HR Analytics– Case Study

Group Members:

1. Amitava Maity
2. Arpan Kumar Nandi
3. Ajay Kumar
4. Rasika Apte

In this case we are required to model probability of attrition using 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.

The following questions are among the ones they would like answered

- What proportion of our staff are leaving?
- Where is it occurring?
- Can we predict future Attrition?
- If so, how well can we predict?

The analysis needs to be done for answering below questions :

- Identify trends in leavers' behaviour and the reasons employees change jobs/organisations
- Identify the employers' perspective on employees' reasons for leaving
- Identify retention strategies that have a positive influence on retention.

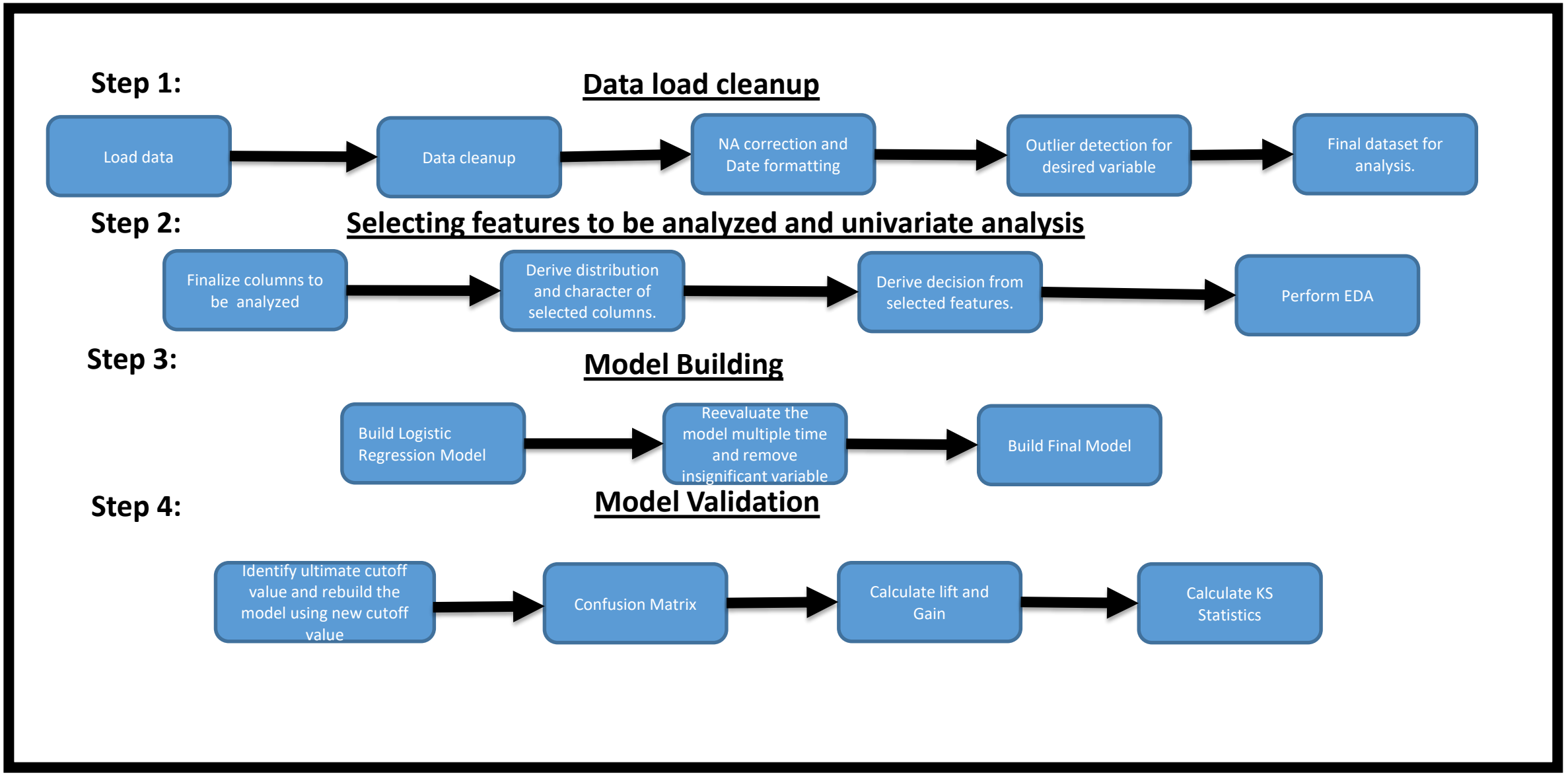
Collect and Manage the Data

After load the datasets, we have found that we have following information for analysis. We have good dataset for 2015, the data consist of -

