HR ANALYTICS CASE STUDY

SUBMISSION

Group Name:

1. Ashish Chandan
2. Kasinath Kalava
3. Ramashankar Nayak
4. Ravi Bhavsar

Business Objective

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.

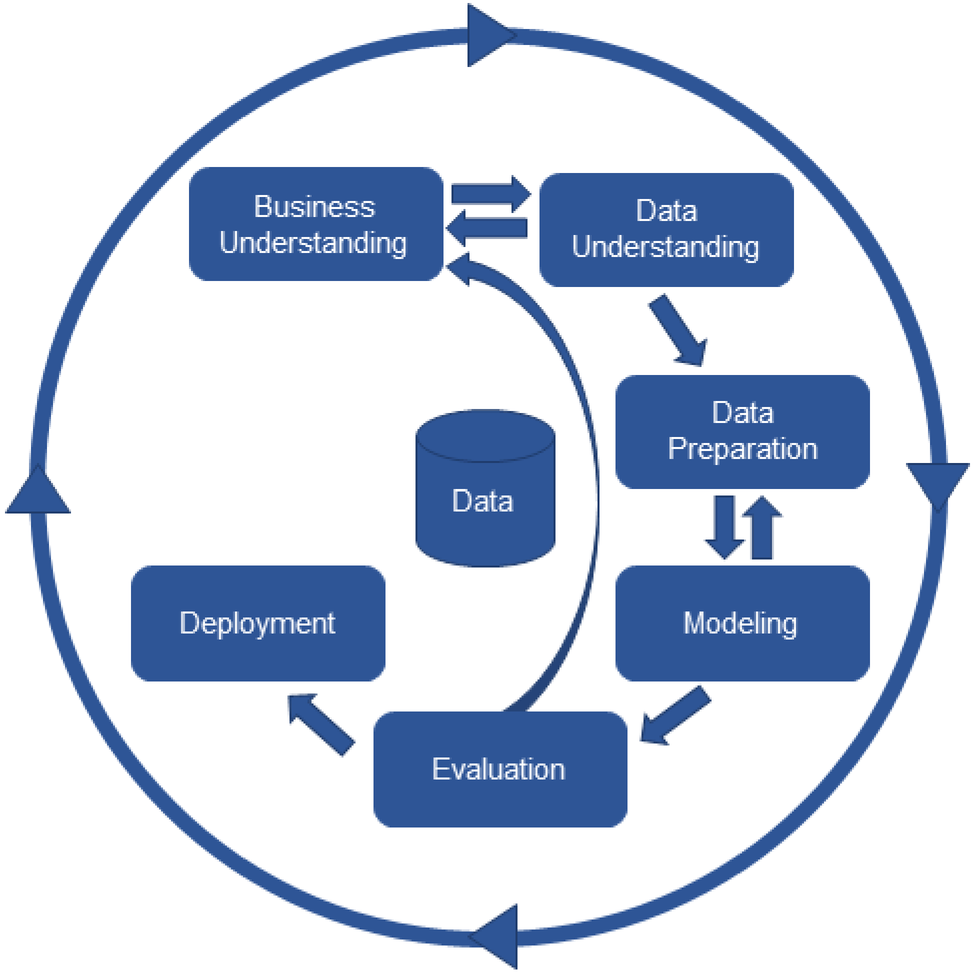
Business Constraints:

- Only 1-year data of the employees has been provided

Objective of analysis is to find out:

- Factors to focus on for curbing attrition rate from growing
- What changes are to be made in workplace environment to minimise attrition
- Priorities among factors attributing high attrition to attend them immediately.

Problem Solving Methodology



- **Data Understanding**: Data has been provided in 5 categories – General data of employee, Employee survey data, Manager survey data, In and out time details for year 2015.
- **Data Preparation**: It includes handling missing values, outliers, duplicates, creation of dummy variable for categorical variables, deriving new metrics (wherever applicable) and formatting data before proceeding with modeling.
- **EDA** – To understand the trends within data against target variable (Attrition).
- **Modeling**: It involves analyzing the nature of predicted variable. As attrition is binomial in nature, so we use glm function available in R to build logistic regression model.
- **Evaluation**: Steps involve to validate the model for the test dataset using Confusion Matrix(determining accuracy, specificity and sensitivity), also comparing outputs from the model with a random model to check the performance and summarizing the results keeping the business success constraints in mind.

ANALYSIS : DATA UNDERSTANDING

1. Datasets provided for analysis

   The Manager Survey Data – Collected from a company wide survey.

   The Employee Survey Data – Collected from a company wide survey.

   In Time Data – Collected from company's attendance Log sheet/ Machine.

   Out Time Data – Collected from company's attendance Log sheet/ Machine.

   General Data – General data includes employees personal data along with education.

2. Attrition from general dataset is the target variable.
3. EnvironmentSatisfaction, JobInvolvement, JobSatisfaction are defined like below

   1 means 'Low', 2 means 'Medium', 3 means 'High‘ and 4 means 'Very High‘

4. WorkLifeBalance is defined like below

   1 means 'Bad', 2 means 'Good', 3 means 'Better' and 4 means 'Best'

