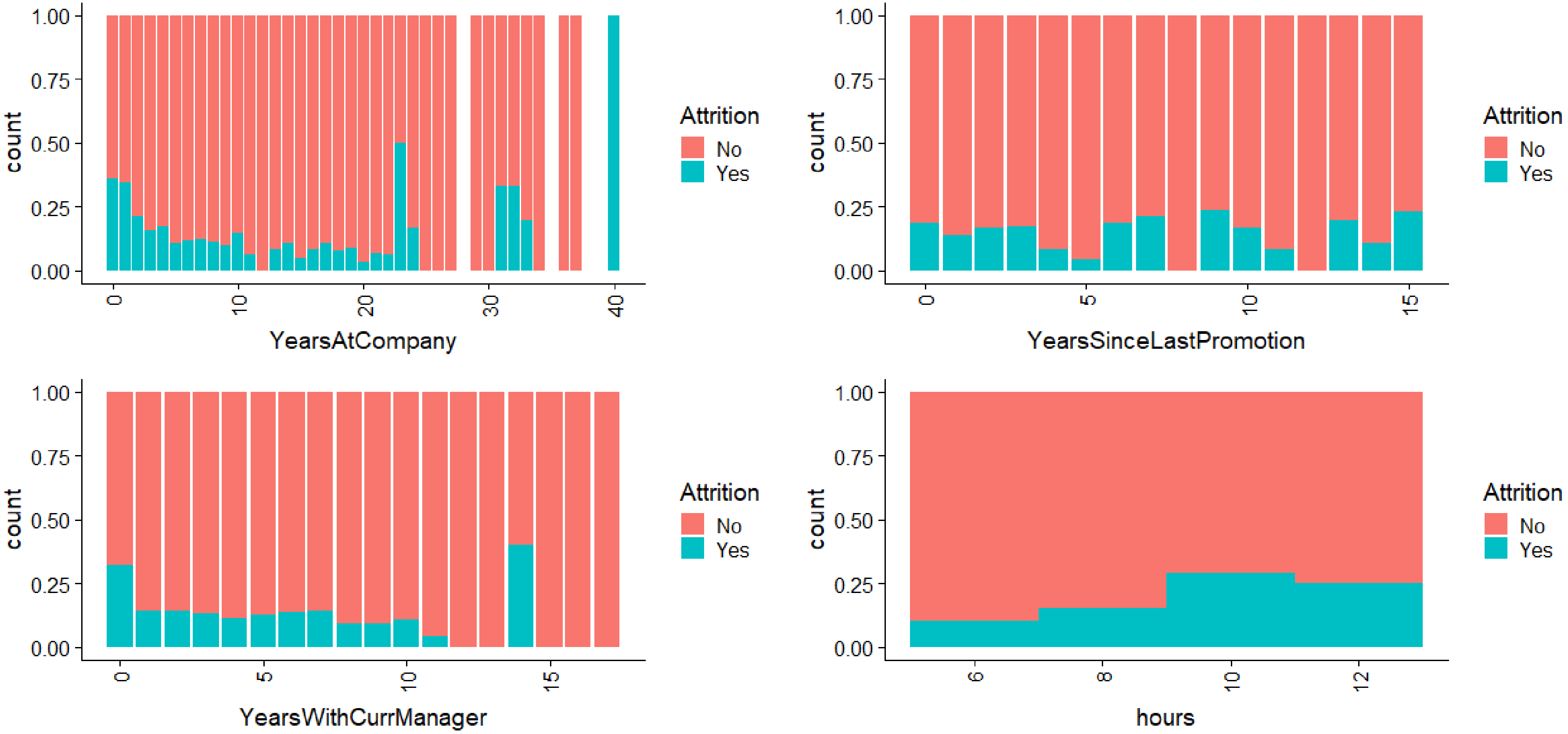
- The attrition rate for research director is marginally higher but there seems to be no clear distinction of Attrition Rate for Job roles
- Single Employees are significantly more likely to attrition
- No clear distinction between monthly income and attrition
- Employees who have worked in less than 2 companies or more than 5 companies are at a higher risk of

