A CASE STUDY ON HR ANALYTICS USING PREDICTIVE MODELLING BY MEANS OF LOGISTIC REGRESSION

PRESENTED BY:

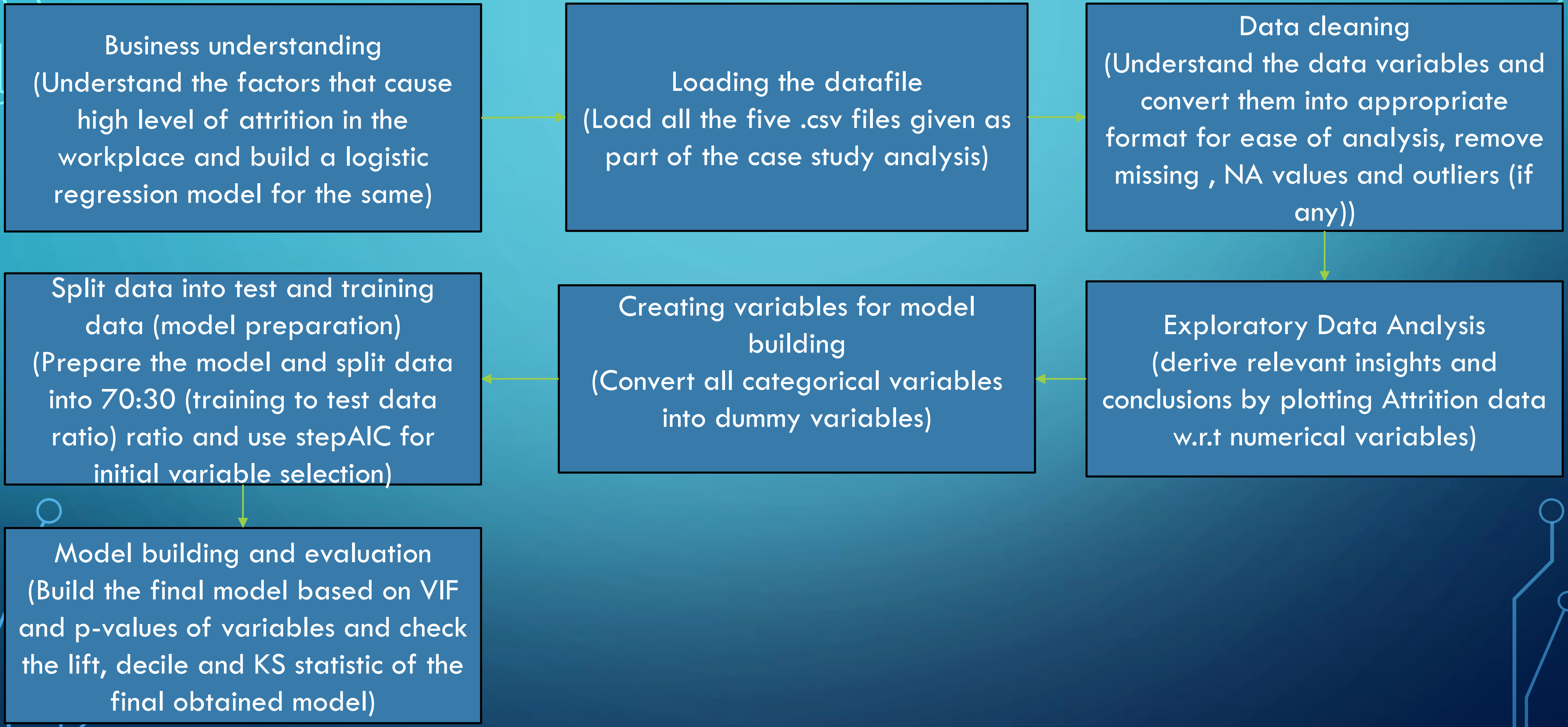
GROUP NAME:

1. VIVEK ANGADI
2. SRI HARSHA
3. UJWAL KIRAN BHARGAVA
4. ARITRA MAJUMDAR

INTRODUCTION, BACKGROUND AND PROBLEM STATEMENT

- A large company XYZ having around 4000 employees has a very high level of attrition (15% of its employees leave every year)
- This leads to major issues like missed project timelines and reputation loss among customers and partners.
- Thus, the main objective of this case study is to model the probability of attrition using logistic regression based on various factors.
- The results obtained using the final model will be used by the management to understand what changes they should make to their workplace in order to reduce this high rate of attrition.