5. Education is defined like below

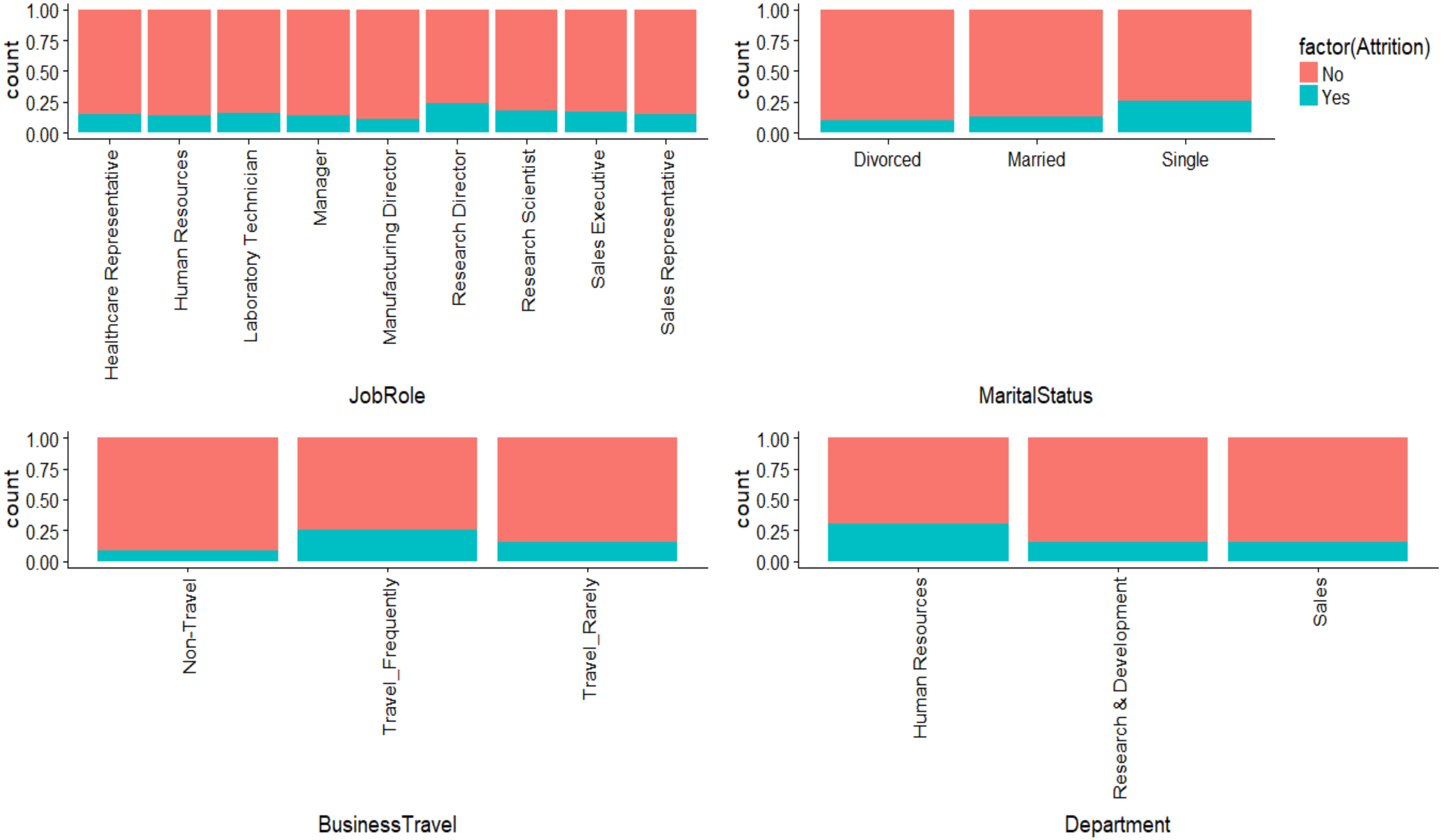
   1 means 'Below College', 2 means 'College', 3 means 'Bachelor‘ , 4 means 'Master' and 5 means 'Doctor'

Analysis : Data Cleaning and Preparation

- All the data files have been merged to form a core data file for analysis, Excluding In time and Out time.
- NAs, Duplicates and Outliers are validated and treated where ever required. Mean and Median are considered to handle NAs. Outliers were treated for TotalWorkingYears, YearsAtCompany and YearsWithCurrManager. However we have skipped treating for Monthlyincome as there is a sudden jump after 90 to 91 percentile in the data (considering 10% of the data is above the sudden jump).
- For Attrition, Gender, Over18 - as these are having 2 levels these being realigned as numerical Yes == 1 and No == 0.
- Average working hours are calculated based on in time and out time datasets for each employee as part of derived metrics, Also segregated the Average working hours as greater than 8 hours(Overt time) and less than 8 hours(Inadequate working hours).
- Total number of leaves are calculated considering NA from in time and out time dataset for each employee (12 Public holidays have been excluded).
- Dummy variables for following categorical predictors having more than 2 levels are created

  EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance, JobRole, MaritalStatus, BusinessTravel, Department,Education, EducationField, JobInvolvement, JobLevel, PerformanceRating
- Final dataset has been achieved after this preparation exercises and Relevant predictors have been scaled to aid in regression modelling.

