HR Analytics Case Study

SUBMISSION

Group Name: ***Data Cubed***

1. *Ashutosh Kumar (DDA1730374)*
2. *Medhavi Shruti (DDA1730035)*
3. *Daniel Narendra (DDA1730118)*
4. *Santanu Dey (DDA1730043)*

Abstract & Problem Statement for HR Analytics Case Study

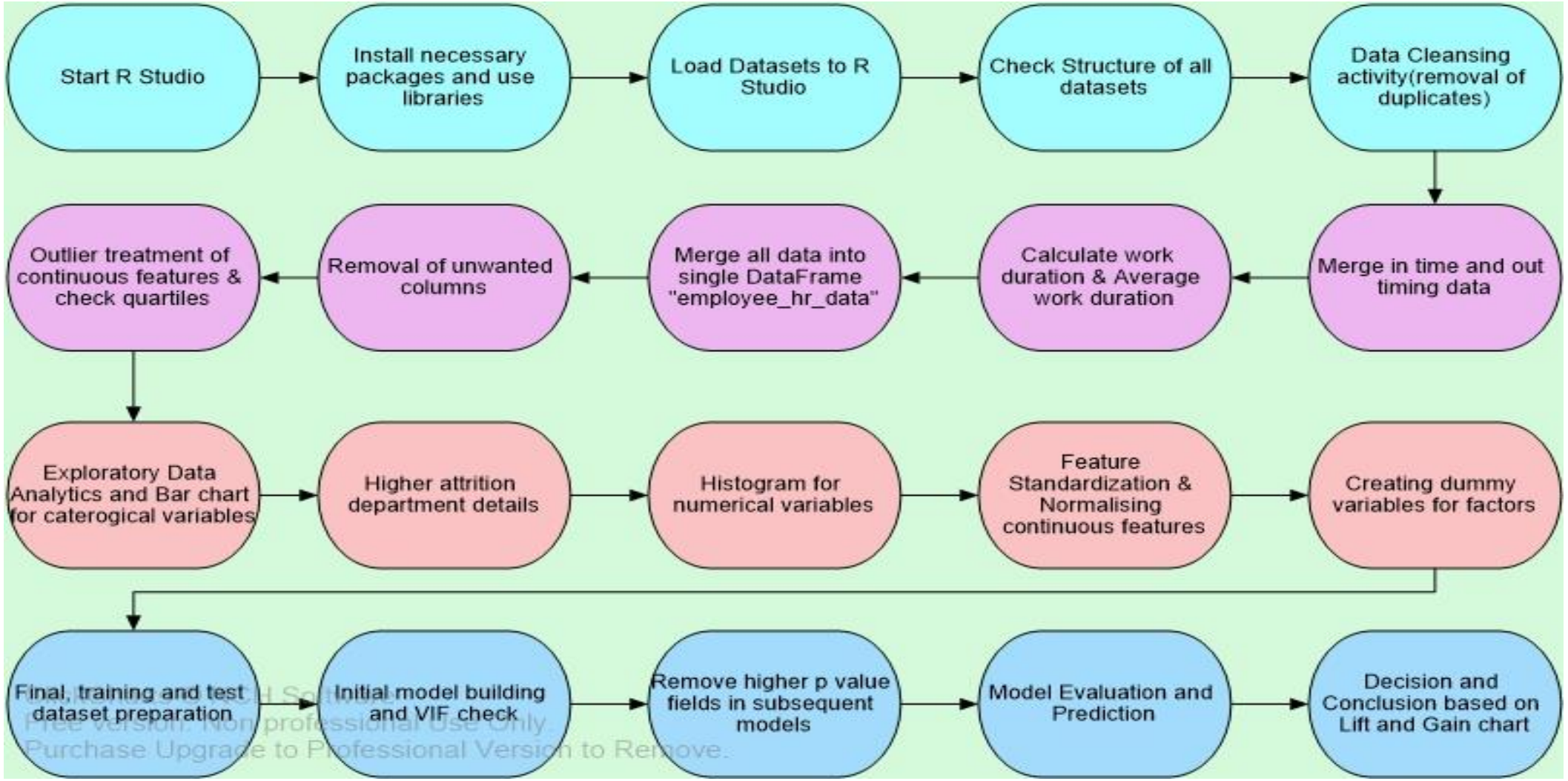
Introduction

- Company XYZ employs 4000 employees with an attrition rate of 15% employee.
- XYZ company need to fill the gap of employee with available talent pool from the job market.
- Management believes that attrition rate is higher and its bad for company business because :
  - Delivery of projects get delayed and there is reputation loss because of skilled and trained resource crunch.
  - Company need to maintain separate department to hire the new talent.
  - It requires significant amount of time to train the new joiners which eventually can impact the business in that duration.

Goals

- Company XYZ contracted HR Analytics firm to understand factors where they need to focus to curb attrition.
- Alternatively, company focus is to retain their employee and reduce the attrition rate.
- As part of HR Analytics firm we need to model the probability of attrition using logistic regression.
- Result of logistic regression will guide company do the certain changes to achieve their goal or to make changes in workplace so that most of employees stays.

