HR ANALYTICS CASE STUDY SUBMISSION

Group Name:

1. **Prasoon Dayal**
2. **Nitin Srivastava**
3. **Satya Mishra**

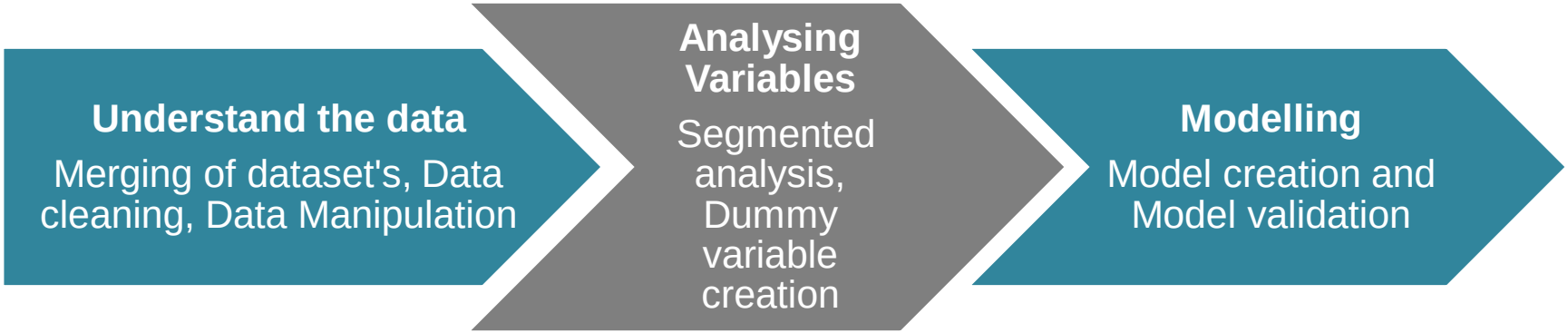
XYZ Company Requirement Summary

1. **XYZ**, employs, at any given point of time, around 4000 employees, but every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. **XYZ** wants to cut down on the attrition rate.

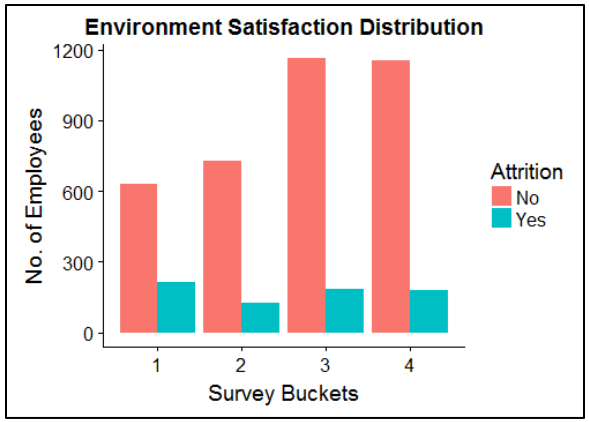
XYZ's Current thoughts on why attrition is bad:

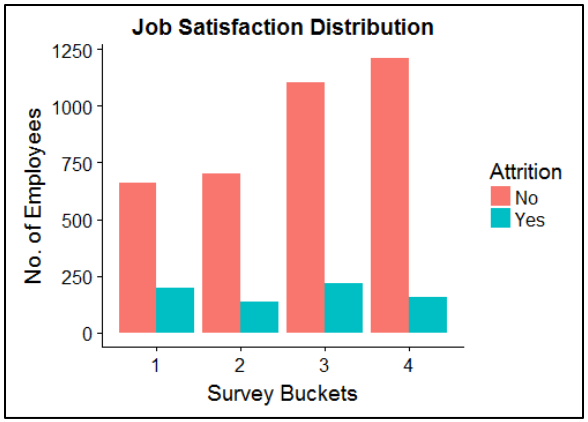
1. Project delays leading to missing timelines hence loss of goodwill in front of consumers and partners.
2. Having to maintain a sizeable HR team for recruitment all the time.
3. New employees need to be trained to be job ready.