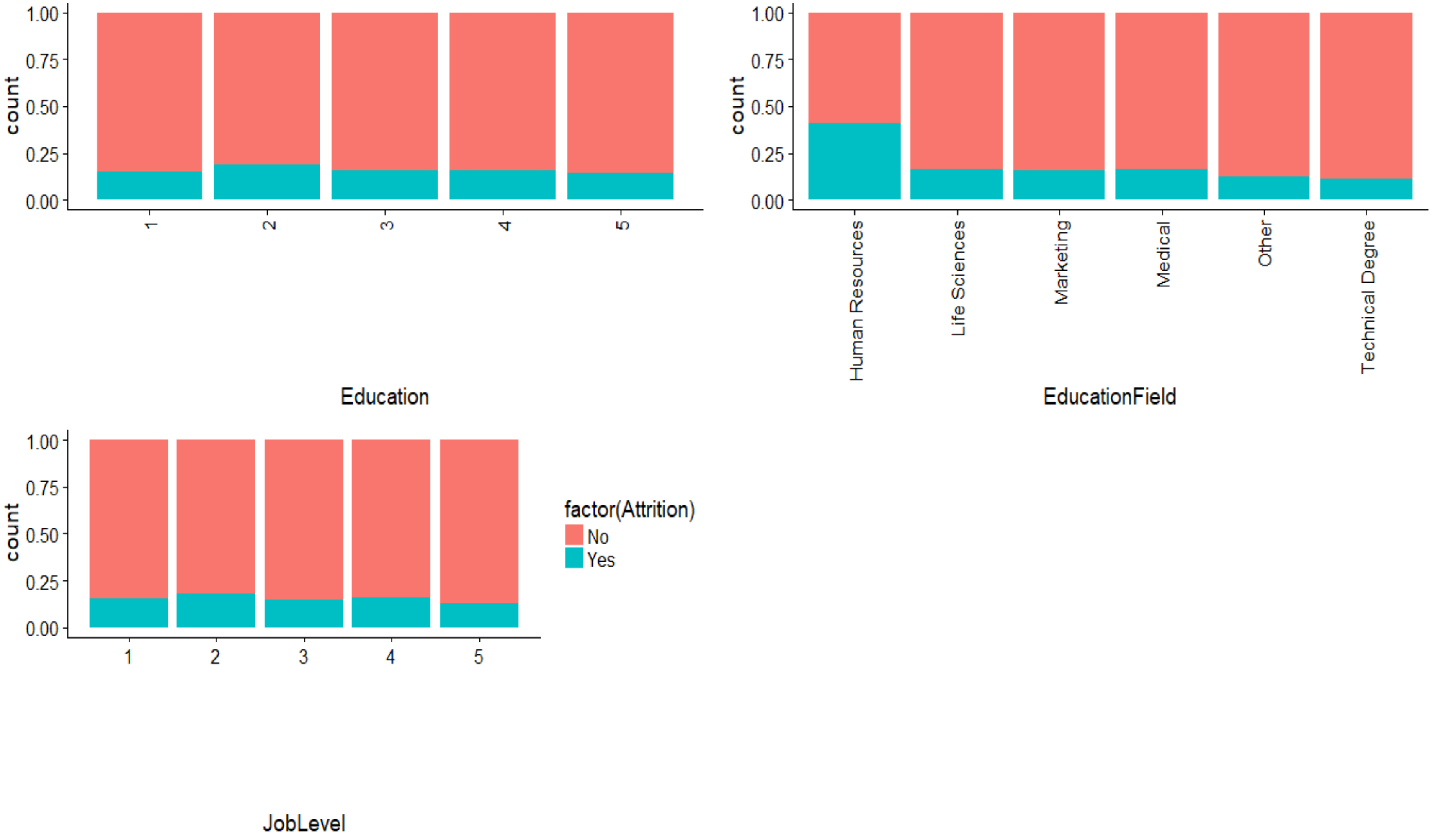
EDA : Exploring categorical variables



1. Job roles with Research Director has the higher attrition
2. Marital status - Single, are leaving the company more compared to married and divorced
3. Employees who had Business travel frequently seems to quit more compared to others
4. Surprising, HR department got more attrition

Bar charts for Categorical variables Job role, Marital status, Business travel and Department against Attrition