PROBLEM SOLVING METHODOLOGY



Business understanding
(Understand the factors that cause high level of attrition in the workplace and build a logistic regression model for the same)

Loading the datafile
(Load all the five .csv files given as part of the case study analysis)

Data cleaning
(Understand the data variables and convert them into appropriate format for ease of analysis, remove missing , NA values and outliers (if any))

Exploratory Data Analysis
(derive relevant insights and conclusions by plotting Attrition data w.r.t numerical variables)

Creating variables for model building
(Convert all categorical variables into dummy variables)

Split data into test and training data (model preparation)
(Prepare the model and split data into 70:30 (training to test data ratio) ratio and use stepAIC for initial variable selection)

Model building and evaluation
(Build the final model based on VIF and p-values of variables and check the lift, decile and KS statistic of the final obtained model)

UNDERSTANDING DATA AND ASSUMPTIONS USED IN CASE STUDY

There are 5 .csv data files used for analysis in this case study.

- employee_survey_data.csv - contains details from employees about how well they are satisfied with their job and work-life balance
- manager_survey_data.csv - contains details from managers about how well employees perform under them
- general_data.csv - contains details about employee location and other aspects like employee behaviour and their involvement in job
- in_time.csv - contains details regarding in time of employees in calendar year 2015
- out_time.csv - contains details regarding out time of employees in calendar year 2015