EDA



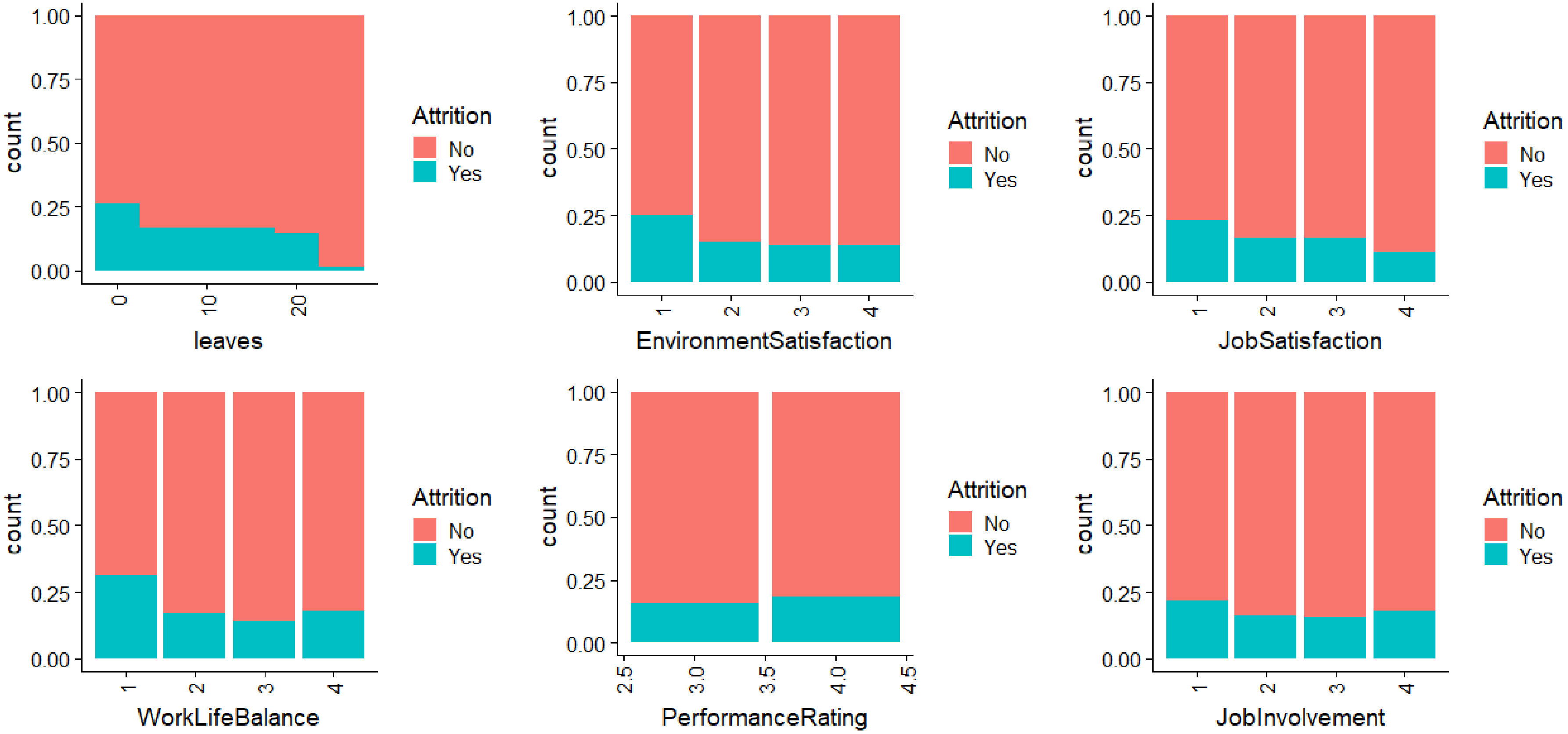
- No clear relation between salary hike and attrition status. Single Employees are significantly more likely to attrition
- Employees with Total Working Years <8 years are at a higher risk of attrition
- No distinct difference in training frequency with respect to attrition
- No clear relation between attrition and Stock level

EDA



- Employees with YearsAtCompany<5 are at a higher risk of attrition
- No clear distinction between last promotion and attrition
- Employees with less than 1 Year worked under current Manager are at a higher rate of attrition
- Employees who worked more hours are at a higher risk of attrition