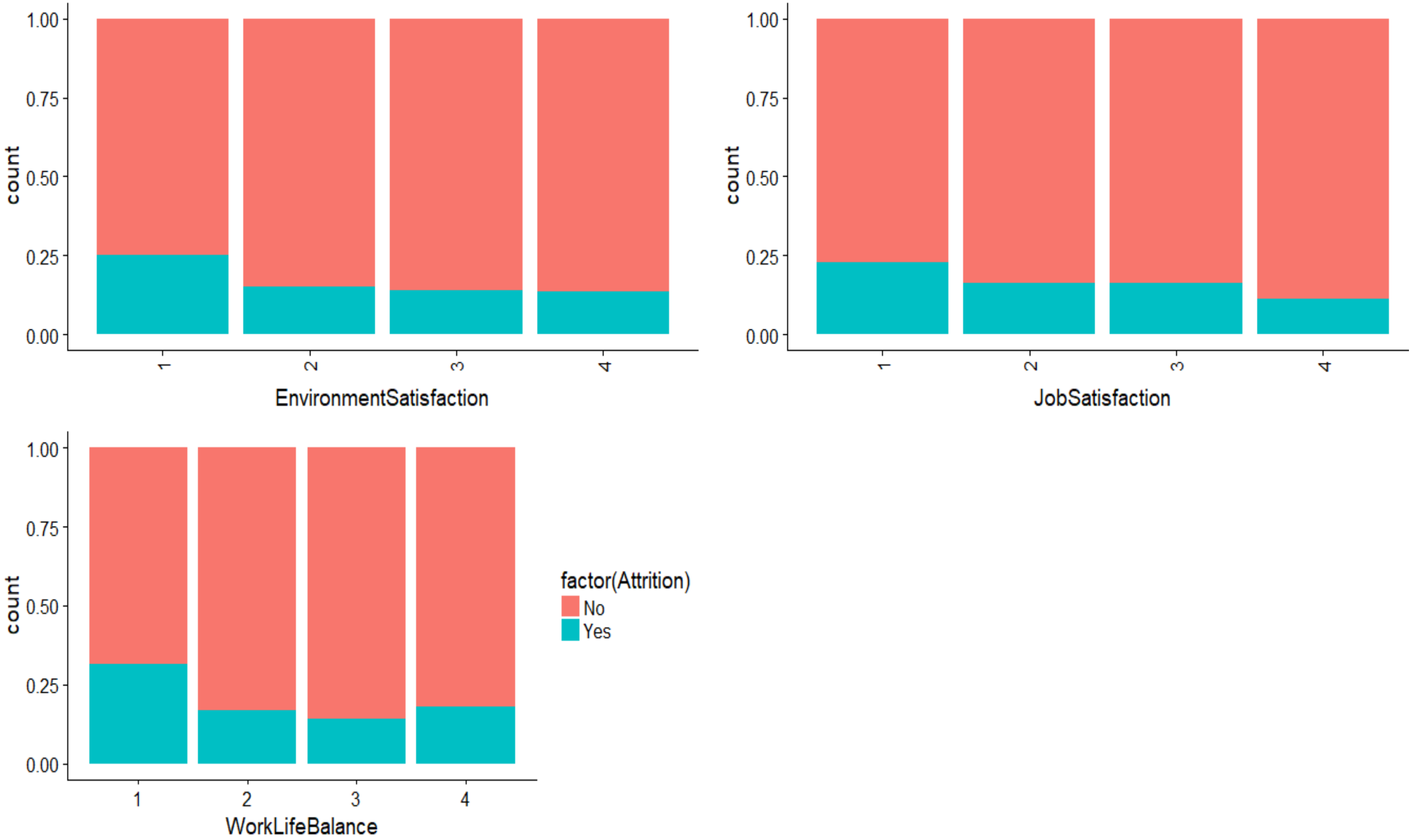
EDA : Exploring categorical variables



1. Education 2 got higher attrition
2. Max percentage of person having Education Filed as Human Resource has left the organisation
3. Job level 2 has got higher attrition

Bar charts for Categorical variables Education, Education field and Job level against Attrition

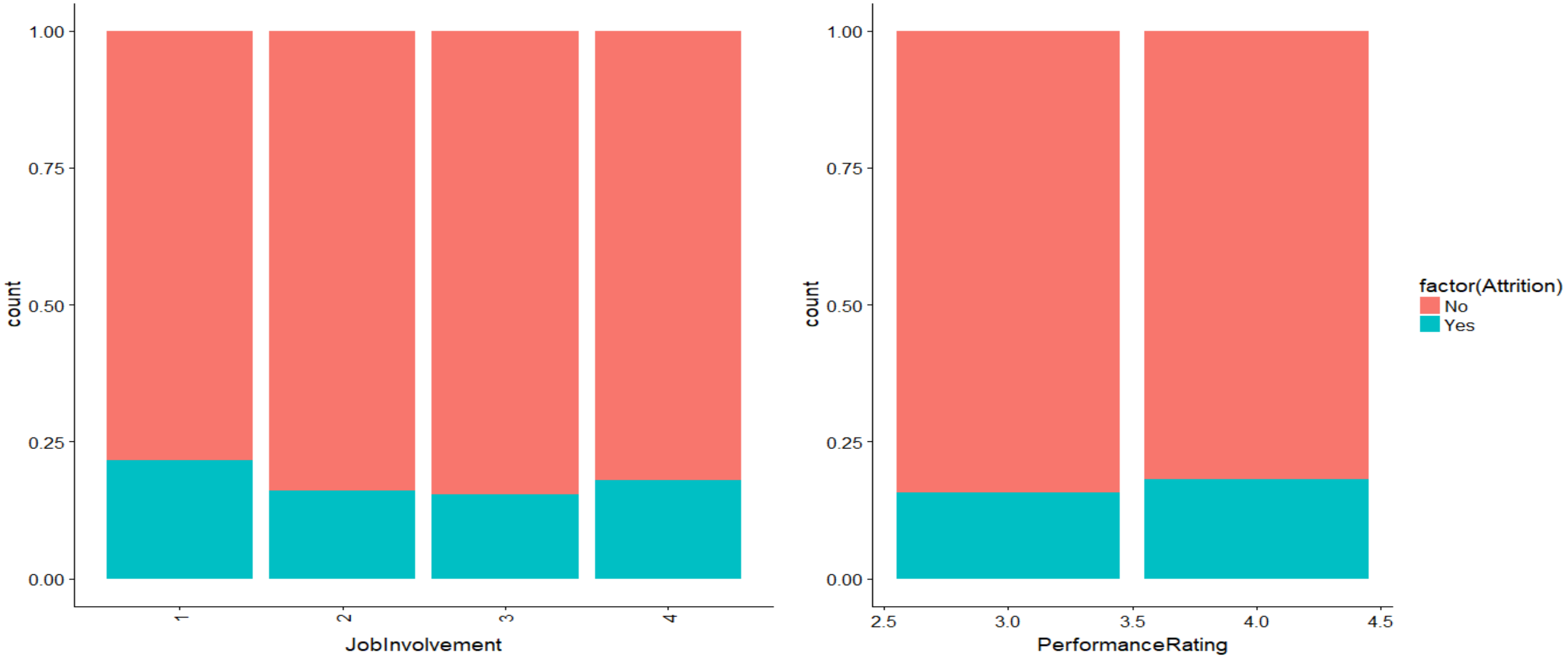
EDA : Exploring categorical variables



Environment satisfaction, Job satisfaction level and Work Life Balance attrition decrease with increase in corresponding index.

Bar charts for Categorical variables Environment Satisfaction, Job Satisfaction and WorkLifeBalance against Attrition