Analysis Approach



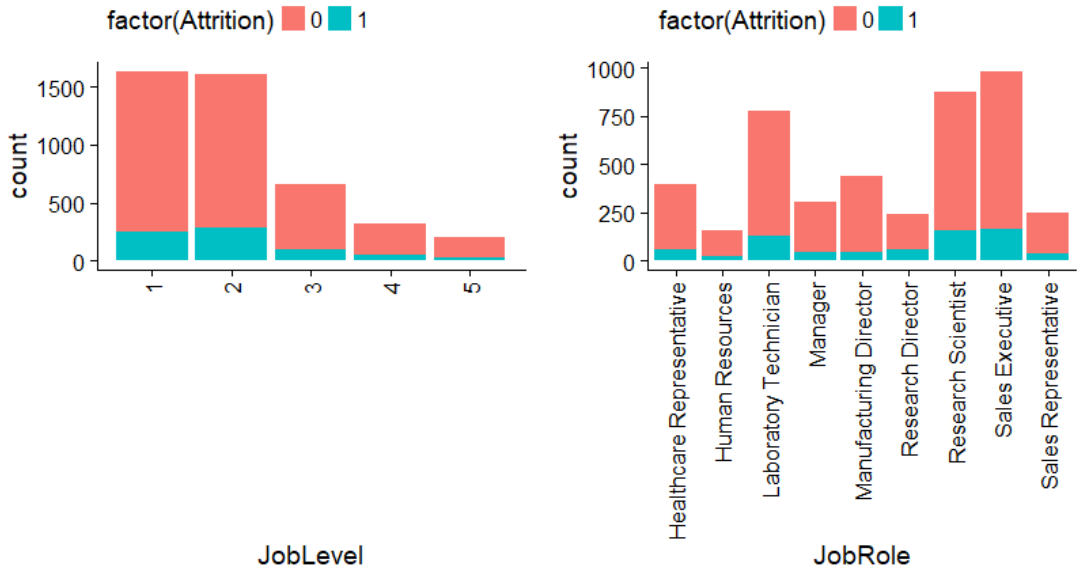
Analysis - Data Cleansing and Data Preparation

- Load all the csv files into data frames :
- general_data.csv to general_data
- employee_survey_data.csv to employee_survey data frame
- manager_survey_data.csv to manager_survey_data data frame
- in_time.csv to in_time data frame
- out_time .csv to out_time data frame
- Made EmployeeID as the name of first column for in_time and out_time data frame where the column name was missing.
- Checked for all the NA values and duplicate in all of the data frames.
- Replaced the NA's in NumCompaniesWorked column in general_data based on the TotalWorkingYears and the YearsAtCompany columns.
- Replaced the NA's in TotalWorkingYears column general_data based on YearsAtCompany and NumCompaniesWorked columns.
- Make use of the information in the sheet in_time and out_time and calculate the average working hours of each employee as this gives insights on the time employees spends at work.
- Merge all the data frames into one after checking and taking EmployeeID as the key in all the data frames.
- Check for Outliers by taking all the continuous variables into one subset data frame. Ignoring the outliers as for these variables there can be some observations which stay apart from others, so no treatment for them.

EDA – Categorical Variables

- We first focussed on **Categorical variables** and its role in Attrition
- Visualized Attrition using Bar charts and drew insights from it
- Attrition based on JobLevel shows that employees at level 1 and 2 are more likely to leave job as compared to other levels.
- Attrition based on JobRole shows that employees working as Lab Technician, Research Scientist, and Sales Executives mostly tend to leave.

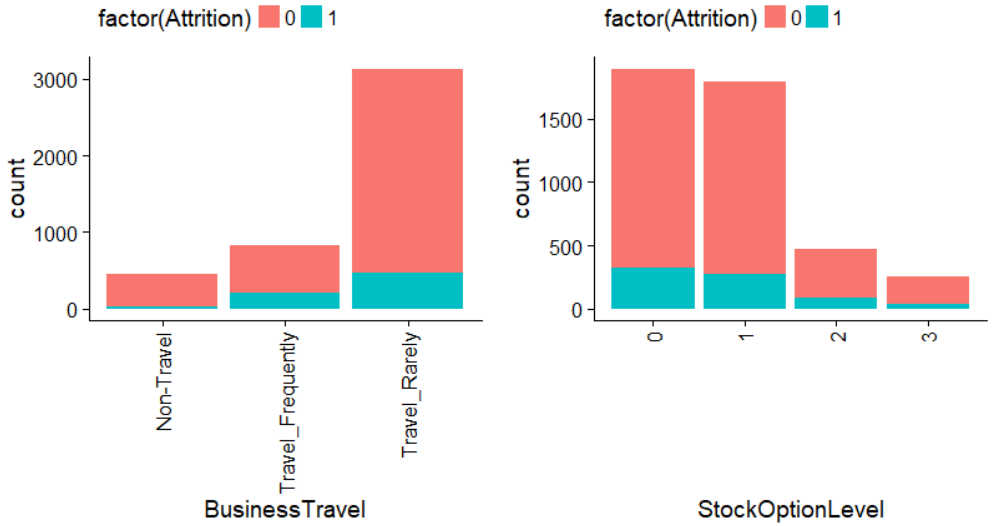
Plot1 : Attrition vs JobLevel and Attrition vs JobRole