EDA



- No clear distinction between leaves and attrition
- Employees with Low Work Environment Satisfaction levels have a very high risk of attrition
- Employees with Low Job Satisfaction levels have a very high risk of attrition
- High for less work balance lift employees
- Seems difference for performance rating and job involvement categories

Variable Selection Approach

- Numeric and categorical variables are available in provided data
- Few numeric variables have only few repeated values which can be converted as categorical variables by grouping in different bins
- Created model using below 2 different approaches and evaluated the models

  Approach 1:

  Converting all numeric variables except income as categorical variables by grouping into different bins

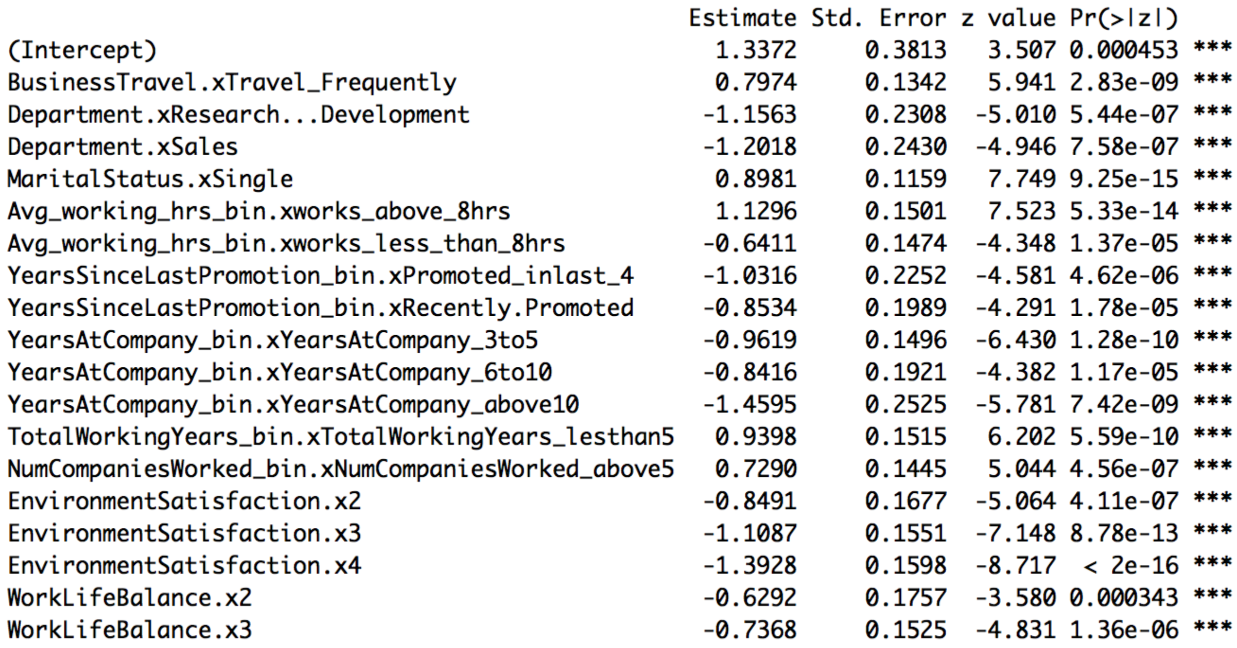
  Approach 2:

  Considering all numeric variables as continuous variables and performing scale for all variables