EDA : Exploring categorical variables

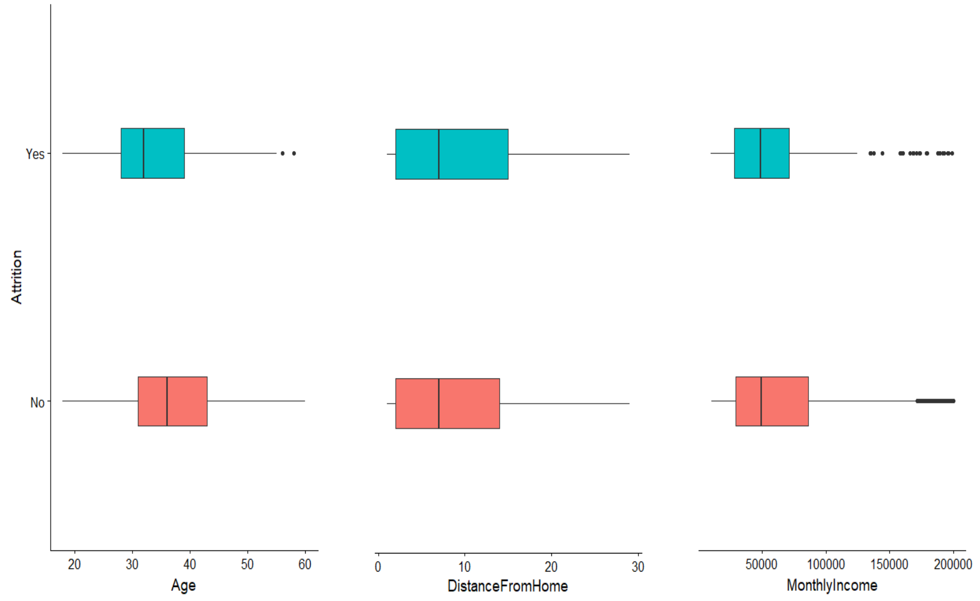


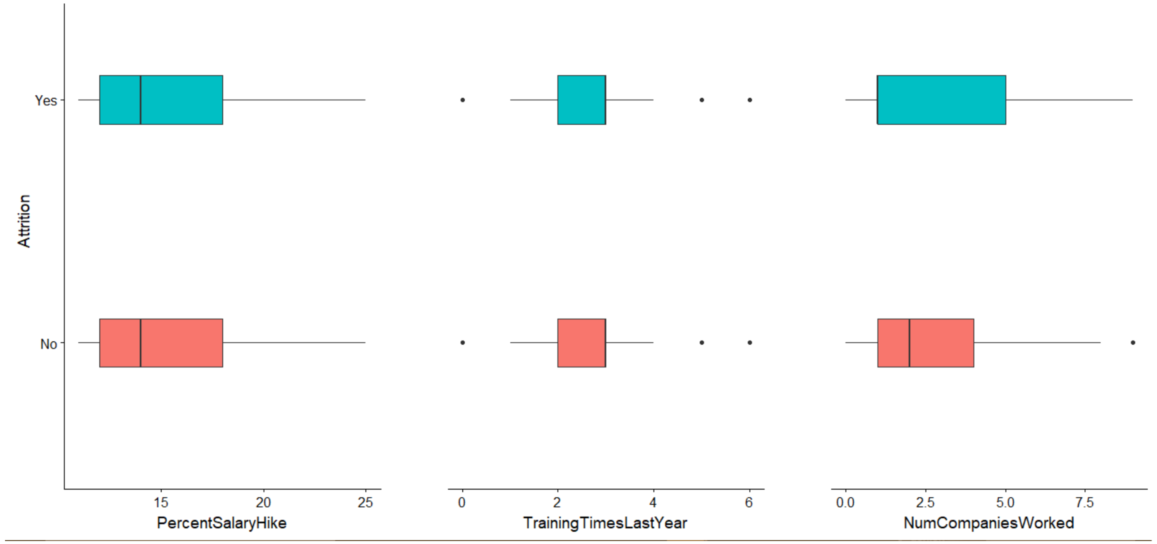
1. Lesser the job involvement, more the attrition
2. Don't look much variation in attrition due to band performance

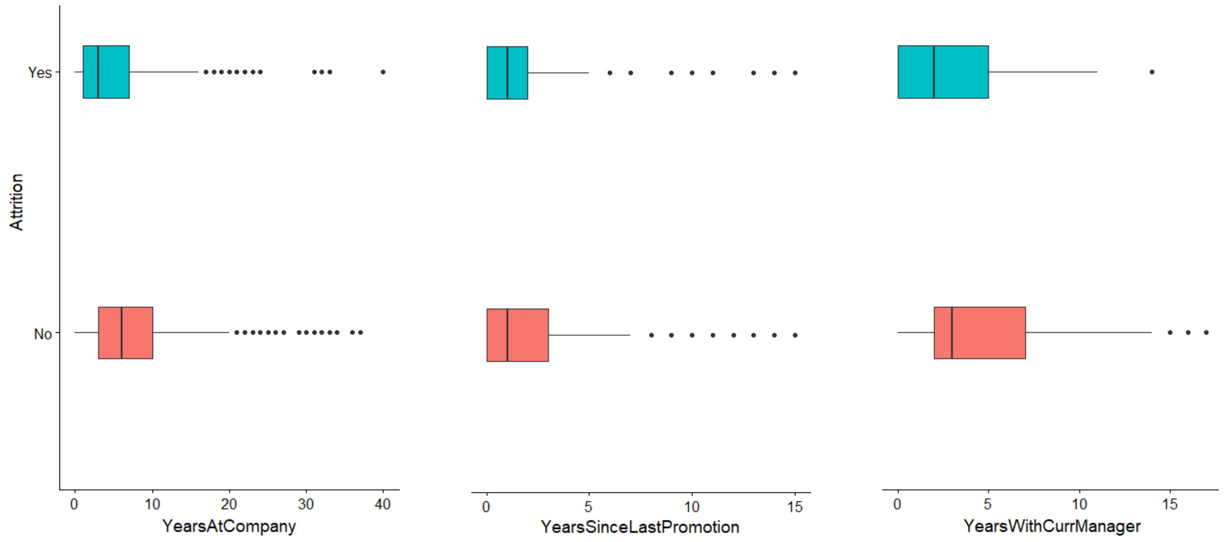
Bar charts for Categorical variables Job Involvement, Performance Rating against Attrition

***Standard Hours, Over18 and Employee Count have only one category (No Variation), hence we have excluded these variables from Analysis