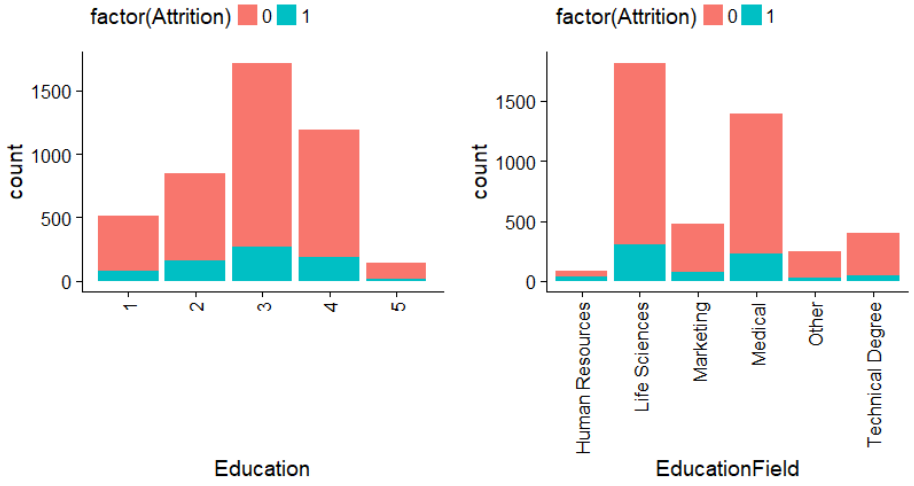
EDA – Categorical Variables (contd.)

- Attrition based on BusinessTravel shows that employees who get to travel less or rarely have high attrition.
- Attrition based on StockOptionLevel shows employees having less level of Stock option tend to leave
- Attrition based on Education shows that employees with having Bachelor or Masters have high attrition.
- Attrition based on EducationField shows employees from Life Sciences and Medicine domain tend to leave.

Plot 2 : Attrition vs Business Travel and Attrition vs Stock Option Level



Plot 3 : Attrition vs Education and Attrition vs EducationField

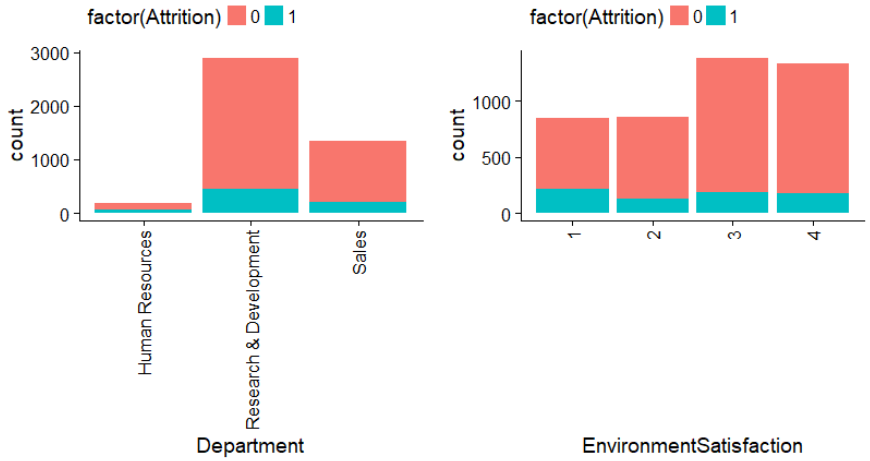


EDA – Categorical Variables (contd.)

- Attrition based on Gender shows that Male Employee tend to leave more than Female.
- Attrition based on MaritalStatus shows Single employees tend to leave more one of the factor being less family dependency.
- Attrition based on Department shows that employees from Research and Development department have high attrition may be because these employees get better opportunity in their field in some other company.
- Attrition based on EnvironmentSatisfaction show that Attrition is almost similar based on Environment Satisfaction Ratings.

Plot 4 : Attrition vs Gender and Attrition vs MaritalStatus

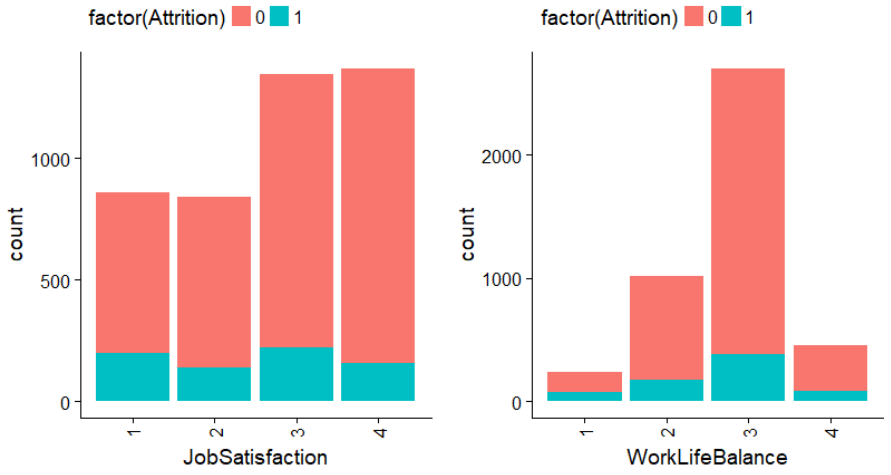
Plot 5 : Attrition vs Department and Attrition vs EnvironmentSatisfaction