1. Employee daily In Out data
2. Employee general data, performance data and survey data

Attributes	Attributes	Attributes
EmployeeID	Age	Attrition
BusinessTravel	Department	DistanceFromHome
Education	EducationField	EmployeeCount
Gender	JobLevel	JobRole
MaritalStatus	MonthlyIncome	NumCompaniesWorked
Over18	PercentSalaryHike	StandardHours
StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear
YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
EnvironmentSatisfaction	JobSatisfaction	WorkLifeBalance
JobInvolvement	PerformanceRating	MeanHoursInOffice

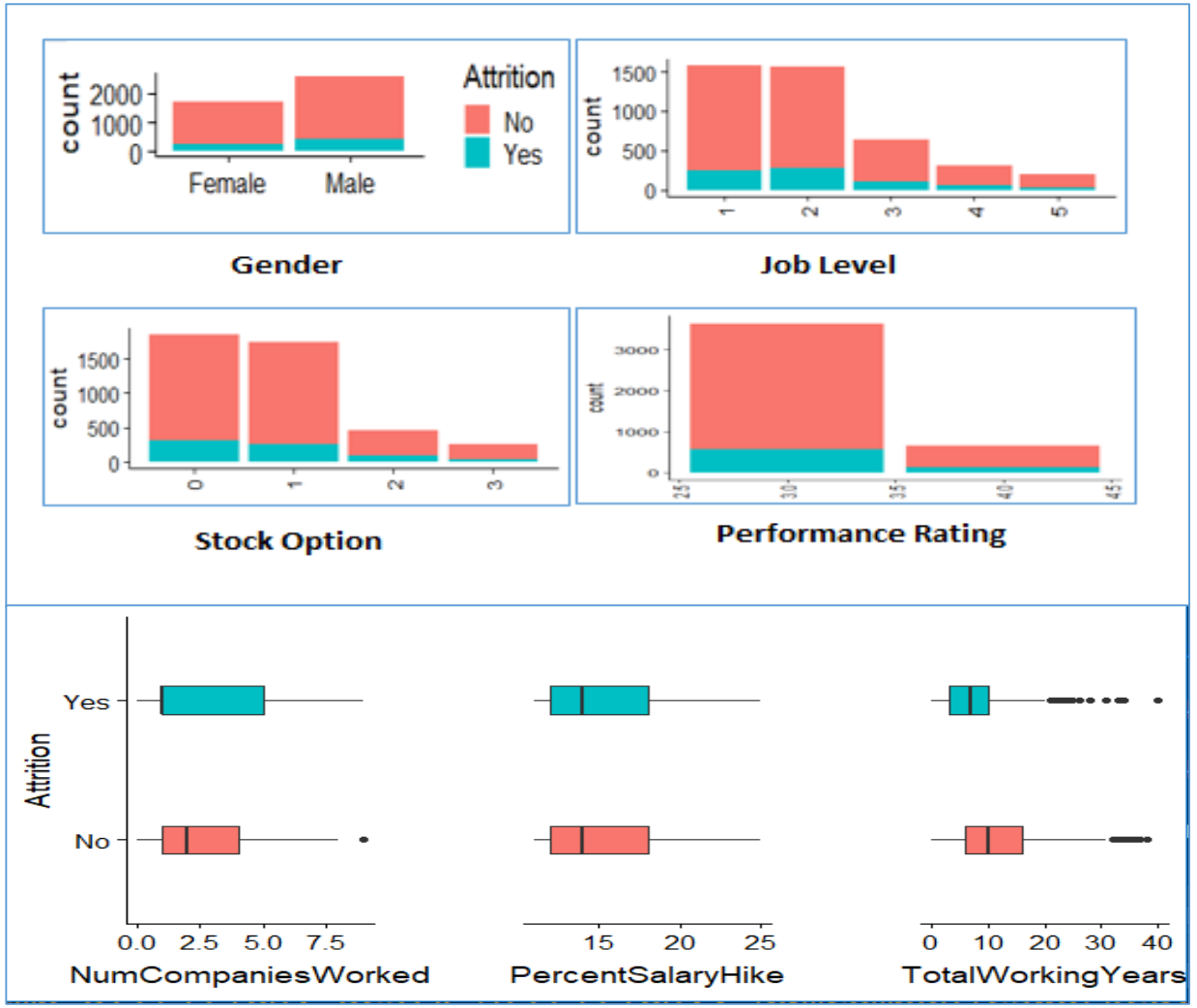
Problem solving methodology



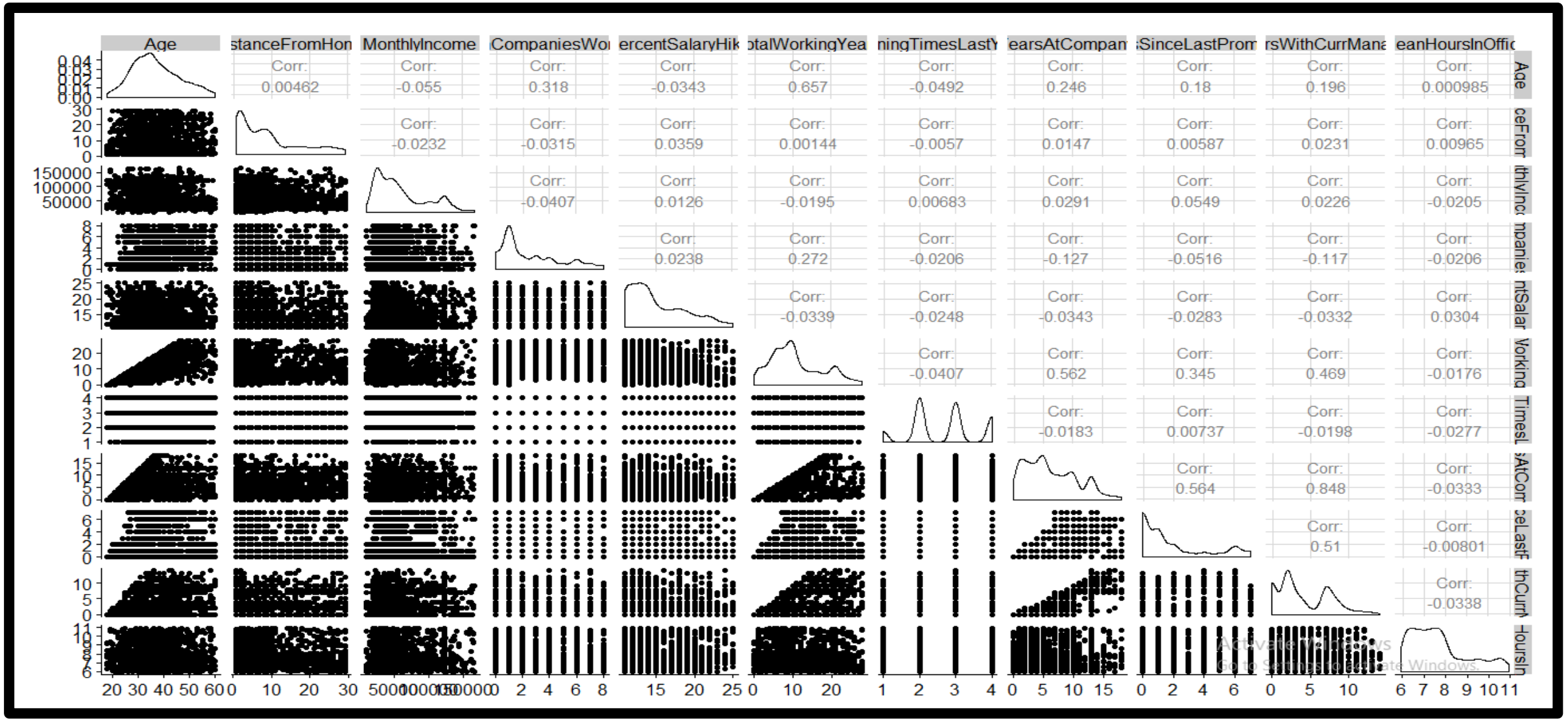
EDA

Summary of EDA

- Male and Single employees are more prone to Attrition
- Employee with job level2 and who are having no stock options are more prone to attrition
- Employee who got comparatively lower rating are more prone to attrition
- There are some outlier in Monthly Income, NumCompaniesWorked, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsSinceLastPromotion, YearsWithCurrManager, MeanHoursInOffice fields, those need to be taken care.
- Employees who have worked in more companies and who have less experience are more prone to attrition
- Employees who spend more years at company are less likely to quit.



Correlation Matrix



Data Cleaning: Model Data Preparation

- Below variables seem to be have strong correlation –
    1. Age & TotalWorkingYears
    2. YearsAtCompany & TotalWorkingYears
    3. YearsAtCompany & YearsWithCurrManager
    4. YearsSinceLastPromotion & YearsWithCurrManager
- Scale continues variable
- Create dummy variable from categorical variable
- Calculate percentage of attrition - 16.16%
- Split full data set into training and test data

Model Building: Logistic Regression

- Logistic Regression :