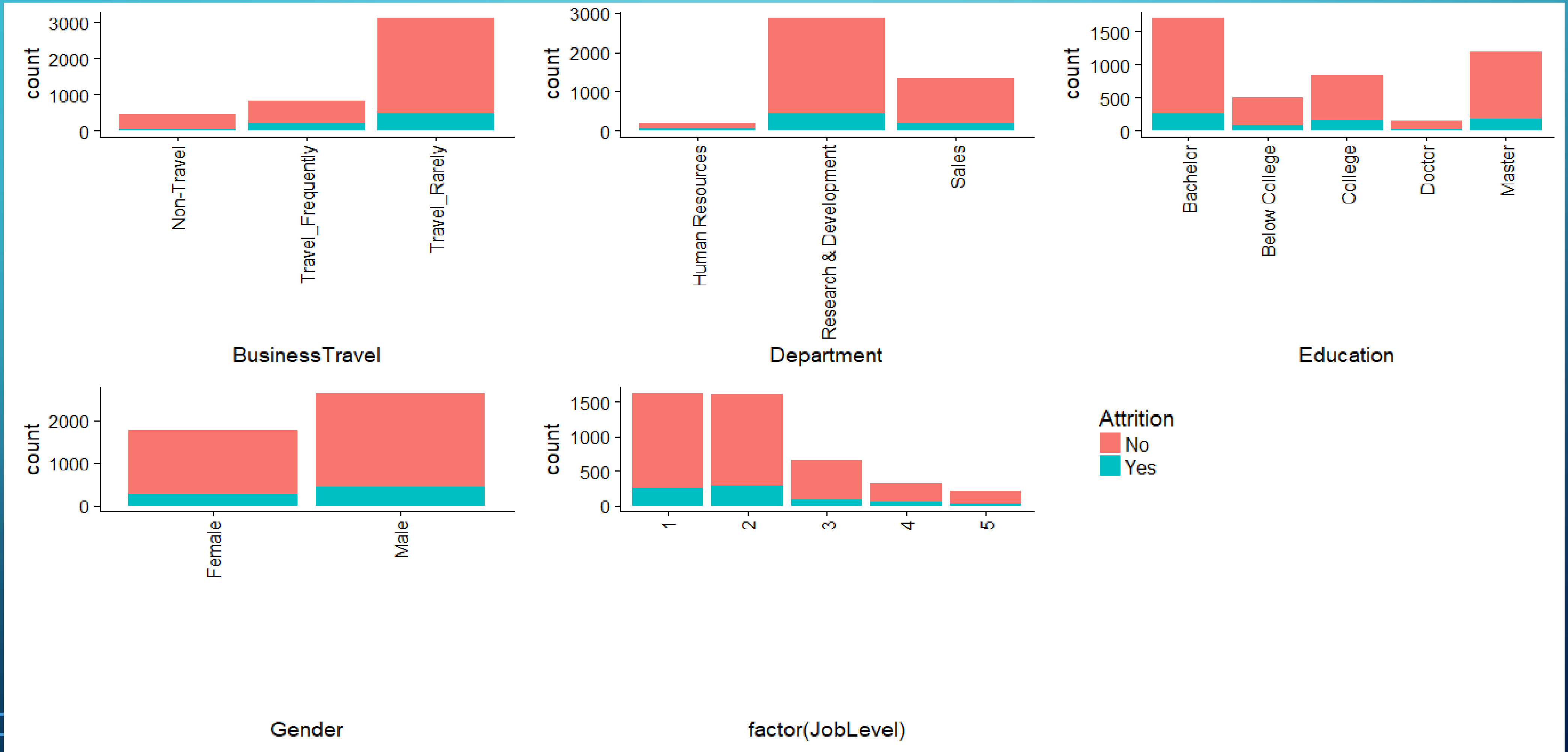
UNDERSTANDING DATA AND ASSUMPTIONS USED IN CASE STUDY

- The columns having all NA's or 0's as row value are discarded.
- The columns having only one unique row value are also discarded.
- All character type variables are converted into factor type for ease of analysis.
- All month-year type variables are converted into appropriate data format for ease of analysis.
- 70% of the data is used for training and remaining for testing in accordance with standard practices.
- For removal of variables from model, VIF >2 and p-value >0.05 are considered in accordance with standard practices.
- Eda is performed based on attrition data w.r.t various factors to get better insights using these results.

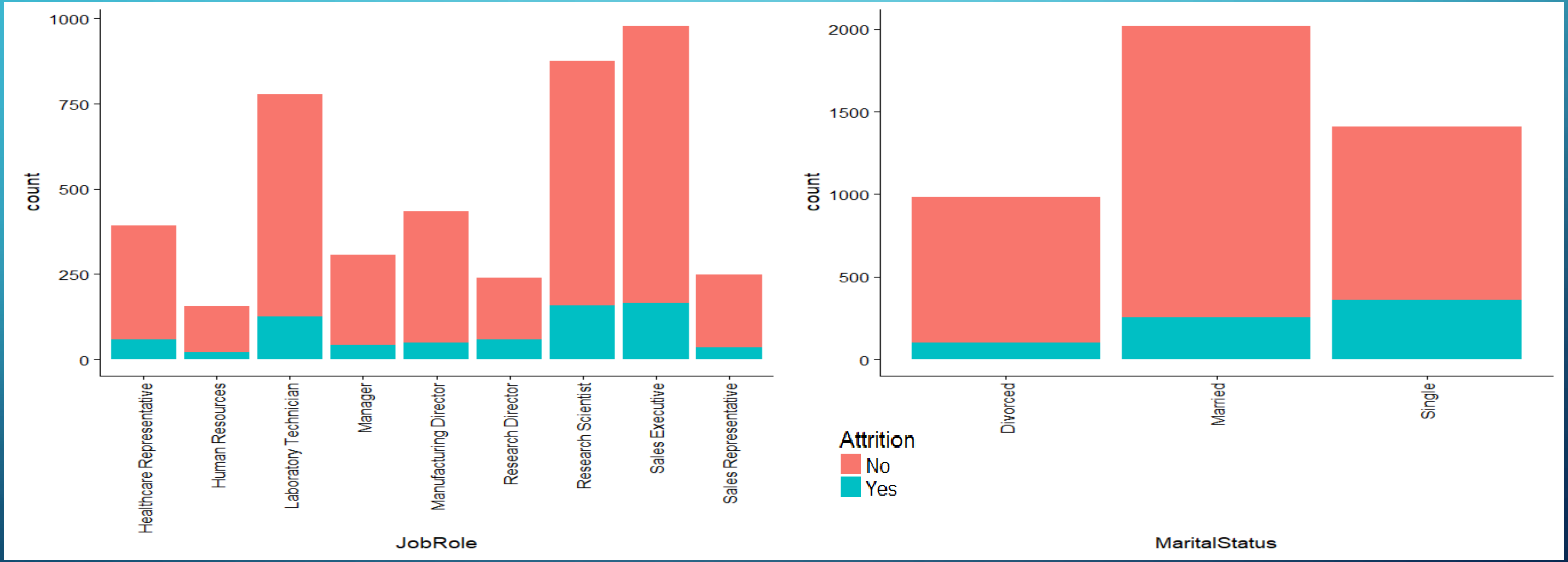
UNDERSTANDING DATA AND ASSUMPTIONS USED IN CASE STUDY

- The columns having NA values (more than 15% of the dataset) in intime and outtime datasets are removed.
- The X prefixes in all columns of intime and outtime datasets are removed for ease of analysis.
- The NA values in individual columns of the employee and general datasets are replaced with their respective medians and means, respectively for ease of analysis.
- Normalising and standardization of continuous features of numeric variables is done for consistency in model building and analysis.