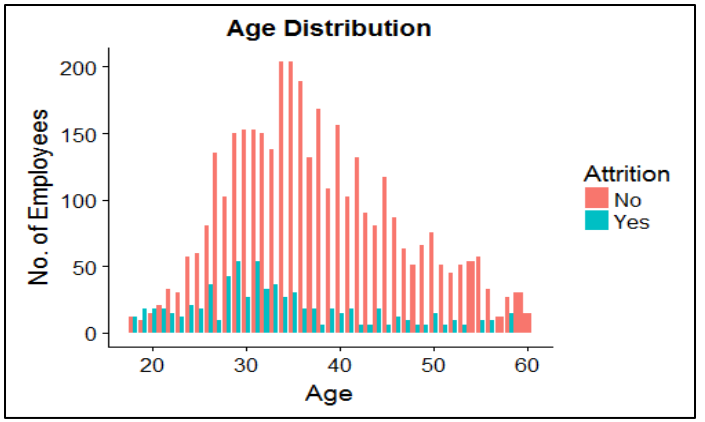
Approach

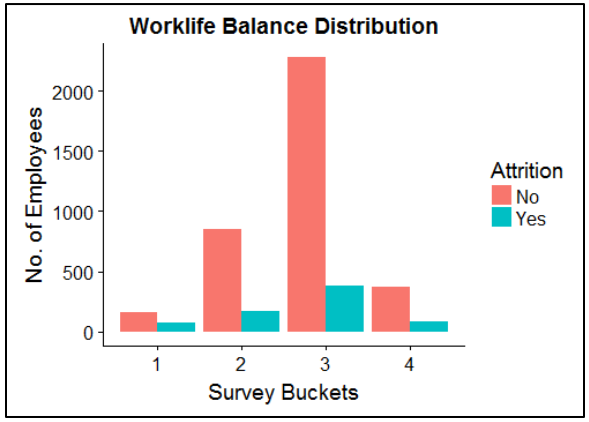


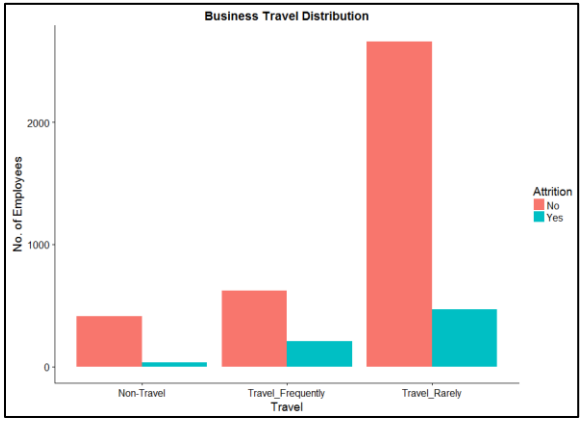
Univariate Analysis

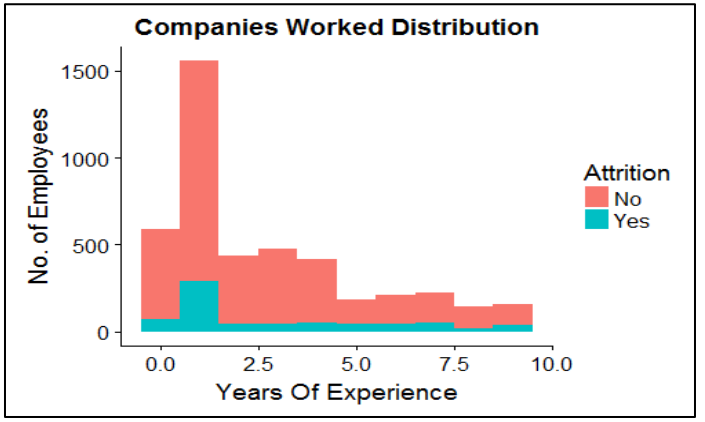


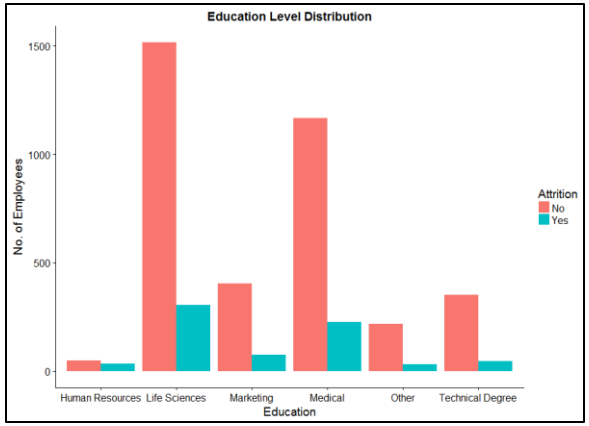