EDA : Exploring Numeric variables

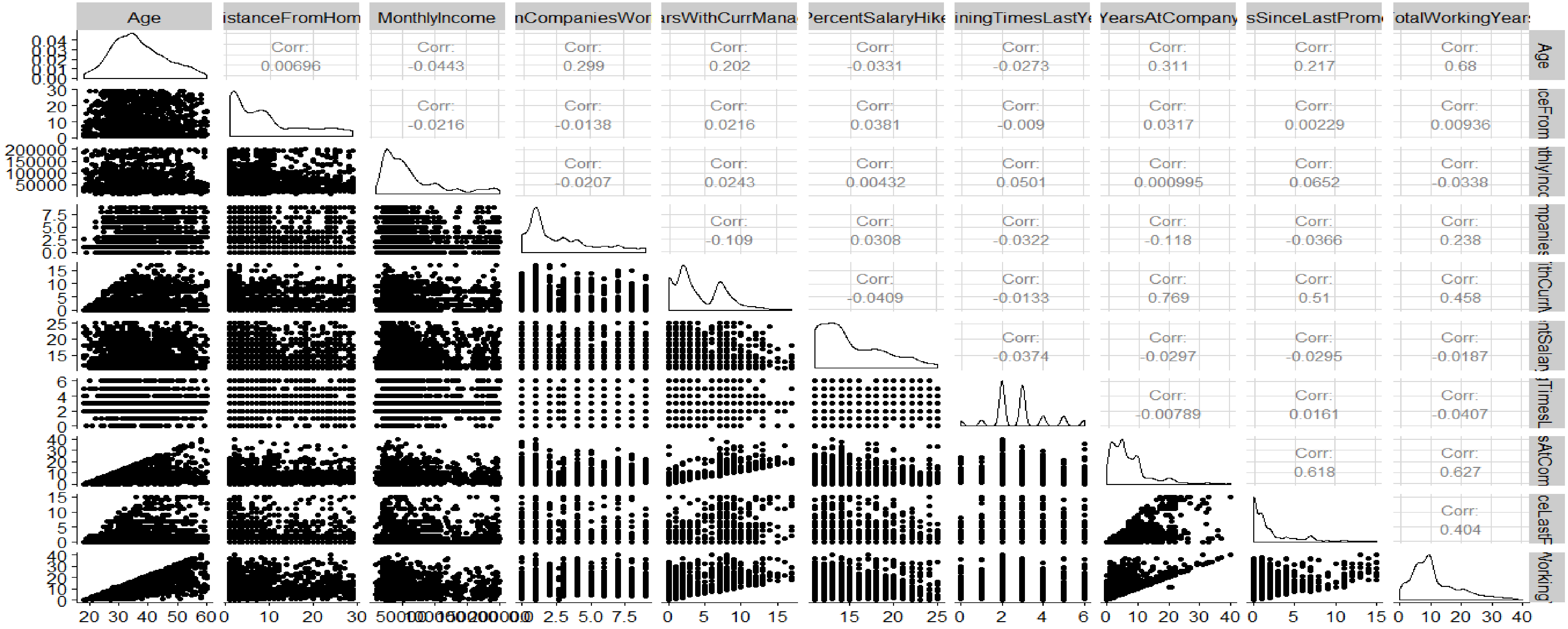






1. Box plots of numeric variables are published here against attrition data
2. Variation of Number of companies worked when the attrition is Yes is high as compared to No
3. Variation of Monthly income and Years since last promotion is quite low when there is attrition
4. If the average years at company / Years with current manager is low, there is fair chance of attrition

EDA : Correlation between Numeric variables



1. Years at company and years with current manager are highly correlated (corr. 0.77) followed by
2. Total working years and years at company (corr. 0.63), then followed by
3. Years since last promotion and Years in company (corr. 0.62)

Model Building using Logistic Regression

- Our Response variable is “Attrition” (1 == Yes, & 0 == No)
- Rest all non constant numeric variables are scaled to aid in regression modelling.
- Splitting data into training and test data set.
- For creating Train and test datasets from final data set:
  - We fixed seed to 100 and Used split ratio of 0.7 for training dataset and remaining data has been assigned to test dataset
- Initial model has been conceived with glm function, then StepAIC has been applied to arrive at standard model which yielded on iterative predictor selection without major reduction in AIC Score.
- Removed the variables having high VIF value and low significance i.e. if p-value > 0.05
- Checked the correlation among variables appropriately and removed from the model accordingly.
- Then based on VIF (variance inflation factor) and P - value (with significance) predictors have been filtered and after another 28 iterations we could achieve our final model. with almost all predictors being significant with lowest VIF are present.

Final Logistic Model - Significant Variables

Variable Name	Coefficient value	VIF value	P Value
Age	-0.45799	1.111878	1.18E-14
TrainingTimesLastYear	-0.18525	1.012758	0.000879
YearsSinceLastPromotion	0.37649	1.529929	6.37E-08
YearsWithCurrManager	-0.63028	1.477179	< 2e-16
avg_ofc_hours	1.29066	1.056913	< 2e-16
EnvironmentSatisfaction.x2	-0.92218	1.416789	2.46E-08
EnvironmentSatisfaction.x3	-0.96936	1.530331	5.24E-11
EnvironmentSatisfaction.x4	-1.19211	1.531721	7.02E-15
JobSatisfaction.x2	-0.59763	1.453038	0.000374
JobSatisfaction.x3	-0.50662	1.597555	0.000477
JobSatisfaction.x4	-1.18175	1.556632	9.70E-14
WorkLifeBalance.x3	-0.36410	1.011311	0.000891
MaritalStatus.xSingle	1.04169	1.047938	< 2e-16

After MODEL 28, We have below final variables

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)

In the final model we have 13 significant variables with positive and negative coefficients.