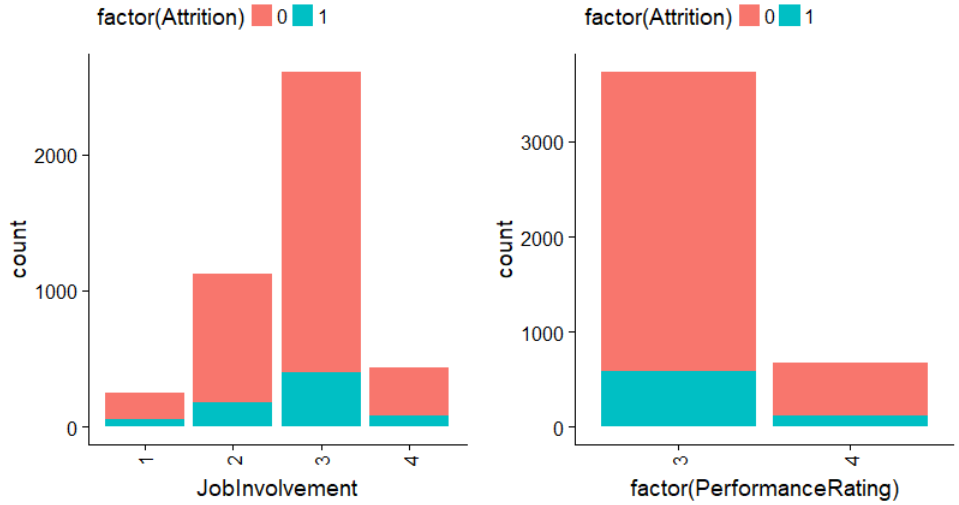
EDA – Categorical Variables (contd.)

- Plot of Attrition based on **JobSatisfaction** show that Attrition is almost similar based on Job Satisfaction ratings.
- Plot of Attrition based on **WorkLifeBalance** shows Single employees tend to leave more one of the factor being less family dependency.
- Plot of Attrition based on **JobInvolvement** shows that employees having High Involvement tend to leave the company.
- Plot of Attrition based on **PerformanceRating** shows that Excellent Performers of the previous year tend to leave more than other employees.

Plot 6 : Attrition vs JobSatisfaction and Attrition vs WorkLifeBalance

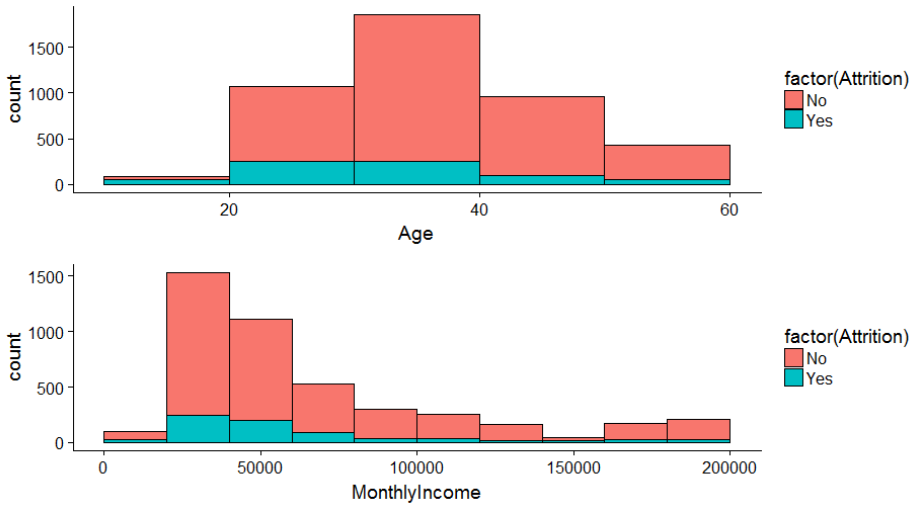


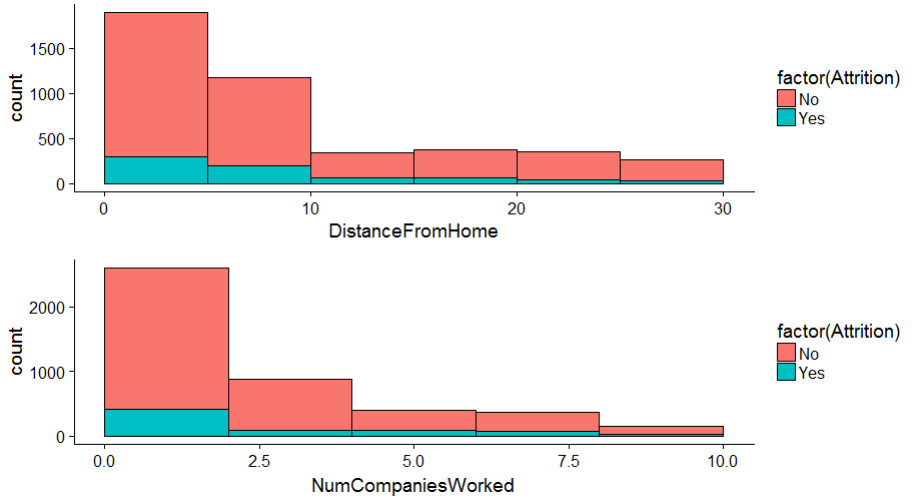
Plot 5 : Attrition vs JobInvolvmnet and Attrition vs Rating