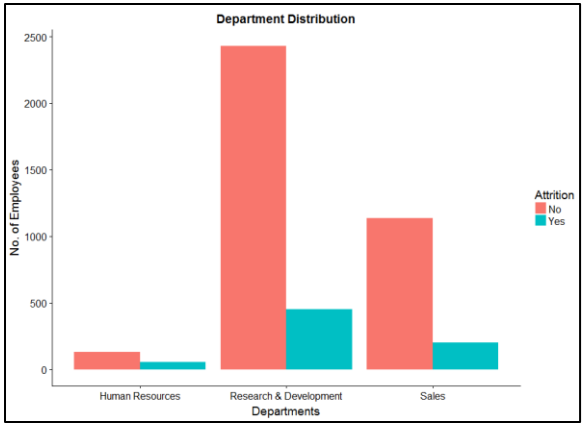


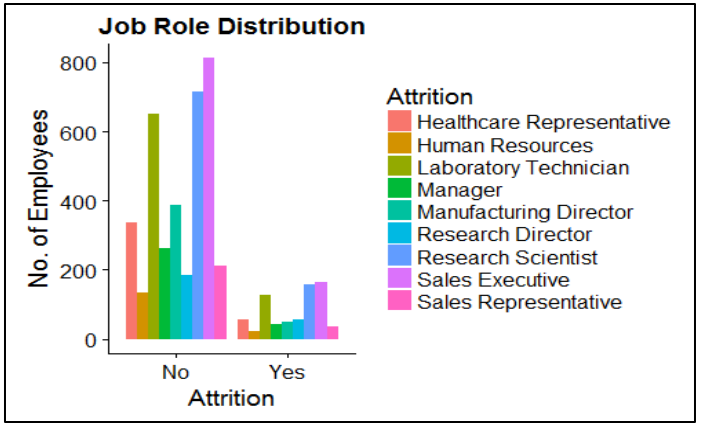




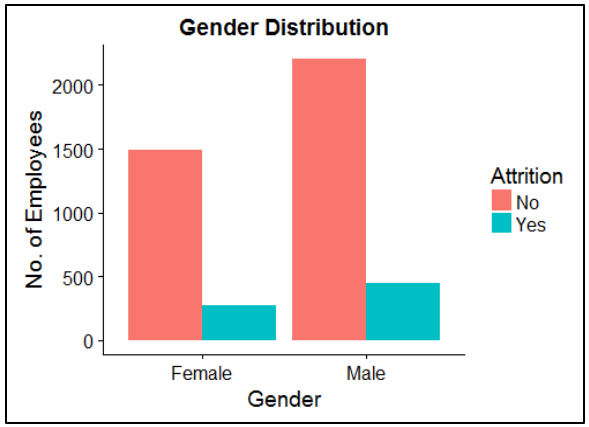


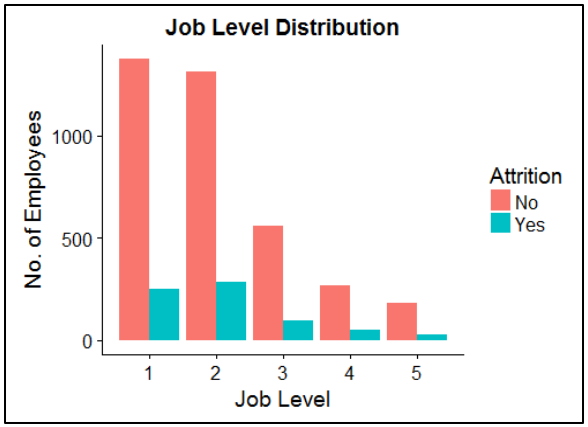


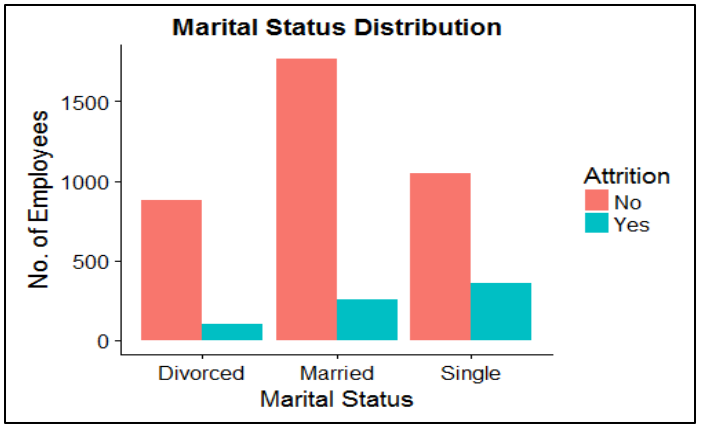




Univariate Analysis (continued)