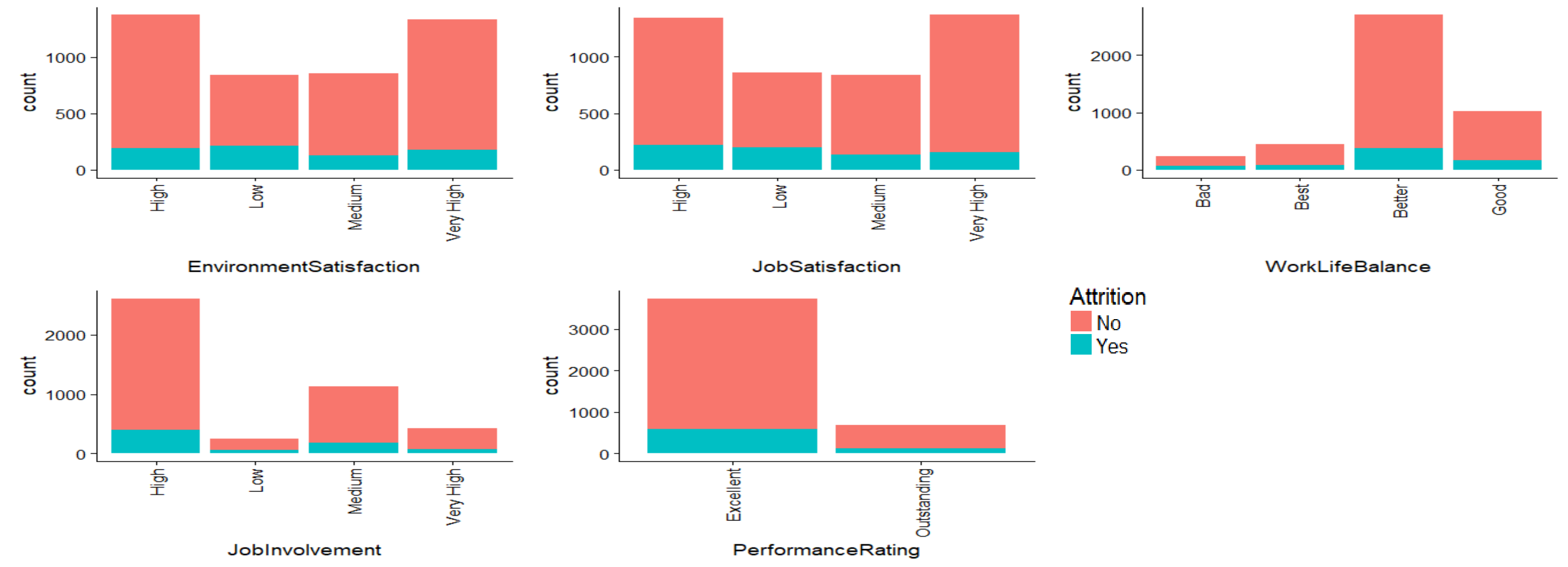
UNIVARIATE DATA ANALYSIS



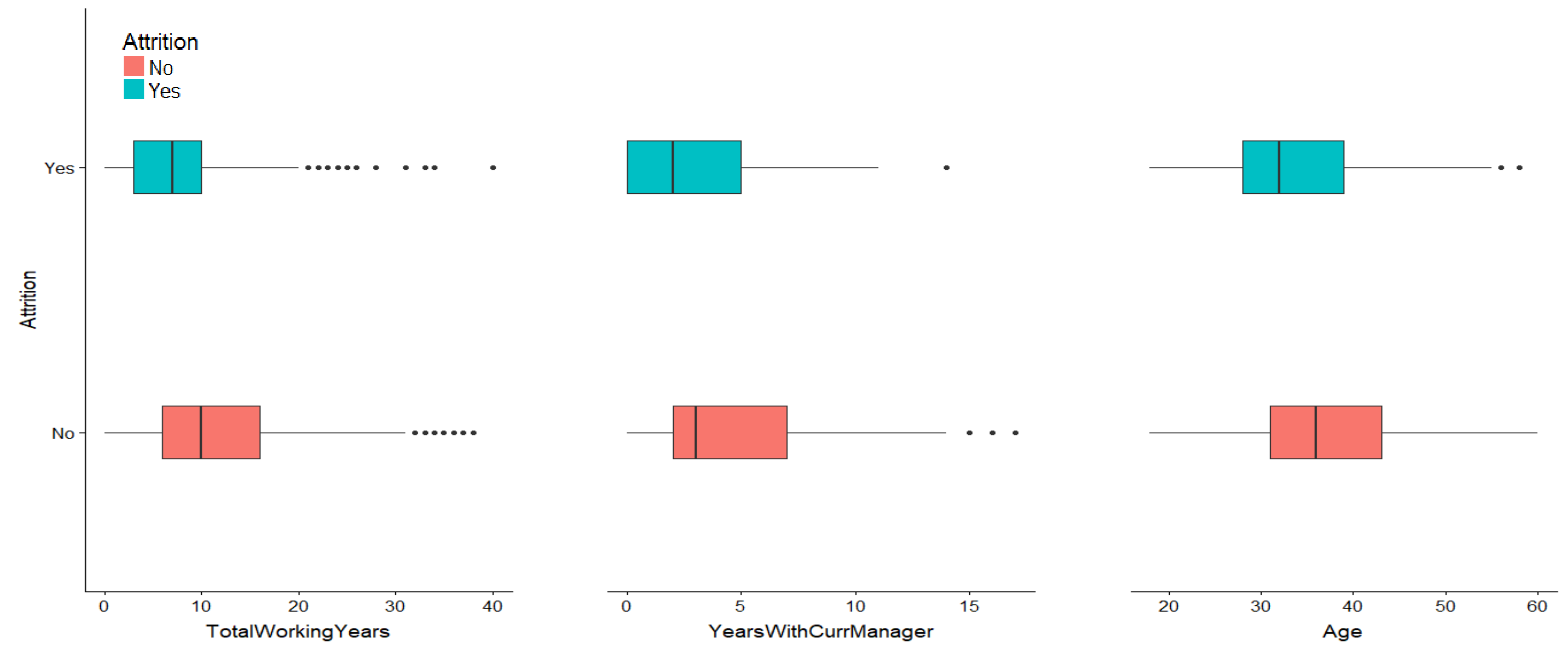
UNIVARIATE DATA ANALYSIS(CONTD..)



UNIVARIATE DATA ANALYSIS(CONTD..)



BOX PLOT FOR CONTINUOUS VARIABLE



CONCLUSIONS FROM THE EDA

- From the plots obtained, it is clear that as the work experience of employees increases the attrition rate decreases. i.e employees with lesser work experience switch more frequently.
- Also, as the years with current manager increase, attrition rate decreases.

DERIVED METRICS USED

The following are the three major derived metrics used in this case study.

- avg_hours – Indicates the average working hours per day for each employee.
- Overtime – Indicates if the employee has worked for more than the usual 8 hours per day.
- Undertime – Indicates if the employee has worked for less than 7 hours period per day.
- Leaves – Indicates the number of holidays taken by the employee in the calendar year (2015).

UNDERSTANDING CATEGORICAL VARIABLES FOR MODEL BUILDING

- All the two-level and multi-level categorical variables in the final merged data set (EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance, Education, JobLevel JobInvolvement, PerformanceRating, Attrition and Gender) are converted into factor types.
- Same approach is used for both the two-level and multi-level categorical variables since for each n-level categorical variable, n-1 dummy variables are created.