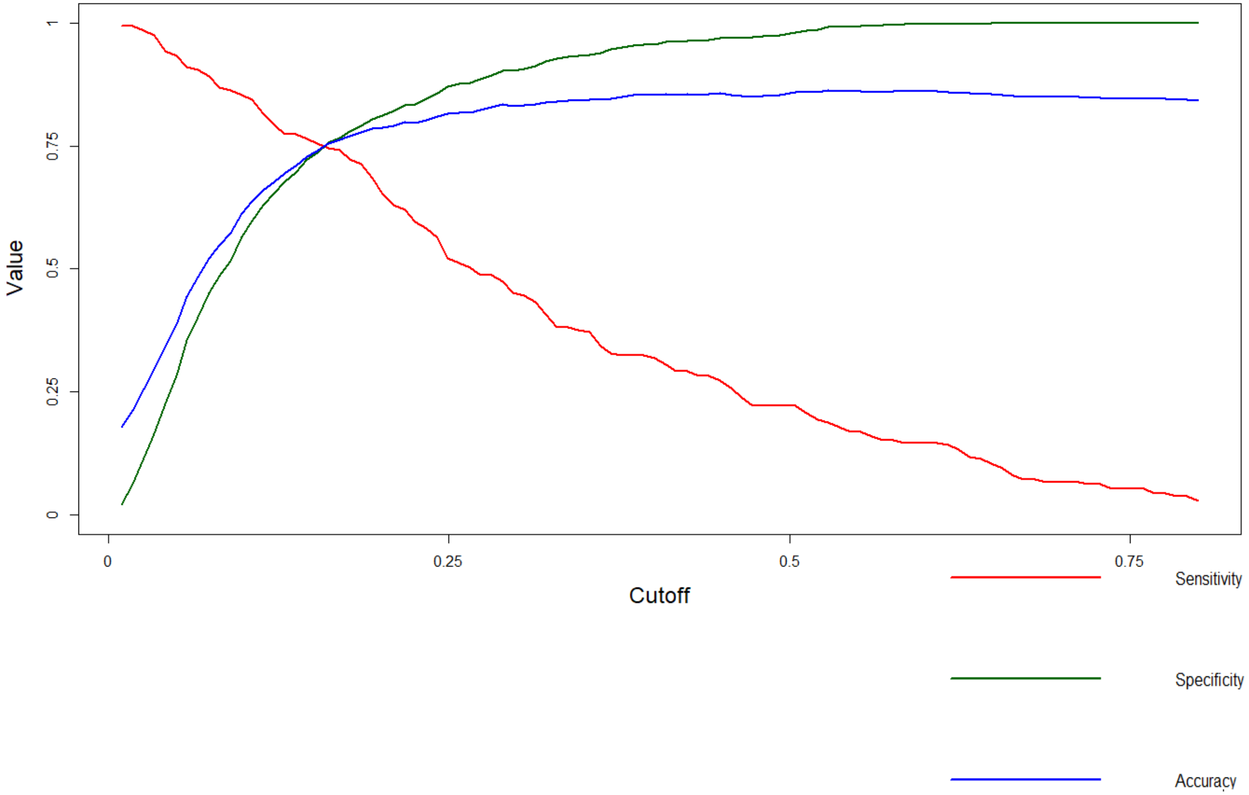
- Positive coefficient means that if positively more the value of the variable more will be the chance of attrition and
- Negative coefficient means they will be affecting attrition rate negatively

Model Evaluation

To evaluate the final model, we execute the model on the test dataset and performed the following model validations:

Finding Accuracy, Specificity and Sensitivity through Confusion Matrix

- In order to find a suitable probability cut-off, we checked the Accuracy, Sensitivity and Specificity from 1% to 80% probability values
- The optimum cut-off probability is the one where the value of specificity and sensitivity are close to each other. Here we have taken a safe range of 0.01. Cut-off probability was found to be ~0.15



Cut-off Probability	Value	Percentage
Accuracy	0.7399849	74%
Sensitivity	0.7558685	75%
Specificity	0.7369369	74%

Model Evaluation contd.

Calculating KS Statistics for test data

KS statistics measures the degree of separation between positive and negative distribution.

The optimal value of KS statistics for a good model should lie between 40-60 and

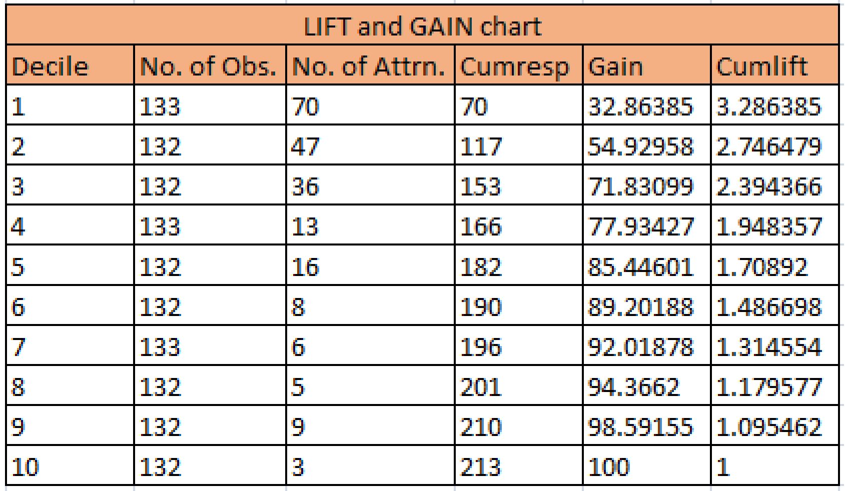
should be within first 3 Deciles. KS-Statistics for our final model is 0.49