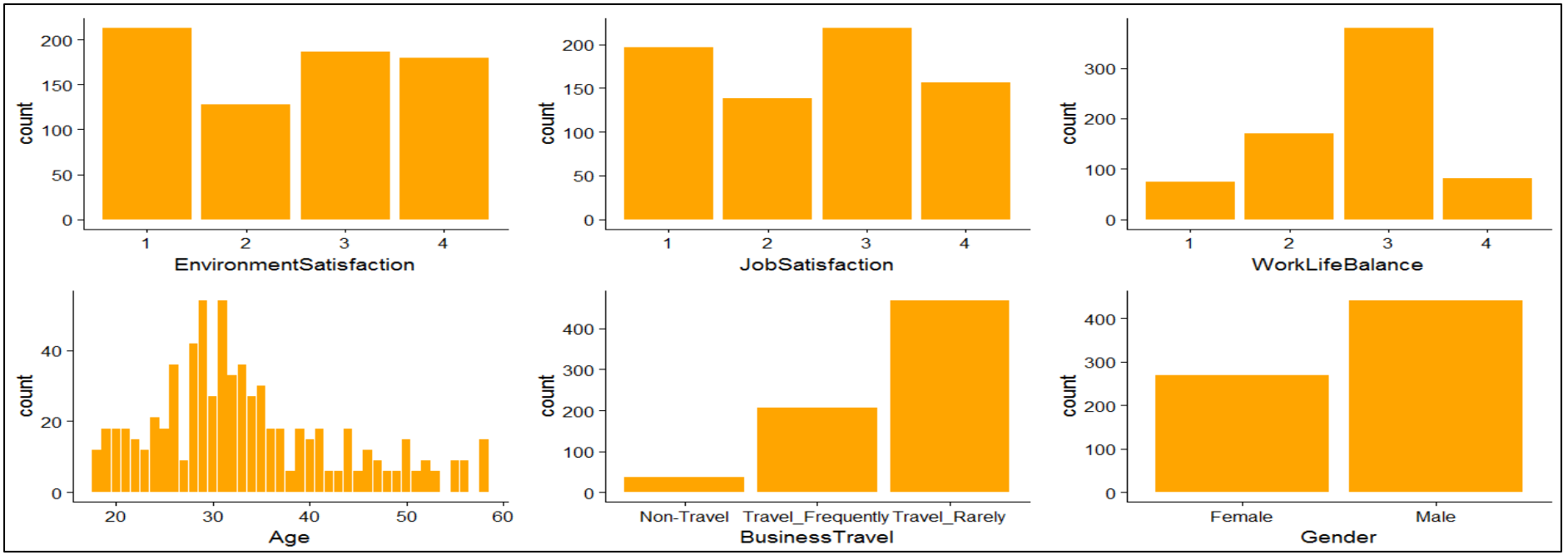




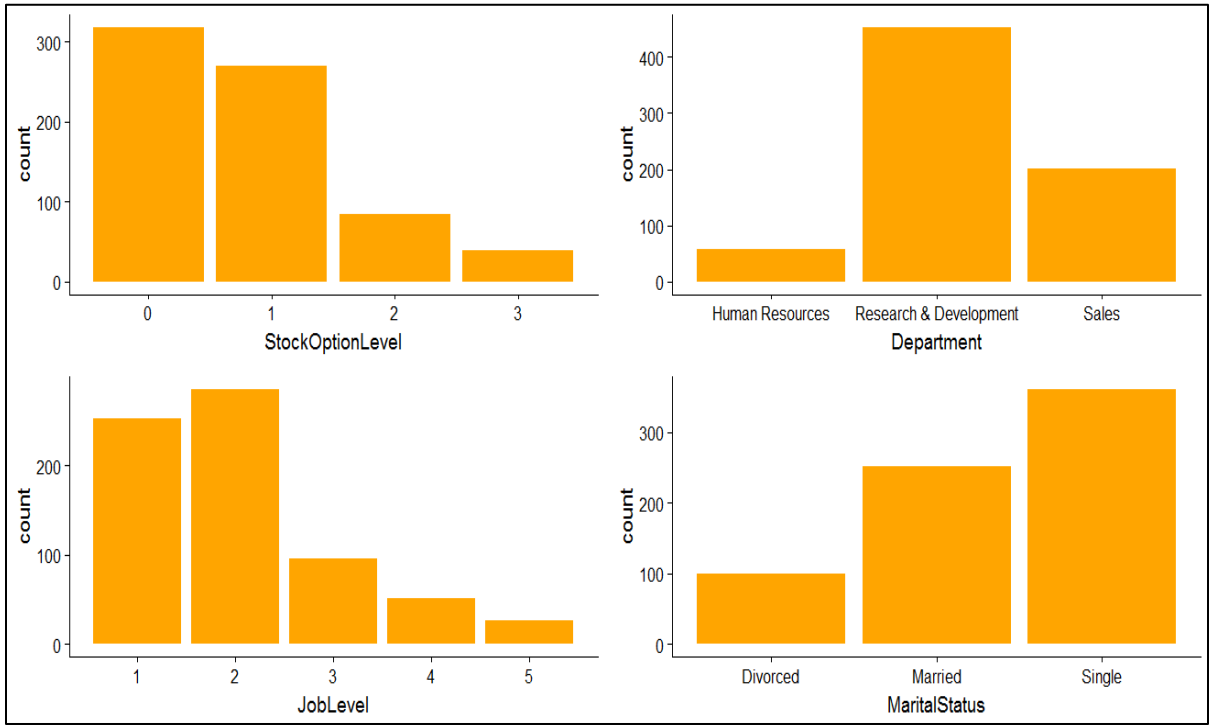


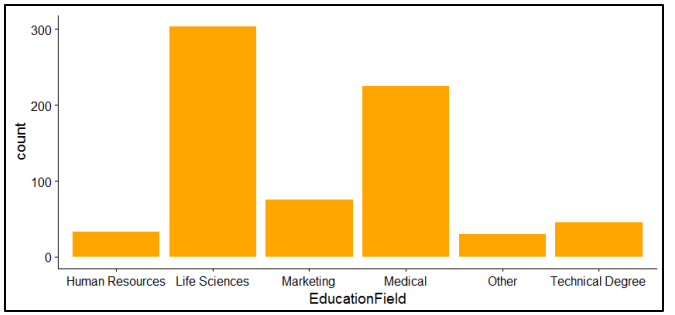
While doing Univariate analysis of attrition over other variables we saw employees with marital status as “SINGLE” and the employees who gave a low rating in Environment Satisfaction have higher attrition rate.