EDA – Numeric Variables

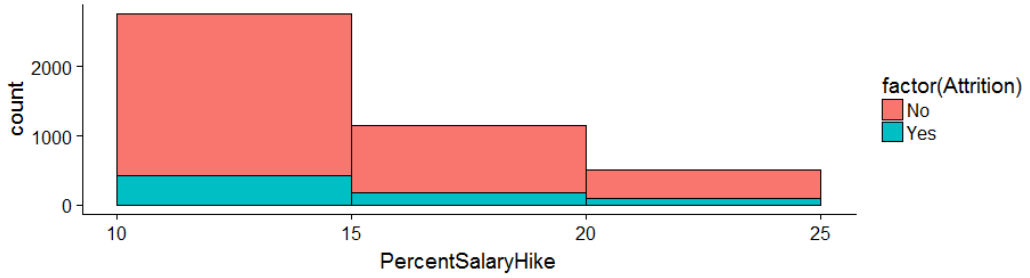
- We then focussed on **Numeric variables** and its role in Attrition
- Visualized numeric variables by histogram
- Attrition based on Age, MonthlyIncome, DistanceFromHome, and NumCompaniesWorked shows:
  - Attrition is high among employees in the Age group 20-40 years, and those with Monthly income 20K to 60K INR per month
  - Attrition is highest among employees whose homes are at a distance of upto 10 kms, and those who have worked in 2 companies

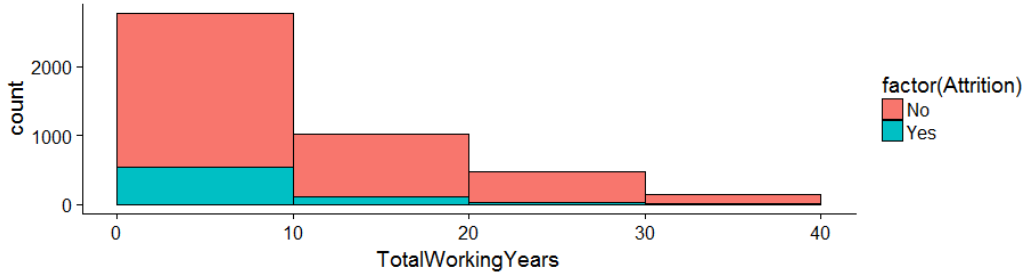


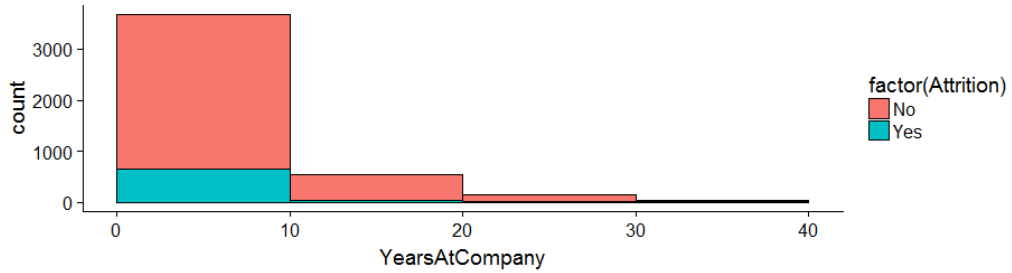


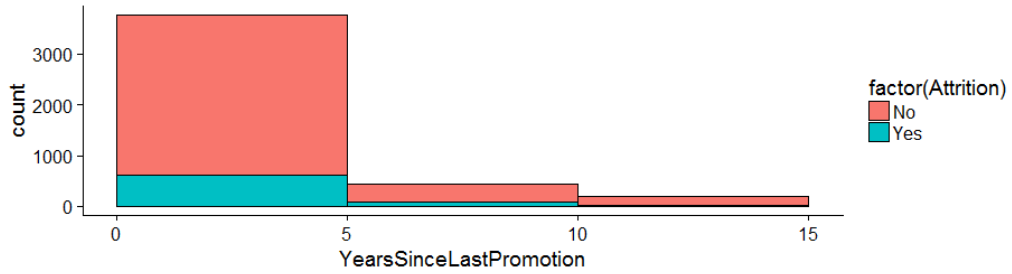
EDA – Numeric Variables (contd.)

- Attrition based on PercentSalaryHike, TotalWorkingYears, YearsAtCompany, and YearsSinceLastPromotion shows:
  - Attrition is highest among employees who get 10-15 percent Salary Hike, and those who have a total of upto 10 years work experience.
  - Attrition is highest among employees who have worked upto 10 years in the current company, and those who have not received a promotion in the last 0-5 years.