Approach 1

- **Significant variables**

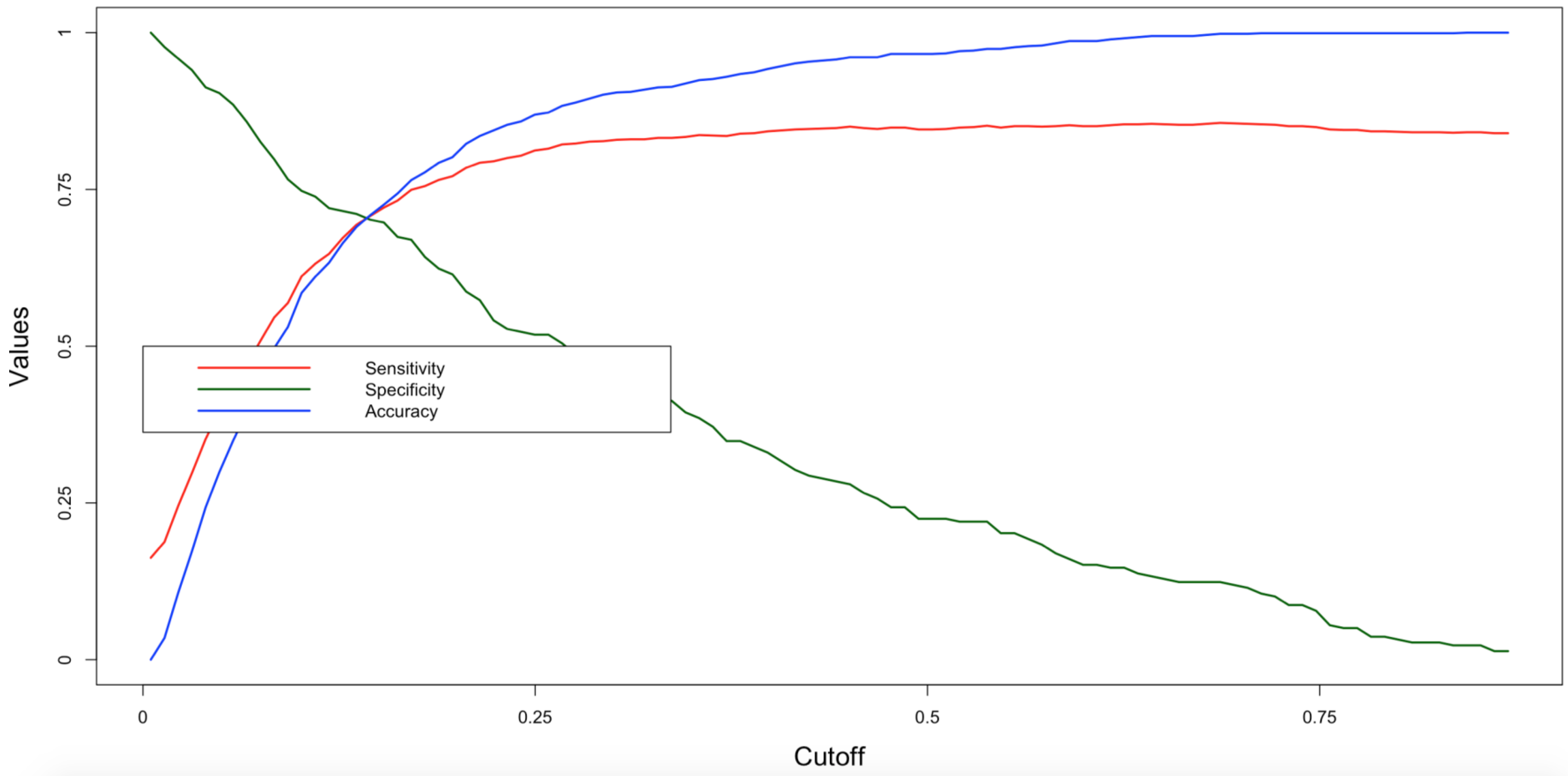
1.BusinessTravel-Travel_Frequently
2.Department-Research&Development,Sales
3.MaritalStatus-Single
4.Avg_working_hrs- above_8hrs,less_than_8hrs
5.YearsSinceLastPromotion - Promoted_inlast_4,Recently.Promoted
6.YearsAtCompany- YearsAtCompany_3to5,YearsAtCompany_6to10,YearsAtCompany_above10
7.TotalWorkingYears-TotalWorkingYears_lesthan5
8.NumCompaniesWorked-NumCompaniesWorked_above5
9.EnvironmentSatisfaction-2,3,4
10.WorkLifeBalance-2,3



```
                                                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                          1.3372     0.3813   3.507 0.000453 ***
BusinessTravel.xTravel_Frequently                    0.7974     0.1342   5.941 2.83e-09 ***
Department.xResearch...Development                  -1.1563     0.2308  -5.010 5.44e-07 ***
Department.xSales                                   -1.2018     0.2430  -4.946 7.58e-07 ***
MaritalStatus.xSingle                                0.8981     0.1159   7.749 9.25e-15 ***
Avg_working_hrs_bin.xworks_above_8hrs                1.1296     0.1501   7.523 5.33e-14 ***
Avg_working_hrs_bin.xworks_less_than_8hrs           -0.6411     0.1474  -4.348 1.37e-05 ***
YearsSinceLastPromotion_bin.xPromoted_inlast_4      -1.0316     0.2252  -4.581 4.62e-06 ***
YearsSinceLastPromotion_bin.xRecently.Promoted      -0.8534     0.1989  -4.291 1.78e-05 ***
YearsAtCompany_bin.xYearsAtCompany_3to5             -0.9619     0.1496  -6.430 1.28e-10 ***
YearsAtCompany_bin.xYearsAtCompany_6to10            -0.8416     0.1921  -4.382 1.17e-05 ***
YearsAtCompany_bin.xYearsAtCompany_above10          -1.4595     0.2525  -5.781 7.42e-09 ***
TotalWorkingYears_bin.xTotalWorkingYears_lesthan5    0.9398     0.1515   6.202 5.59e-10 ***
NumCompaniesWorked_bin.xNumCompaniesWorked_above5    0.7290     0.1445   5.044 4.56e-07 ***
EnvironmentSatisfaction.x2                          -0.8491     0.1677  -5.064 4.11e-07 ***
EnvironmentSatisfaction.x3                          -1.1087     0.1551  -7.148 8.78e-13 ***
EnvironmentSatisfaction.x4                          -1.3928     0.1598  -8.717  < 2e-16 ***
WorkLifeBalance.x2                                  -0.6292     0.1757  -3.580 0.000343 ***
WorkLifeBalance.x3                                  -0.7368     0.1525  -4.831 1.36e-06 ***
```