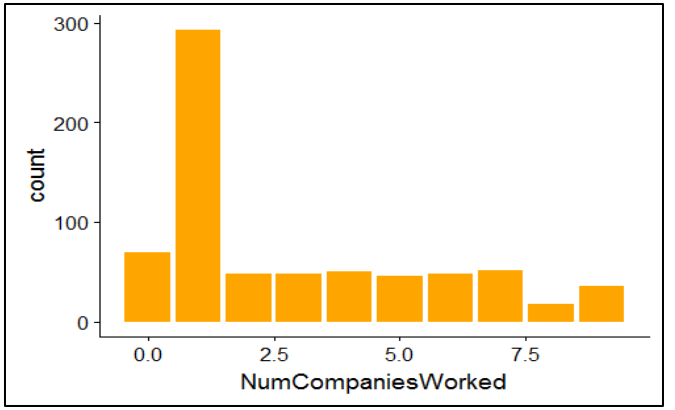
Segmented Analysis (Where Attrition is “YES”)

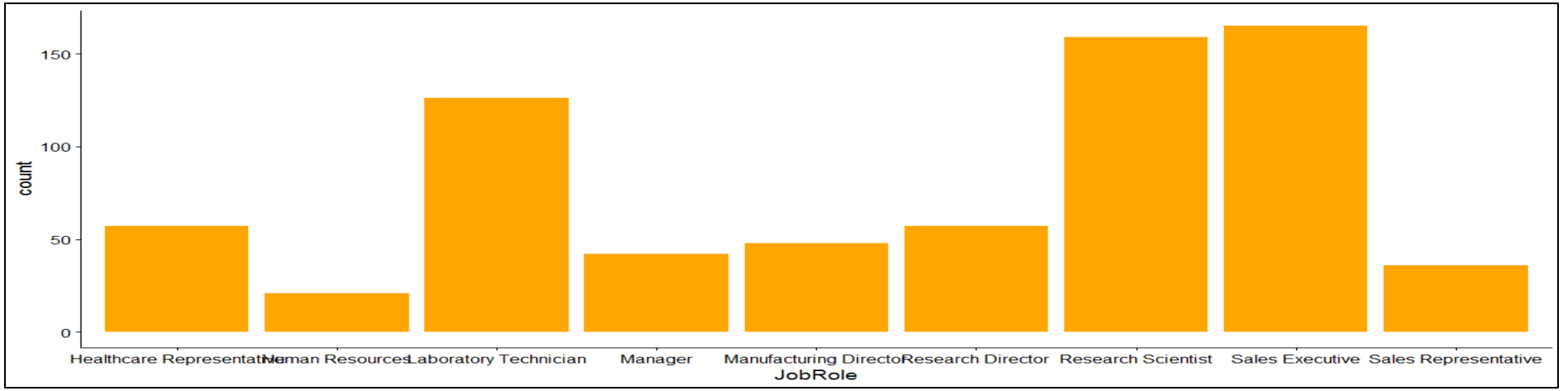


Segmented Analysis (Where Attrition is “YES”)