EDA – Numeric Variables (contd.)

- Attrition based on YearsWithCurrManager, and avg_work_hours shows:
  - Attrition is highest among employees among employees who have worked upto 5 years with the current manager, and is similar for those with 6-11 average working hours.

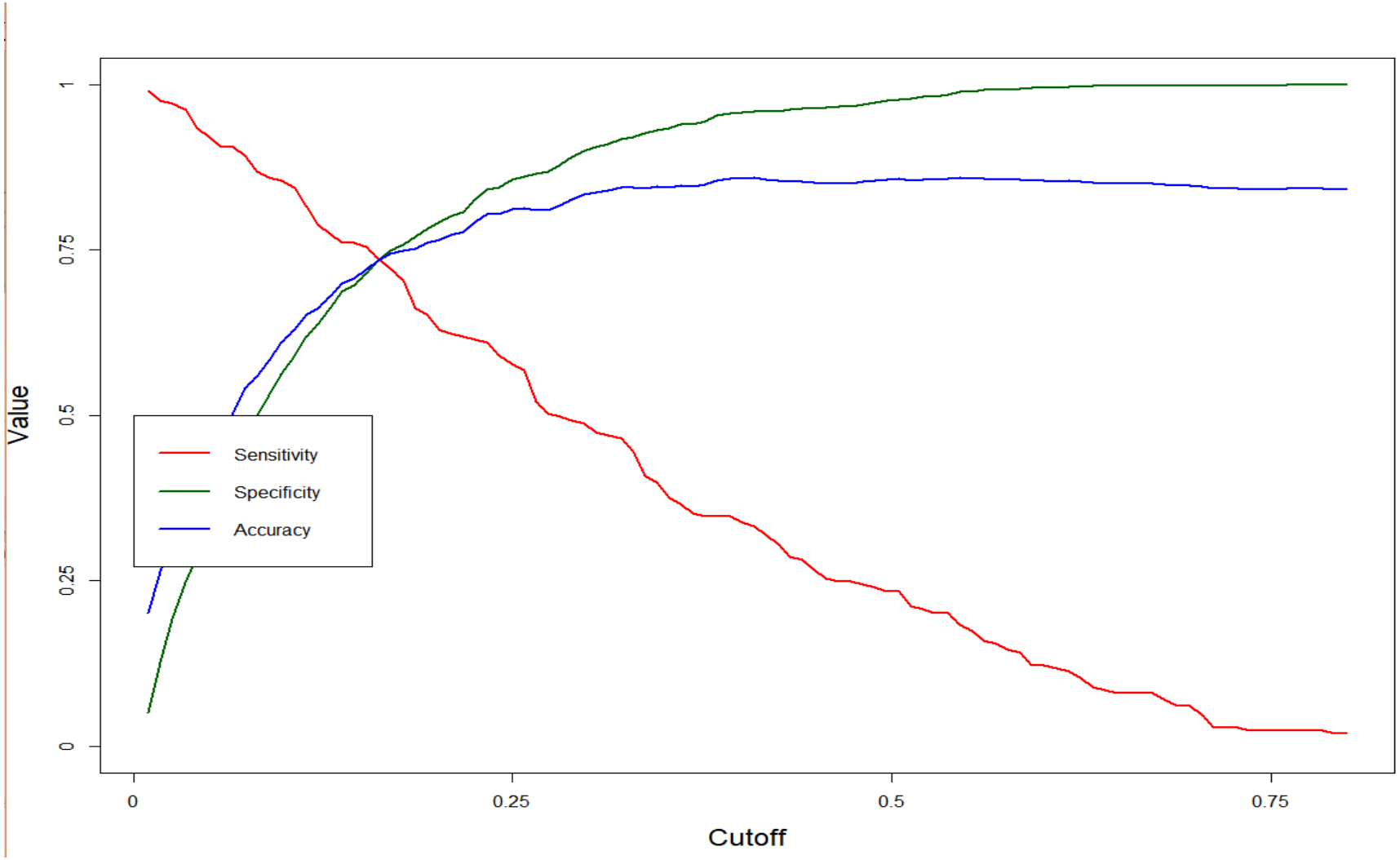
Model Building, Evaluation & Prediction Approach

- Logistic regression is used for model building and prediction
- R generalized linear model glm() function is used for initial model building with training dataset for family type Binomial
- To perform stepwise selection in model stepAIC function is used to derive model 2 in scope of both the direction
- Check for the multicollinearity using R VIF check function in model 2
- Variables with high VIF cannot be excluded as high VIF variables have high significance as well
- Excluded variable with highest p-value at each step
- Performed 19 iterations to get the final model for training data set.
- Using final model, prediction is done for test data set
- Find out the optimal probability cut-off.
- Sensitivity, specificity, accuracy is calculated and plotted for the cut-off value for test data
- The same is repeated for training + test data to derive the evaluation metrics

