HR ANALYTICS CASE STUDY
PROBABILITY OF ATTRITION

Group Name : Best Riders

1. Member name –Anuj Arya
2. Member name –Asim Pattnaik
3. Member name –Mohammed Suhail Y
4. Member name –Rakesh Bosu

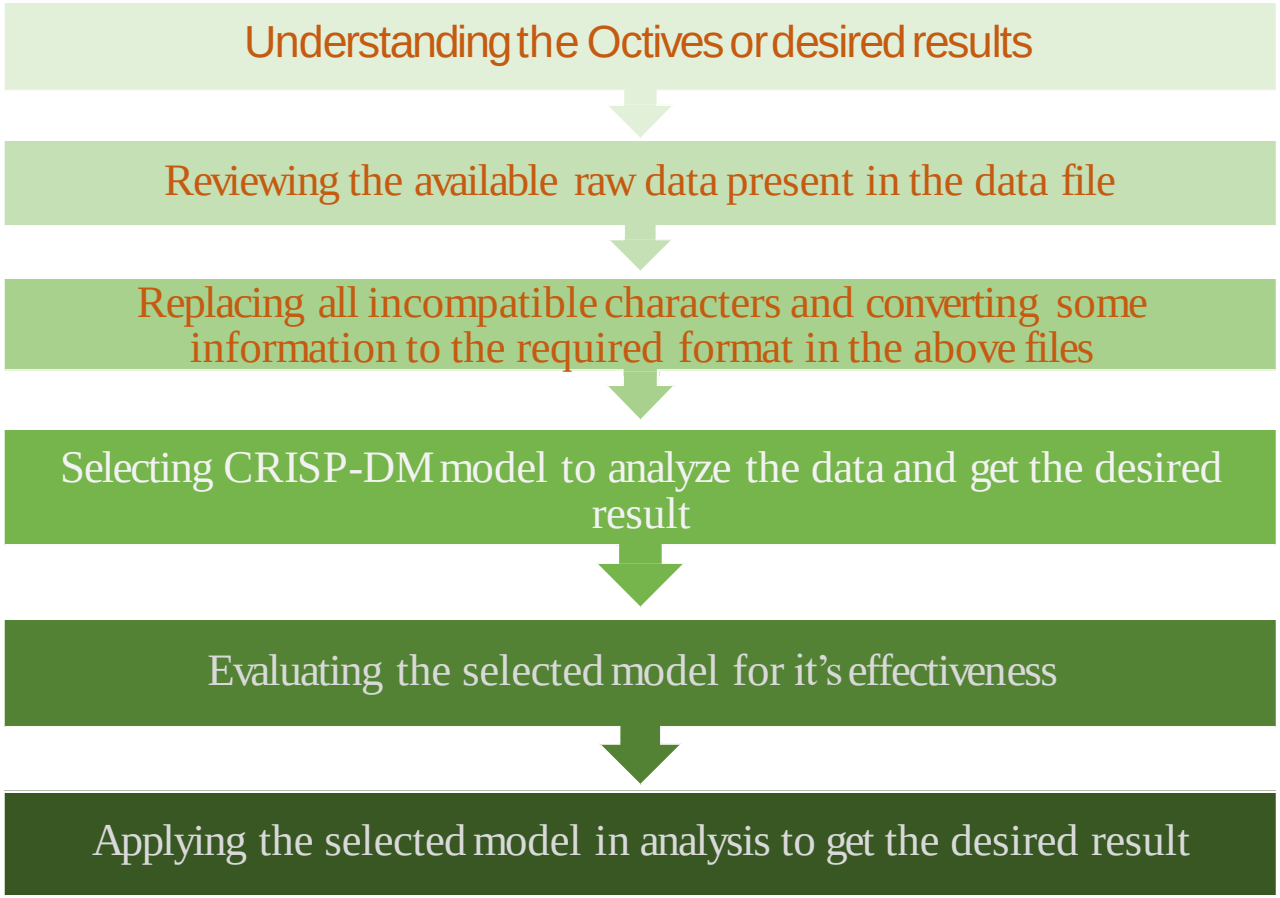
Problem Statement

A large company named XYZ, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company

AIM:

Required to model the probability of attrition using a logistic regression. The results thus obtained will be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay.