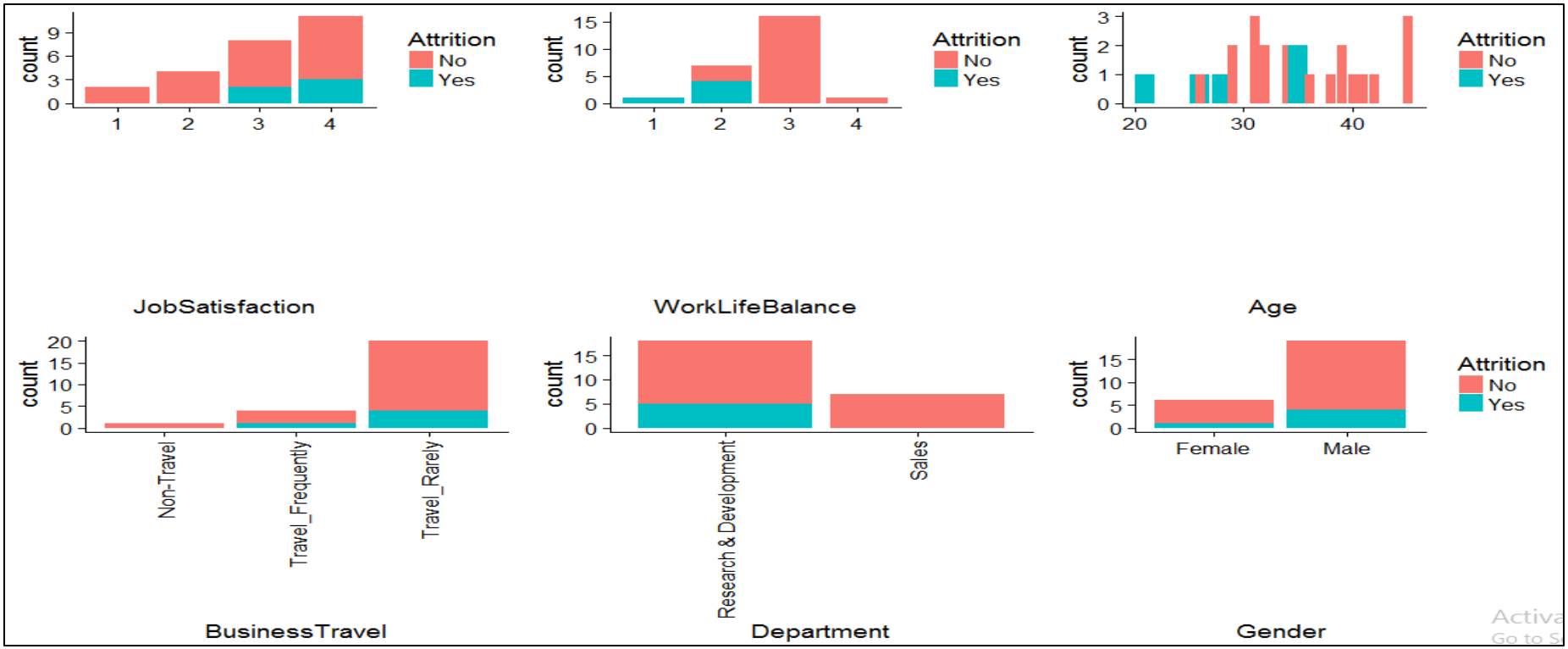


While doing segmented analysis of attrition (where attrition is YES) over other variables we do not see any of the variable having a major impact. The attrition across the variable’s tells the same story .

Summary Of Univariate Analysis

1.We looked into the different variable that compromise of both categorical and continuous types, we saw employees with marital status as “SINGLE” and the employees who gave a low rating in Environment Satisfaction have higher attrition rate.
2.There are no such variable that single handedly influence the attrition rate.
3.We understand that in XYZ company most of the employees work as laboratory technician and sales executive.
4.The age variable has normally distributed curve with most of the attrition happening in the age group of 25 to 35.
5.Most of the employees work in the research and development department and come from life science background.
6.The trends that follow for overall data is also same for segmented analysis on attrition as YES.

Distribution of NA values for different variables on Attrition



Distribution of NA values for different variables on Attrition

We see that Environment Satisfaction has 25 NA's. In the same way JobSatisfaction , WorkLifeBalance, NumCompaniesWorked & TotalWorkingYears have 20, 38, 19, 9 NA's respectively (graphs included in R Code file).

In order to avoid bias, we selected WOE analysis for EnvironmentSatisfaction, JobSatisfaction , WorkLifeBalance, NumCompaniesWorked and replaced all values with respective WOE values including NA. For TotalWorkingYears , observation with NA values were removed.

Binning and Treatment of outlier values

A. For Monthly Income replaced salary greater than 137756.0 with 137756.0.
B. Binning of continuous variables shown in below table helped to treat them as categorical variables.
C. Bins were creates as below:

TotalWorkingYears	0-2	3-4	5-7	8-9	10-12	13-16	17-22	23+
YearsAtCompany	0-1	3-4	5-8	9-14	15+			
YearsSinceLastPromotion	1-3	4+						
YearsWithCurrManager	1-2	3	4-8	9+				
PercentSalaryHike	11	12	13-14	15-18	19-20	21+		
DistanceFromHome	1-2	3-10	11+					
Age	18-25	26-33	34-37	38+				

Dummy variable creation

After WOE analysis, dummy variables were created for all categorical and continuous variables. These included Education, Environment Satisfaction, Job Satisfaction, Work Life Balance, Business Travel, Department, Education Field, Gender, Job Level, Job Role, Marital Status, Number Companies Worked, Stock Option Level, Training Times Last Year, Job Involvement, Performance Rating, Percent Salary Hike, Total Working Years, Years At Company, Years Since Last Promotion, Years With Current Manager, Distance From Home, Age & Attrition.