- So, it does not matter w.r.t. two-level categorical variables as only one dummy variable is created for each of them.
- Hence, the NULL values obtained as a result (if any) can be ignored.
- Here, AttritionYes is used as the dependent variable for modelling as we are only concerned about employees who are actually leaving the company.

UNDERSTANDING THE MODEL BUILDING

- The seed is set to 100 and split ratio 0.7 is used.
- This is because, as per standard practice, 70% of data is used for training purpose and the remaining for test purpose.
- Using stepAIC, variables are iteratively removed for model building for model_3.
- By removing variable based on VIF and p-values, the final model is built after 17 iterations, and thus 20 models are created.

FINAL MODEL

- The following is the final model used for analysis.

```
model_20<-glm(formula = AttritionYes ~
WorkLifeBalanceGood+JobSatisfactionLow+WorkLifeBalanceBest+
JobSatisfactionVery.High+NumCompaniesWorked+YearsWithCurrManager+
WorkLifeBalanceBetter+BusinessTravelTravel_Frequently+
YearsSinceLastPromotion+EnvironmentSatisfactionLow+TotalWorkingYears
+MaritalStatusSingle+overtime,family = "binomial",data = train)
```

- The final model (model_20) is chosen since all variables have p-value less than 0.001.
- This is because, number of variables in the final model should not be too large.