CRISM-DM Modeling



Data Exploration

4410 employees records with following information

- General data (Age, Gender, Income, Experience,Attrition etc.)
- Employee survey about environment & job satisfaction, work life balance
- Manager survey about job involvement & performance rating.
- Employees in-time and out-time data

Continuous Variables :

- Age
- DistanceFromHome
- EmployeeCount
- MonthlyIncome
- NumCompaniesWorked
- PercentSalaryHike
- StandardHours
- TotalWorkingYears
- TrainingTimesLastYear
- YearsAtCompany,
- YearsSinceLastPromotion
- YearsWithCurrManager

Data Exploration

Ordered Categorical Variables:

- Education
- JobLevel
- MaritalStatus
- StockOptionLevel
- EnvironmentSatisfaction
- JobSatisfaction
- WorkLifeBalance
- JobInvolvement
- PerformanceRating

Unordered Categorical Variables:

- BusinessTravel
- Department
- EducationField
- Gender
- JobRole
- Over18

Data Preparation and Processing

- Missing Values:
  - Replaced missing values in below columns with Median
    - EnvironmentSatisfaction
    - JobSatisfaction
    - WorkLifeBalance
  - Replaced missing values in "NumCompaniesWorked" with "TotalWorkingYears" with Mean Values

- Removed outliers in below columns based on quantile function
    - TotalWorkingYears
    - YearsAtCompany
    - YearsWithCurrManager

- Derived new metrics(Average Office Hours and Leave Count) based on employee In and Out time data.

- Scaled continuous variables and created dummy variables for categorical variables for modelling

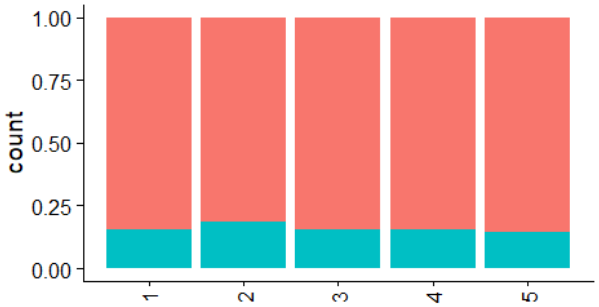
In and Out Time:

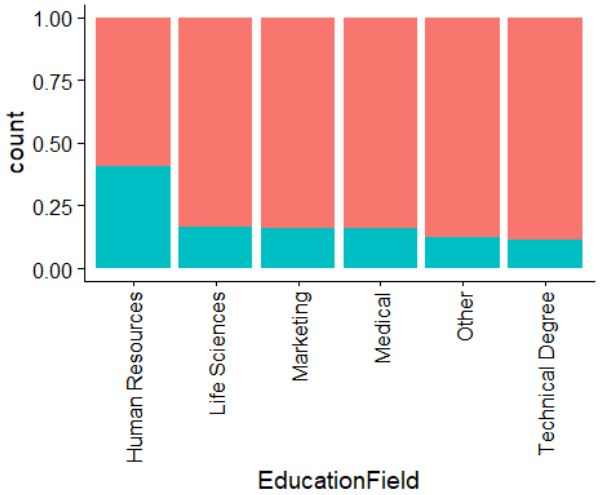
Leave Count : Calculate the number of hours that the person was on leave (excluding public holidays)

Average office hours: Calculate the hours spent by employee in office check and if it is greater than 8 hours.