Note – EmployeeId and Attrition was skipped during this analysis.
Attrition values were converted to numeric 1 and 0, with 1 resembling' Yes'

Scaling variables

The only variable that was scaled is Monthly Income

Model creation activity

A. We used logistic regression to predict the classification of attrition as YES or NO.
B. The final data set that was used for model building activity had 82 variables.
C. 70% of the dataset was used to train the classification model and the rest of 30% was kept for testing phase
D. Step AIC function was used to eliminate non significant variables at one go.
E. A total of 30 models were created to arrive at the final model to be used.
F. The removal of variables was based on High VIF and insignificance value of (p>0.05)

Inference

The final model showed below factors affecting attrition of an employee:

EnvironmentSatisfaction 'Low'
JobSatisfaction 'Low'
WorkLifeBalance 'Low'
BusinessTravel Frequently
JobLevel at scale 2
JobRole Manufacturing Director
Jobrole Research Director
Marital Status Single

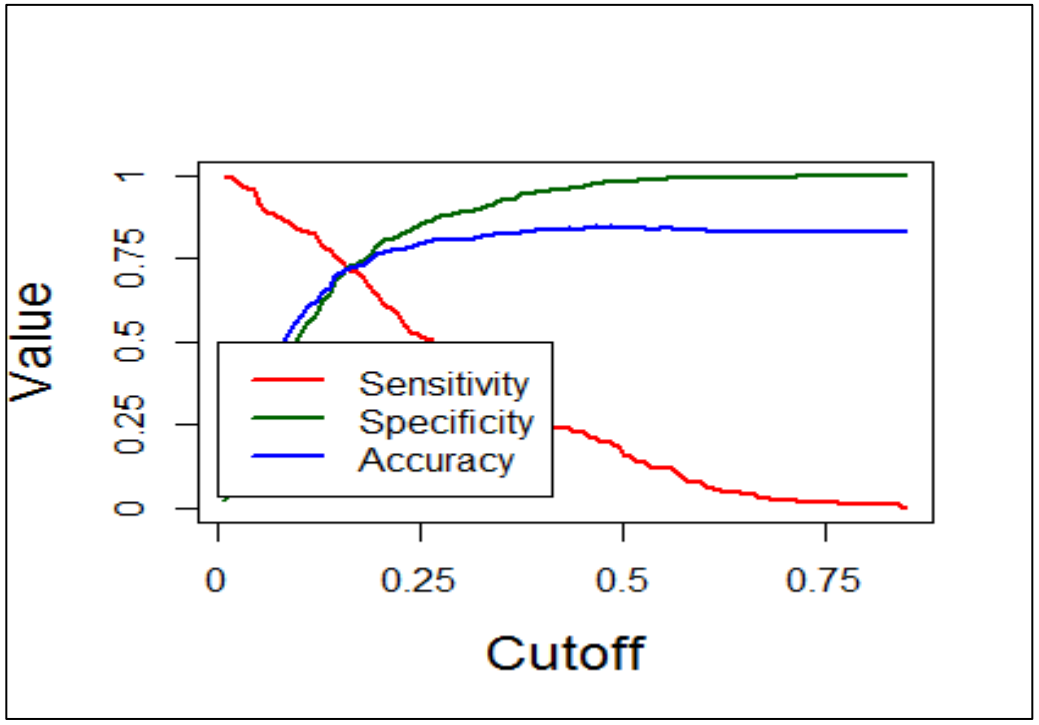
Inference contd.

Those who have not worked in any company before
No. of trainings conducted for employee last year - 6
'High' level of involvement in job
Years spent by employee at company in range of 5-8
Years spent by employee with current manager is in range of 1-2
Age is 38 or above
Actual Working Hours is more than 8.5 hours.

Model validation activity

A. We have used the predict function to get the predictions.
B. We have mentioned the type as response as that will give us the probability values.
C. The probability obtained has a range from 0.5% to 90%
D. We choose the probability threshold as 0.50 to do the prediction and saw the accuracy to be around 85%, sensitivity to be around 19% and specificity to be around 97%.
E. As per our requirement we want the sensitivity to give us a much higher level that 19%

Getting the optimum probability threshold



A. An user defined function was created to plot a line graph with the three parameters namely Accuracy, Sensitivity & Specificity.

B. The optimal probability threshold came out to 0.18580 for the best prediction

C. The final score that was achieved for all the three parameters are: -

i. Accuracy – 74.03
ii. Specificity - 73.66%
iii. Sensitivity – 74.10%

Additional Evaluation Metrics

We considered below statistics for our model evaluation in addition to sensitivity, specificity and accuracy.