Initial model Model_1: Summary of the first model is given below

Null deviance	2661.4
Residual deviance	2012.4
AIC	2124.4
Coeff	0.64001

Model_2: Use stepAIC function and create 2nd model

Insignificant variable: JobLevel.x2 is insignificant

Null deviance	2661.4
Residual deviance	2028.4
AIC	2098
Coeff	0.75905

Model Building: Logistic Regression

After having 28 iteration we came up with our final model –

Model_28: Summary of the model is given below

Null deviance	2661.4
Residual deviance	2039.9
AIC	2257.9
Coeff	-2.0935

Fields	Std. Error	Z Value	P value	Star
NumCompaniesWorked	0.05548	6.478	2.52E-07	***
TotalWorkingYears	0.07665	-11.057	< 2E-16	***
YearsSinceLastPromotion	0.06001	5.32	9.31E-11	***
MeanHoursInOffice	0.05108	12.344	< 2E-16	***
BusinessTravel.xTravel_Frequently	0.12767	5.054	1.04E-07	***
MaritalStatus.xSingle	0.11043	7.865	2E-16	***
EnvironmentSatisfaction.x4	0.12246	-3.613	4.32E-07	***
JobSatisfaction.x4	0.12684	-5.589	3.70E-15	***

Model Evaluation

- ✓ General Confusion Matrix: Use probability cutoff of 50%
  - Sensitivity : 0.17225
  - Specificity : 0.97965
- ✓ General Confusion Matrix: Use probability cutoff of 40%
  - Sensitivity : 0.32057
  - Specificity : 0.96392

Sensitivity has been increased when we decrease the cutoff, here we are keen to calculate eventual sensitivity value because this is our objective of this case study.

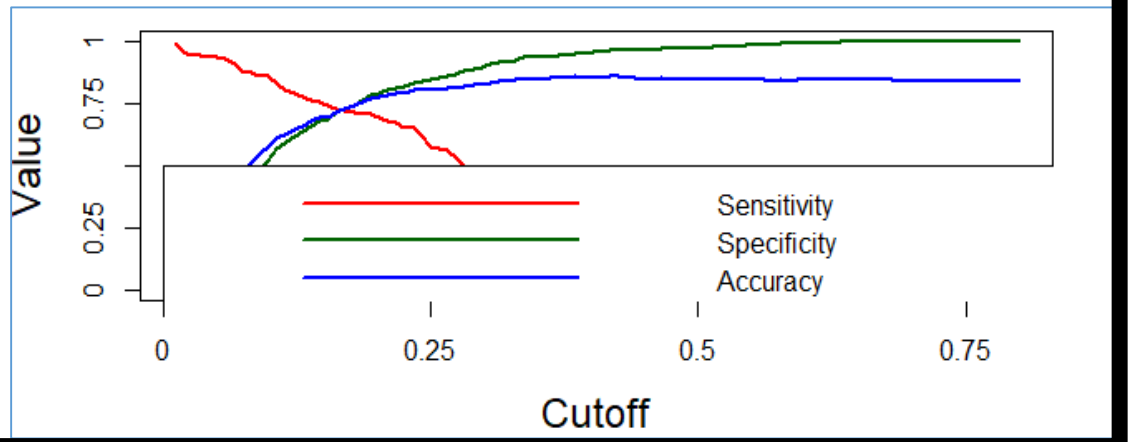
- ✓ Draw plot for sensitivity, specificity and accuracy to identify the cutoff value for this model
  - Cuttoff = 0.1696

Actual	Predicted	
	No	Yes
No	1059	22
Yes	173	36

For Cutoff 50%

Actual	Predicted	
	No	Yes
No	1042	39
Yes	142	67

For Cutoff 40%



Model Validation

- Build final model using cutoff value of 0.1696

Accuracy	**0.7295**
Sensitivity	0.7368
Specificity	0.7280

bucket	total	totalresp	Cumresp	Gain	Cumlift
1	129	73	73	34.9	3.49
2	129	39	112	53.6	2.68
3	129	35	147	70.3	2.34
4	129	18	165	78.9	1.97
5	129	9	174	83.3	1.67
6	129	10	184	88	1.47
7	129	4	188	90	1.29
8	129	9	197	94.3	1.18
9	129	2	199	95.2	1.06
10	129	10	209	100	1
		Lift and Gain			

- Lift and Gain
- In lift and gain chart 5th Decile is having cumulative gain 83.3% which indicates that it is good model.
- KS Statistics of this model is 46.49% that means it's a good model

Summary

Throughout the analysis, I have learned several important things

- Features such as Year Since Last Promotion, Total Working Year, Mean Hours In Office, Job Satisfaction, Environmental Satisfaction, Years with Current Manager appear to play a role in employee attrition.
- There seems to be a relation between marital status and attrition
- There is a tension between probability threshold and the number of employees who are accurately predicted as potential churners. A high probability threshold would end in a high number of errors. The business relevance is predict attrition well, rather than non attrition hence a lower probability threshold is chosen.
- Model is biased towards finding the employee who are about to leave the company
- The confusion matrix shows that of all the people who are going to leave the company, our algorithm identifies about 72% of them accurately.