Final Model

	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	-1.21634	0.34715	-3.504	0.000459	***
Age	-0.33383	0.07851	-4.252	2.12E-05	***
NumCompaniesWorked	0.35247	0.05602	6.291	3.14E-10	***
TotalWorkingYears	-0.4828	0.10642	-4.537	5.71E-06	***
YearsSinceLastPromotion	0.51719	0.07655	6.756	1.41E-11	***
YearsWithCurrManager	-0.49073	0.08484	-5.784	7.28E-09	***
avg_work_hours	0.52634	0.0529	9.95	< 2e-16	***
BusinessTravel.xTravel_Frequently	1.84118	0.27667	6.655	2.84E-11	***
BusinessTravel.xTravel_Rarely	1.13658	0.26108	4.353	1.34E-05	***
Department.xResearch...Development	-1.17443	0.22085	-5.318	1.05E-07	***
Department.xSales	-1.23533	0.23264	-5.31	1.10E-07	***
MaritalStatus.xSingle	1.01243	0.11294	8.964	< 2e-16	***
EnvironmentSatisfaction.x2	-0.97585	0.16816	-5.803	6.51E-09	***
EnvironmentSatisfaction.x3	-0.95802	0.15043	-6.369	1.91E-10	***
EnvironmentSatisfaction.x4	-1.21984	0.15538	-7.851	4.14E-15	***
JobSatisfaction.x4	-0.78513	0.13131	-5.979	2.24E-09	***
WorkLifeBalance.x3	-0.38437	0.11187	-3.436	0.000591	***