A. KS-Statistic
B. Gain – Lift Analysis and Chart
C. ROC Curve

Our analysis presented below KS-stats table. (refer R-code)

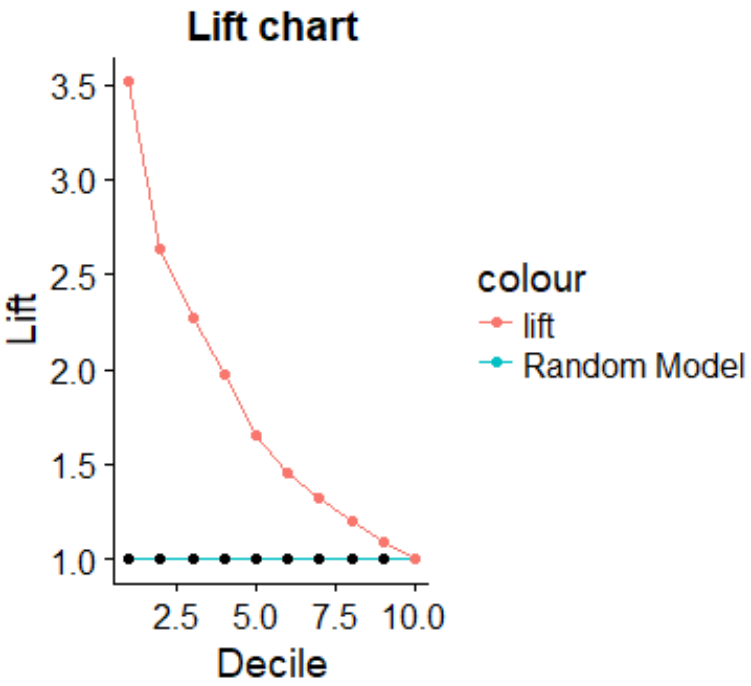
Decile	AttritionCount	CumulativeAttrition	GainPercent	NonAttritionCount	CumulativeNonAttrition	GainPercentNonAttrition	KS_stat	GainRandomModel	lift
1	72	72	35.12195122	60.1	60.1	5.385304659	29.73664656	10	3.512195122
2	36	108	52.68292683	96.1	156.2	13.99641577	38.68651106	20	2.634146341
3	32	140	68.29268293	100.1	256.3	22.96594982	45.32673311	30	2.276422764
4	22	162	79.02439024	110.1	366.4	32.83154122	46.19284903	40	1.975609756
5	7	169	82.43902439	125.1	491.5	44.04121864	38.39780575	50	1.648780488
6	10	179	87.31707317	122.1	613.6	54.98207885	32.33499432	60	1.455284553
7	10	189	92.19512195	122.1	735.7	65.92293907	26.27218288	70	1.317073171
8	8	197	96.09756098	124.1	859.8	77.04301075	19.05455022	80	1.201219512
9	5	202	98.53658537	127.1	986.9	88.43189964	10.10468572	90	1.094850949
10	3	205	100	129.1	1116	100	0	100	1

Inference

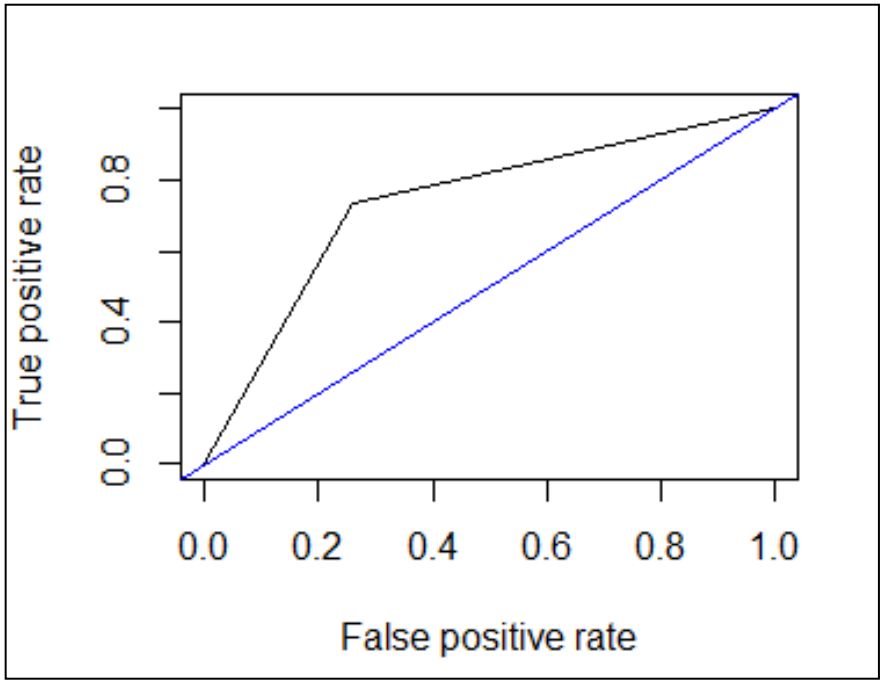
KS-Stat table showed that KS-statistic was 46.19 in 4th decile.

This implies that our model can help us in identifying 79% of employees who will leave, by targeting 40% of workforce.

Lift Chart



ROC Curve



Inference

Lift Chart and ROC Curve shows that our model is able to distinguish between employees who will leave and those who will not.