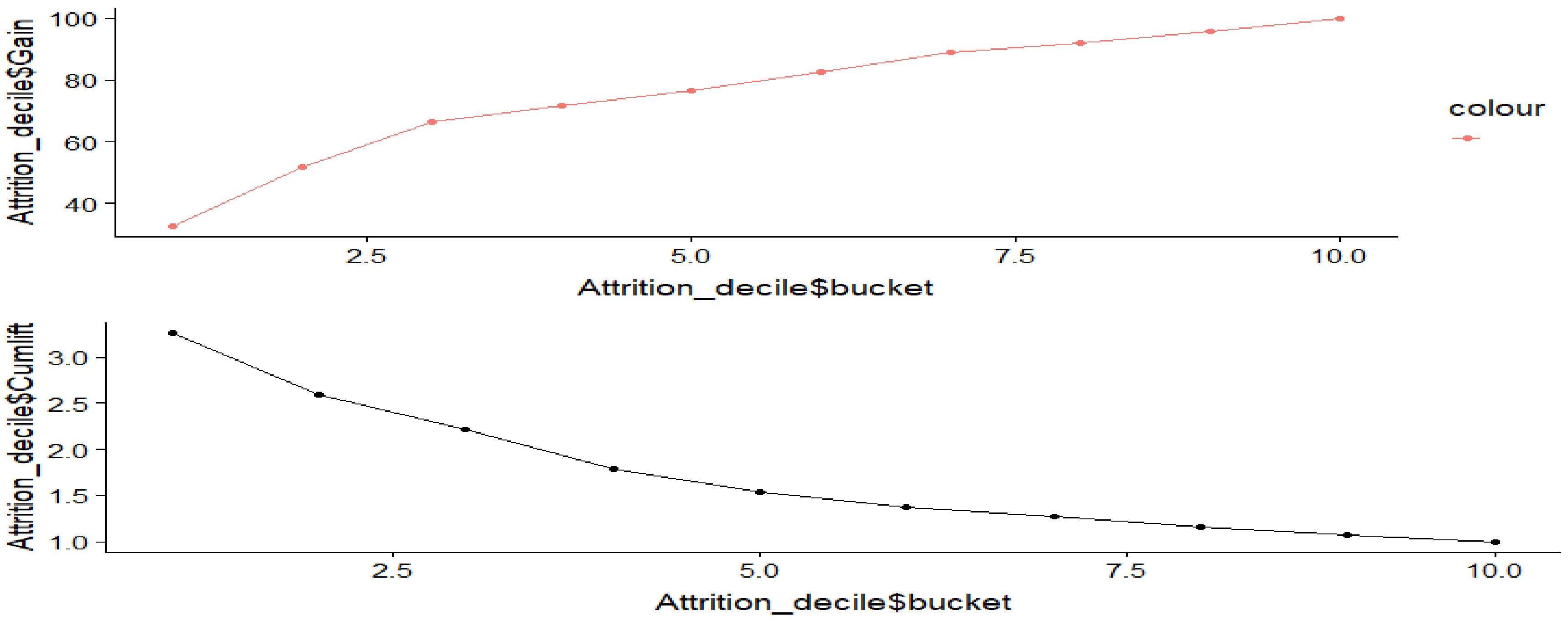
Approach 1 - Graphs

- Optimal probability cutoff chosen is 0.14
- Accuracy 0.699478
- Sensitivity 0.7110092
- Specificity 0.6972395