Data Analysis

Comparing Education, Education Field and Job Level



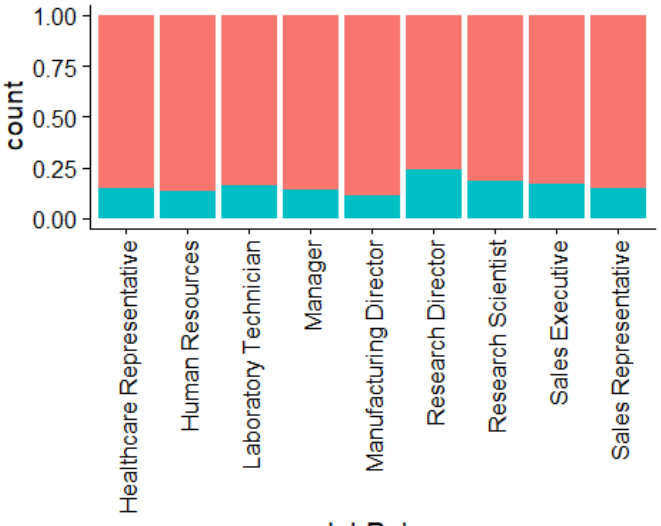


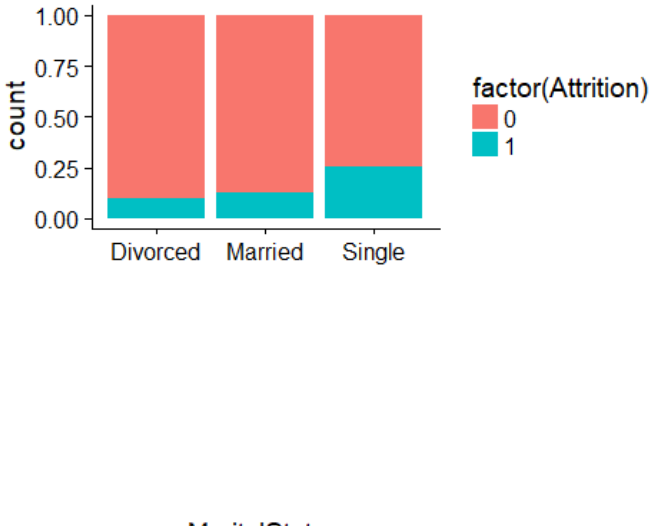
Higher Attrition is found among

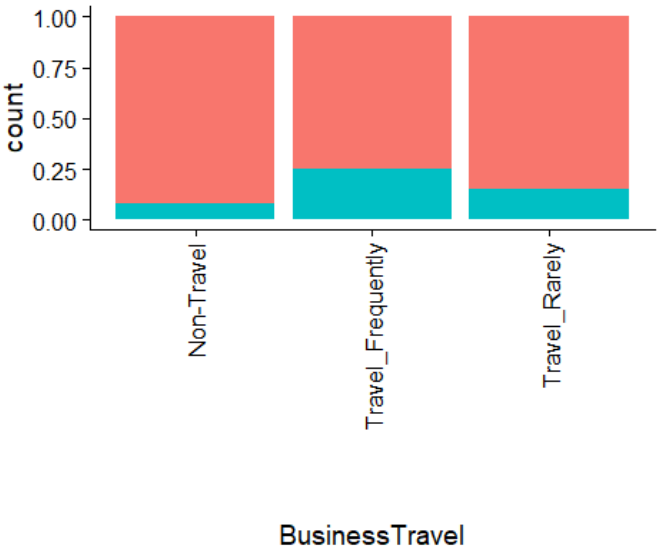
- Education 2
- Education Filed -Human Resource
- Job level 2

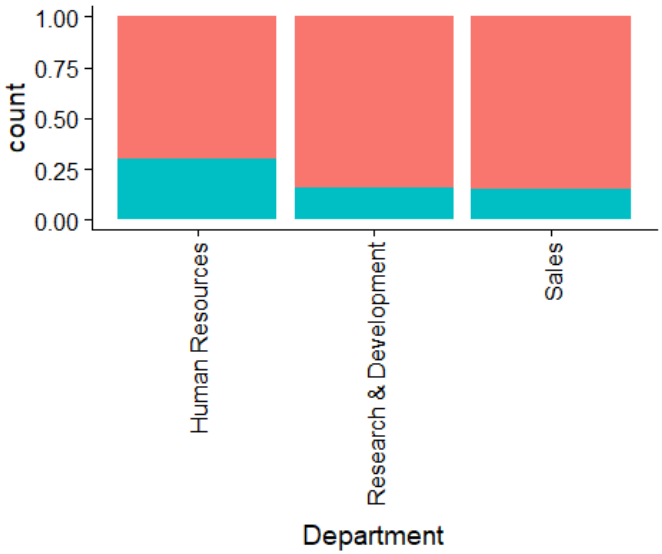
Data Analysis

Comparing Jobrole, Marital status, Business travel and Department









Higher Attrition is found among

- Job roles with Research Director
- Marital status - Single
- Employees with Business travel
- HR Department