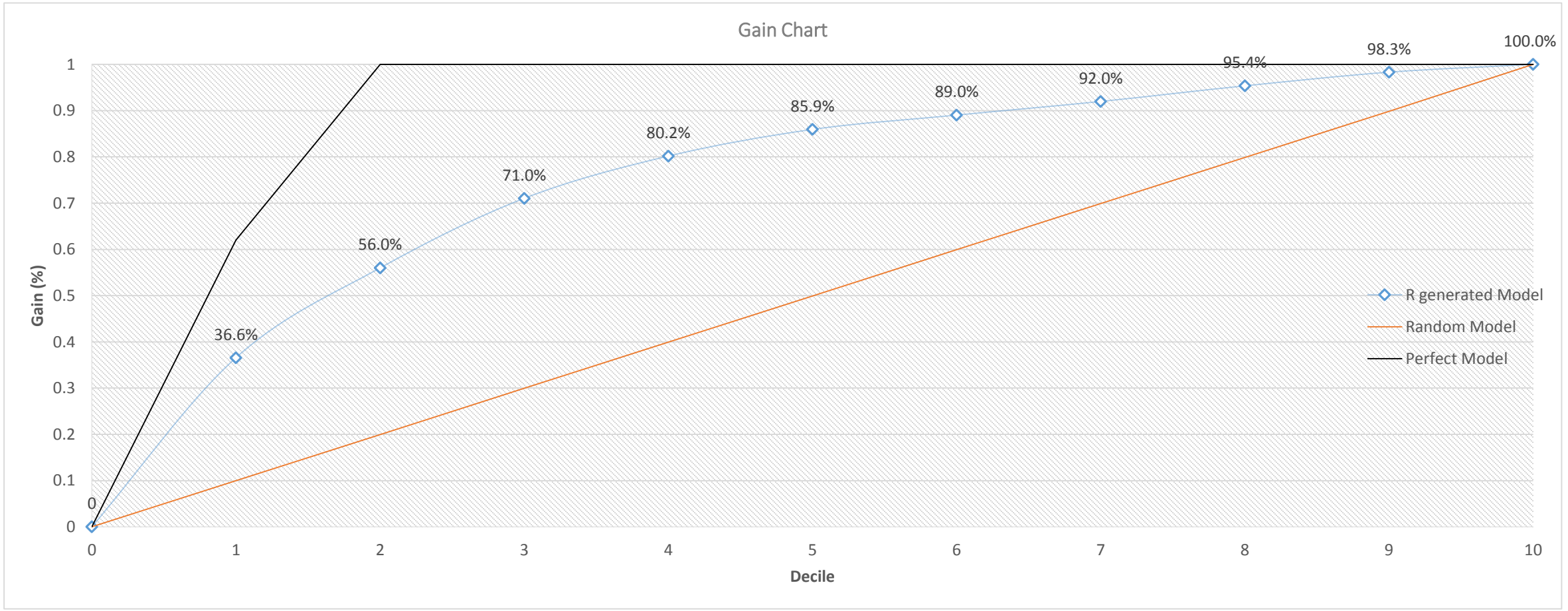
Sensitivity, Specificity and Accuracy



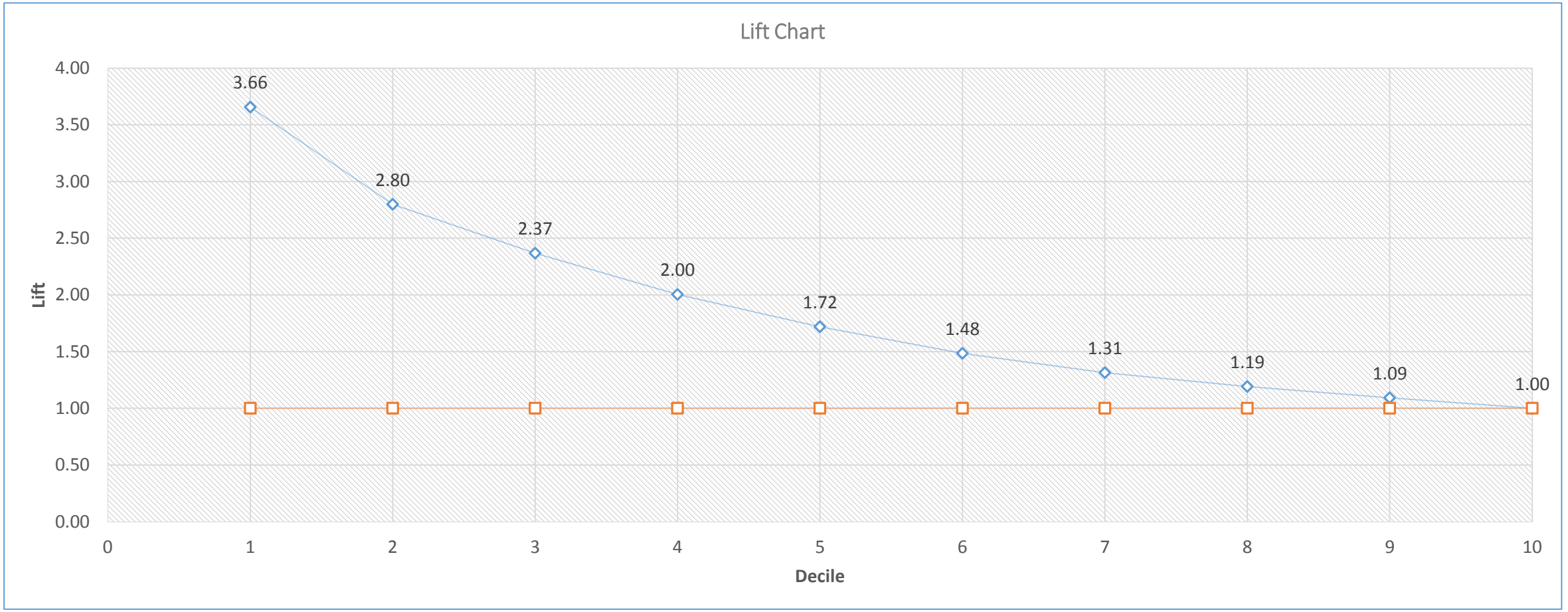
Model Evaluation

Metric	Test Data	Train + Test Data
Accuracy	73.47%	74.56%
Sensitivity	73.24%	77.22%
Specificity	73.51%	74.05%
KS-statistic	46.75%	51.26%
Cutoff	0.162	0.162

Gain chart


Gain Chart
Gain (%)
Decile
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
1
2
3
4
5
6
7
8
9
10
0
36.6%
56.0%
71.0%
80.2%
85.9%
89.0%
92.0%
95.4%
98.3%
100.0%
R generated Model
Random Model
Perfect Model

Lift chart


Lift Chart
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
Lift
3.66
2.80
2.37
2.00
1.72
1.48
1.31
1.19
1.09
1.00
0
1
2
3
4
5
6
7
8
9
10
Decile

Conclusion

The following variables impact the Attrition (positively or negatively) as per our model:

- Age (-): The employees whose age is comparatively less tend to leave more
- NumCompaniesWorked (+): The more the companies one has worked in, the more the chance of one's attrition
- TotalWorkingYears (-): The more the total experience, the less the chance of leaving
- YearsSinceLastPromotion (+): The more the years since last promotion, the more the chance of attrition
- YearsWithCurrManager (-): The more the years with current manager, the less the chance of attrition
- avg_work_hours (+): The more the average working hours, the more the attrition
- BusinessTravel.xTravel_Frequently (+): The employees who travel frequently tend to leave
- BusinessTravel.xTravel_Rarely (+): The employees who travel rarely also tend to leave
- Department.xResearch...Development (-): The employees in R&D department tend not to leave
- Department.xSales (-): The employees in Sales department tend not to leave
- MaritalStatus.xSingle (+): Single employees tend to leave
- EnvironmentSatisfaction.x2 (-): More the environment satisfaction, less the attrition
- EnvironmentSatisfaction.x3 (-): More the environment satisfaction, less the attrition
- EnvironmentSatisfaction.x4 (-): More the environment satisfaction, less the attrition
- JobSatisfaction.x4 (-): More the job satisfaction, less the attrition
- WorkLifeBalance.x3 (-): More the work life balance, less the attrition

Conclusion (contd.)

The HR department should focus on the following aspects to reduce attrition:

- Most of the new hires should be those who have worked in 2 or more companies previously, or have 10+ years of experience
- Avoid change of team/manager for employees to increase chances of retention
- Regularly promote employees or increase opportunities of promotion based on skillsets or goals achieved
- Reduce average working hours by focusing on flexible work timings
- Balance travel opportunities among employees
- Increase headcount in Sales department
- New hires should preferably be married
- Introduce measures to improve Environment Satisfaction, Job Satisfaction, and WorkLifeBalance feedbacks from employees

Thanks