Approach 1 - Graphs

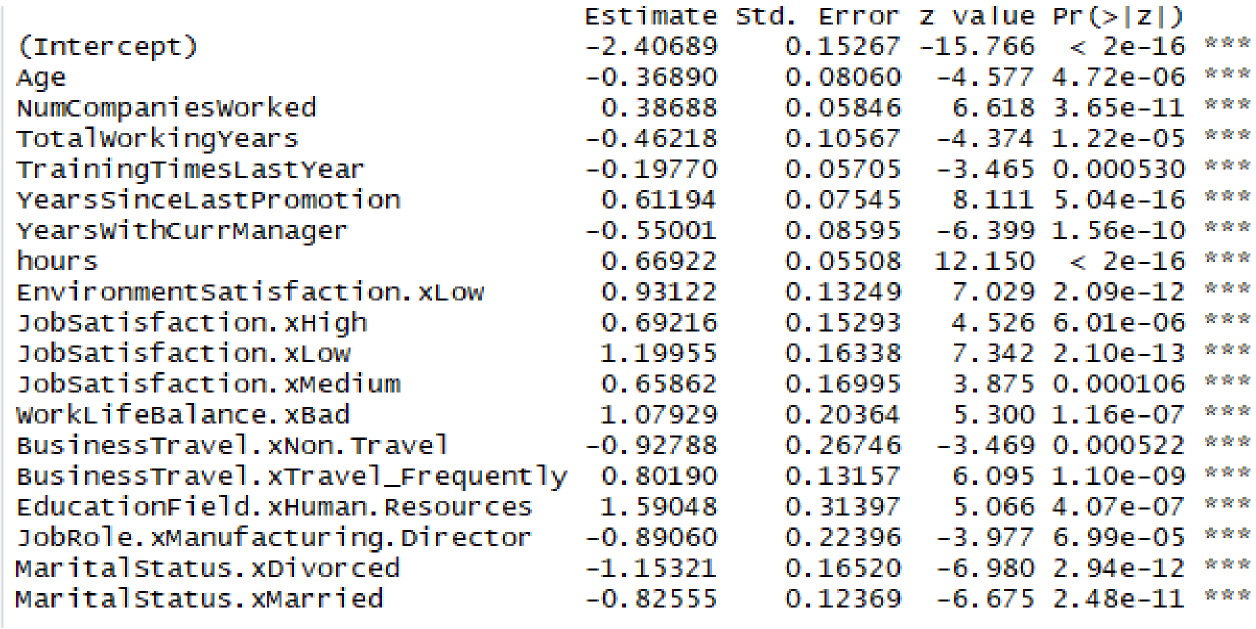
- Lift and Gain



Approach 2

- **Significant variables**

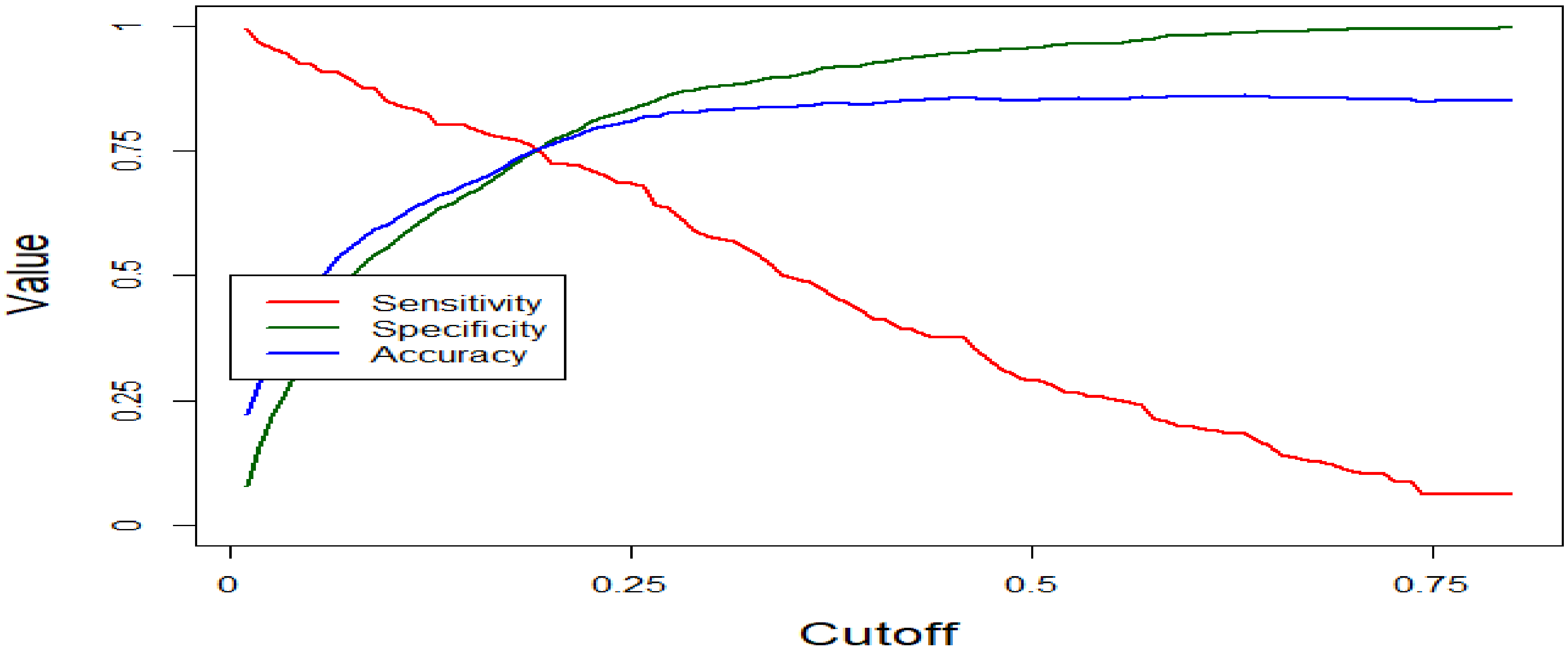
  Got 18 significant variables in approach 2



```
                                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      -2.40689    0.15267 -15.766  < 2e-16 ***
Age                              -0.36890    0.08060  -4.577 4.72e-06 ***
NumCompaniesWorked                0.38688    0.05846   6.618 3.65e-11 ***
TotalWorkingYears                -0.46218    0.10567  -4.374 1.22e-05 ***
TrainingTimesLastYear            -0.19770    0.05705  -3.465 0.000530 ***
YearsSinceLastPromotion           0.61194    0.07545   8.111 5.04e-16 ***
YearsWithCurrManager             -0.55001    0.08595  -6.399 1.56e-10 ***
hours                             0.66922    0.05508  12.150  < 2e-16 ***
EnvironmentSatisfaction.xLow      0.93122    0.13249   7.029 2.09e-12 ***
JobSatisfaction.xHigh             0.69216    0.15293   4.526 6.01e-06 ***
JobSatisfaction.xLow              1.19955    0.16338   7.342 2.10e-13 ***
JobSatisfaction.xMedium           0.65862    0.16995   3.875 0.000106 ***
WorkLifeBalance.xBad              1.07929    0.20364   5.300 1.16e-07 ***
BusinessTravel.xNon.Travel       -0.92788    0.26746  -3.469 0.000522 ***
BusinessTravel.xTravel_Frequently 0.80190    0.13157   6.095 1.10e-09 ***
EducationField.xHuman.Resources   1.59048    0.31397   5.066 4.07e-07 ***
JobRole.xManufacturing.Director  -0.89060    0.22396  -3.977 6.99e-05 ***
MaritalStatus.xDivorced          -1.15321    0.16520  -6.980 2.94e-12 ***
MaritalStatus.xMarried           -0.82555    0.12369  -6.675 2.48e-11 ***
```