Final Model - Predictor Variables

```
> summary(model_28)

Call:
glm(formula = Attrition ~ Age + TrainingTimesLastYear + YearsSinceLastPromotion +
    YearsWithCurrManager + avg_ofc_hours + EnvironmentSatisfaction.x2 +
    EnvironmentSatisfaction.x3 + EnvironmentSatisfaction.x4 +
    JobSatisfaction.x2 + JobSatisfaction.x3 + JobSatisfaction.x4 +
    WorkLifeBalance.x3 + MaritalStatus.xSingle, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7591  -0.5691  -0.3757  -0.2171   3.2749

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.13129    0.16331  -6.927 4.29e-12 ***
Age                        -0.45799    0.05934  -7.718 1.18e-14 ***
TrainingTimesLastYear      -0.18525    0.05568  -3.327 0.000879 ***
YearsSinceLastPromotion     0.37649    0.06962   5.408 6.37e-08 ***
YearsWithCurrManager       -0.63028    0.07616  -8.275  < 2e-16 ***
avg_ofc_hours               1.29066    0.11186  11.538  < 2e-16 ***
EnvironmentSatisfaction.x2 -0.92218    0.16538  -5.576 2.46e-08 ***
EnvironmentSatisfaction.x3 -0.96936    0.14768  -6.564 5.24e-11 ***
EnvironmentSatisfaction.x4 -1.19211    0.15315  -7.784 7.02e-15 ***
JobSatisfaction.x2         -0.59763    0.16798  -3.558 0.000374 ***
JobSatisfaction.x3         -0.50662    0.14501  -3.494 0.000477 ***
JobSatisfaction.x4         -1.18175    0.15873  -7.445 9.70e-14 ***
WorkLifeBalance.x3         -0.36410    0.10958  -3.323 0.000891 ***
MaritalStatus.xSingle       1.04169    0.11138   9.353  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2728.0  on 3086  degrees of freedom
Residual deviance: 2222.5  on 3073  degrees of freedom
AIC: 2250.5

Number of Fisher Scoring iterations: 5
```

Model Evaluation (Cut Off- 50%)

Probability cut off at 50%:

- Accuracy of the model - 85%
- Sensitivity (True Positive Rate) - 22%
- Specificity (True Negative Rate) - 98%

		Predicted Attrition	
		No	Yes
Actual Attrition	No	1080	30
	Yes	159	54

Analysis:

- Even though the accuracy of the model is high, the sensitivity of the model is very low. Since we need Attrition rate, we need to maximize the sensitivity of the model.

Model Evaluation (Cut Off- 40%)

Probability cut off at 40%:

- Accuracy of the model - 85%
- Sensitivity (True Positive Rate) - 31%
- Specificity (True Negative Rate) - 95%

		Predicted Attrition	
		No	Yes
Actual Attrition	No	1063	47
	Yes	145	68

Analysis:

- Though Sensitivity has increased, it's still low.

Model Evaluation (Optimal Cut Off)

Based on analysis, the optimal Cut Off value is 15.36%

At Optimal Cut Off:

- Accuracy of the model - 74%
- Sensitivity (True Positive Rate) - 75%
- Specificity (True Negative Rate) - 74%

		Predicted Attrition	
		No	Yes
Actual Attrition	No	818	292
	Yes	52	161

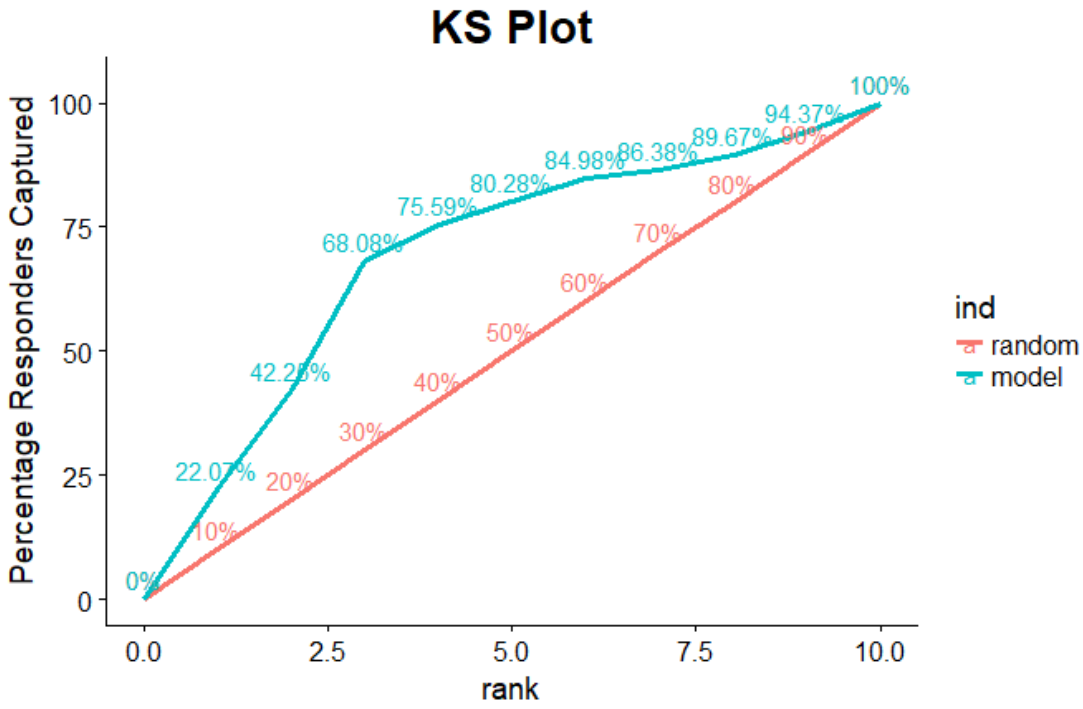
Analysis:

- We get a high sensitivity rate of 75%

Model Evaluation – KS Statistic

- KS Test measures to check whether model is able to separate events and non-events. In our model, it checks whether the our model is able to distinguish between employees who will leave and employee who won't leave.

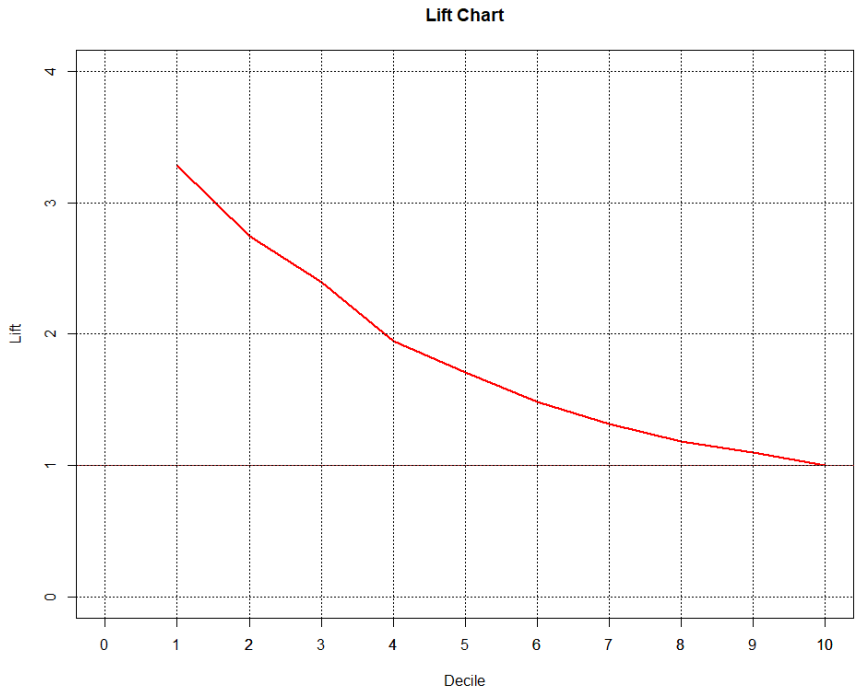
- Ideally, the KS score lies between 40 and 70. In this case, KS score > 40 (i.e. 49.2%), which is good model.



Model Evaluation – Lift And Gain

Lift Chart: Based on lift chart, we identified our model is outperforming a random model. We can predict true positive rate more efficiently using our model compared to a random model.

Gain Chart: Based on gain chart, we identified our model is performing well as shown by KS Statistic chart.



Conclusion

Based on the Logistic Regression model, the following attributes have been identified as key factors that contribute to the high attrition rate of employees and could help the management to take appropriate actions to reduce attrition rate

Age	- Younger employees have high risk of leaving
TrainingTimesLastYear	- More trainings, then employee tends to stay
Years Since Last Promotion	- More frequent promotions, then employee tends to stay
Years with Current manager	- The higher the number of years with same manager, then employee tends to stay
Environment Satisfaction	- The higher, the lesser chances of leaving
Job Satisfaction	- The higher, the lesser chances of leaving
Work life balance	- The higher, the lesser chances of leaving
Average working hours	- The higher the employee works above 8 hours, the higher chances are of him leaving
Marital status single	- Single Employees have higher chances of leaving the company.