Lift and Gain Chart

It helps to measure the effectiveness of the model by calculating the percentage

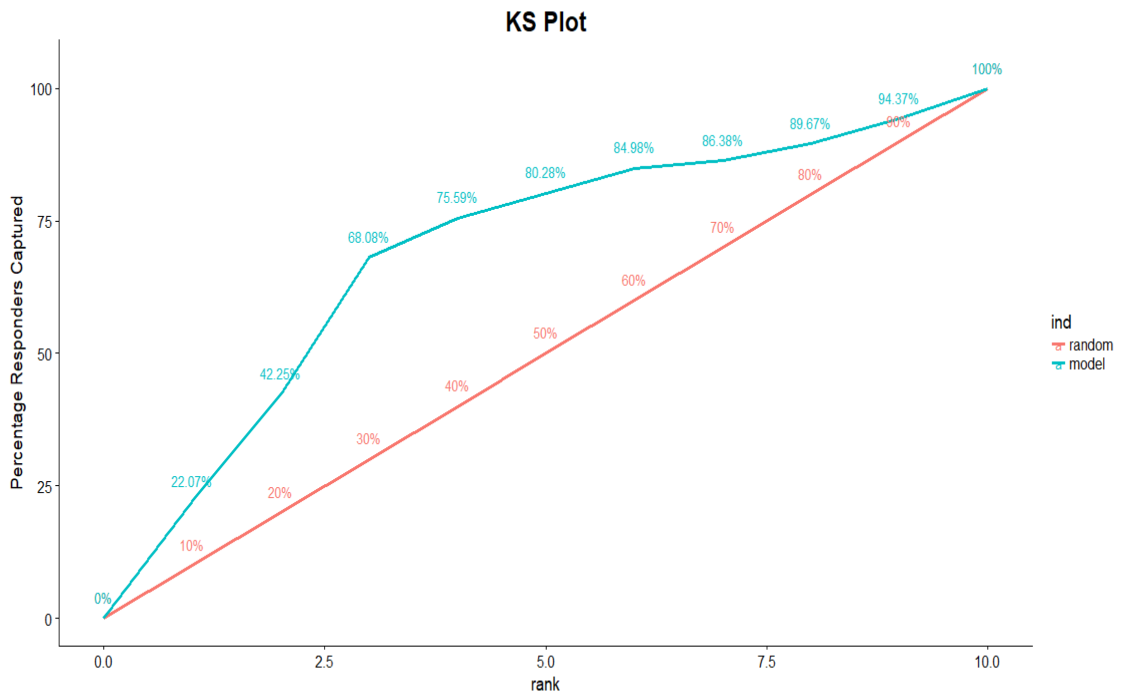
of events captured in each decile.

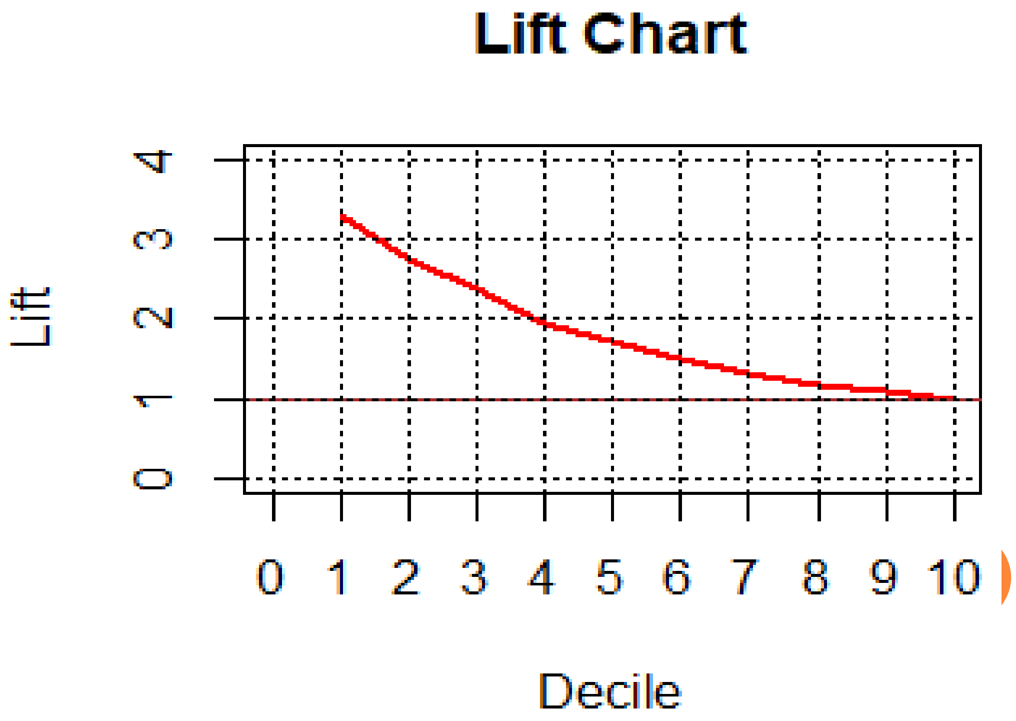
Here, we can see that even by targeting top 30% of the employees, we can predict attrition approx 70% of the time.



LIFT and GAIN chart

Decile	No. of Obs.	No. of Attrn.	Cumresp	Gain	Cumlift
1	133	70	70	32.86385	3.286385
2	132	47	117	54.92958	2.746479
3	132	36	153	71.83099	2.394366
4	133	13	166	77.93427	1.948357
5	132	16	182	85.44601	1.70892
6	132	8	190	89.20188	1.486698
7	133	6	196	92.01878	1.314554
8	132	5	201	94.3662	1.179577
9	132	9	210	98.59155	1.095462
10	132	3	213	100	1





CONCLUSION

Factors	Results/ Suggestion
Age	Employees with less AGE group are prone to quit the company
TrainingTimesLastYear	If employee is getting more frequent trainings, then employee is not quitting the organisation
YearsSinceLastPromotion	If Employees are getting frequent promotions then lesser chance of quitting the oraganisation when compared to employees whose promotions are delayed
YearsWithCurrManager	Employee who works with same manager for longer time has less chances of quitting the organisation
avg_ofc_hours	The more an employee works overtime on an average the more chances that employee will leave the organization
EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance	The better these are for employees the lesser Employee will quit the organisation
MaritalStatus	Employees who are single are prone to leave the company more