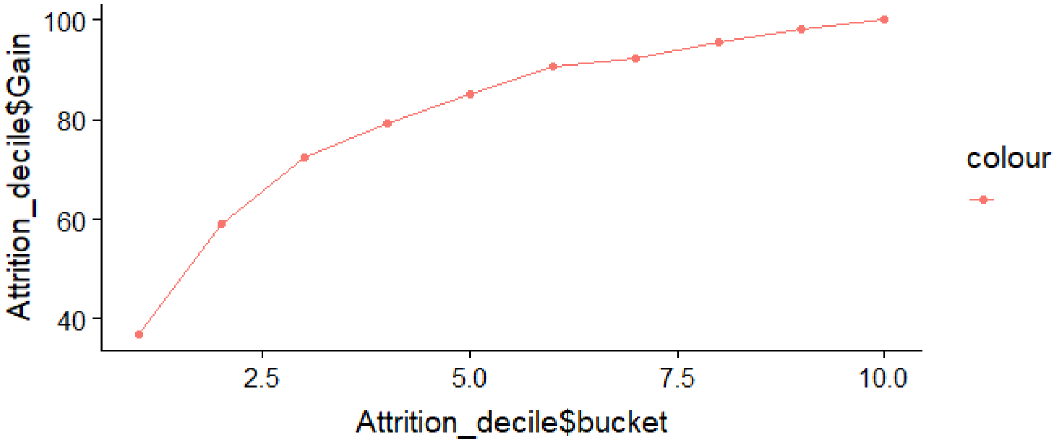
Approach 2 - Graphs

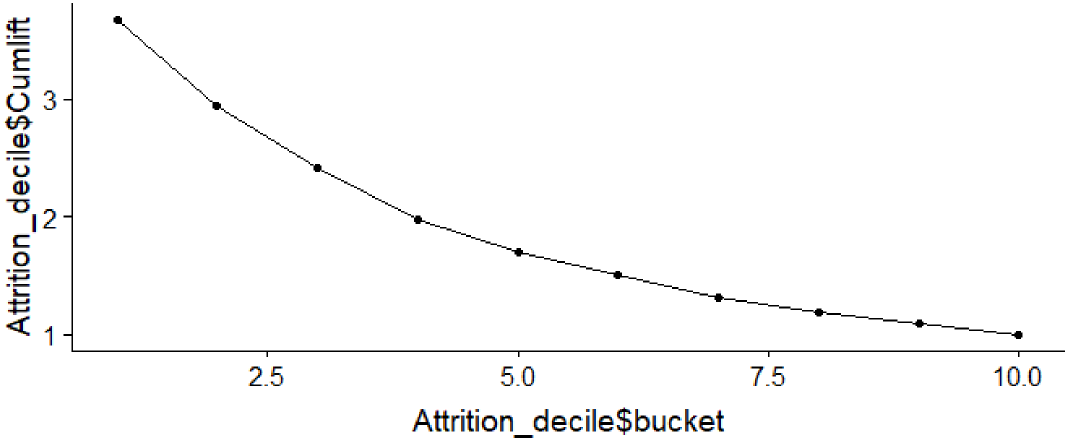
- Optimal probability cutoff chosen is 0.19
- Accuracy 0.855
- Sensitivity 0.28
- Specificity 0.961



Approach 2 - Graphs

- Lift and Gain


Attrition_decile$Gain
Attrition_decile$bucket
colour


Attrition_decile$Cumlift
Attrition_decile$bucket