CONCLUSIONS AND FINAL OBSERVATIONS BASED ON THE FINAL MODEL

From the final model obtained (model-20), it is clear that,

- Factors like work life balance, job satisfaction, environment satisfaction, years with current manager, marital status and business travel affect the attrition rate.
- People having low environmental and job satisfaction, working overtime and unmarried people are more prone to leaving the company.
- People who travel frequently for business purposes are also more prone to leaving the company.

CONCLUSIONS AND FINAL OBSERVATIONS BASED ON THE FINAL MODEL

- Thus based on the conclusions obtained from the final model, it is advisable that more employee welfare schemes should be introduced by the management to improve employee satisfaction and morale.
- Greater work life balance should be ensured to help retention of new talent in the workplace
- The frequency of the business travel of employees should be reduced.
- This helps in greater employee retention and thus, reduces the rate of attrition.

MODEL ASSESSMENT AND EVALUATION

- Confusion Matrix

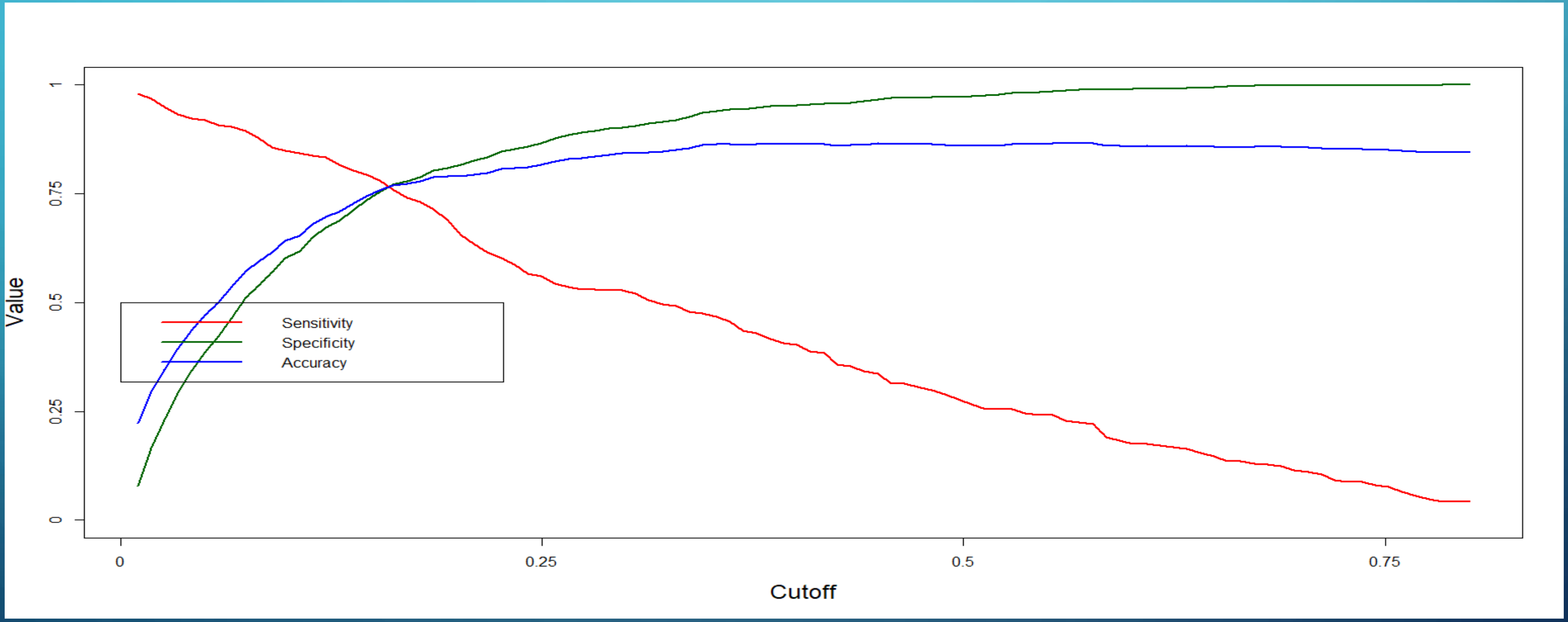
Prediction	No	Yes
No	2848	171
Yes	850	540

- Table of accuracy values

Characteristic	Value
Accuracy	0.7684
Sensitivity	0.7595
Specificity	0.7701
KS Statistic	0.3319
Optimum Probability Cut-off	0.1616

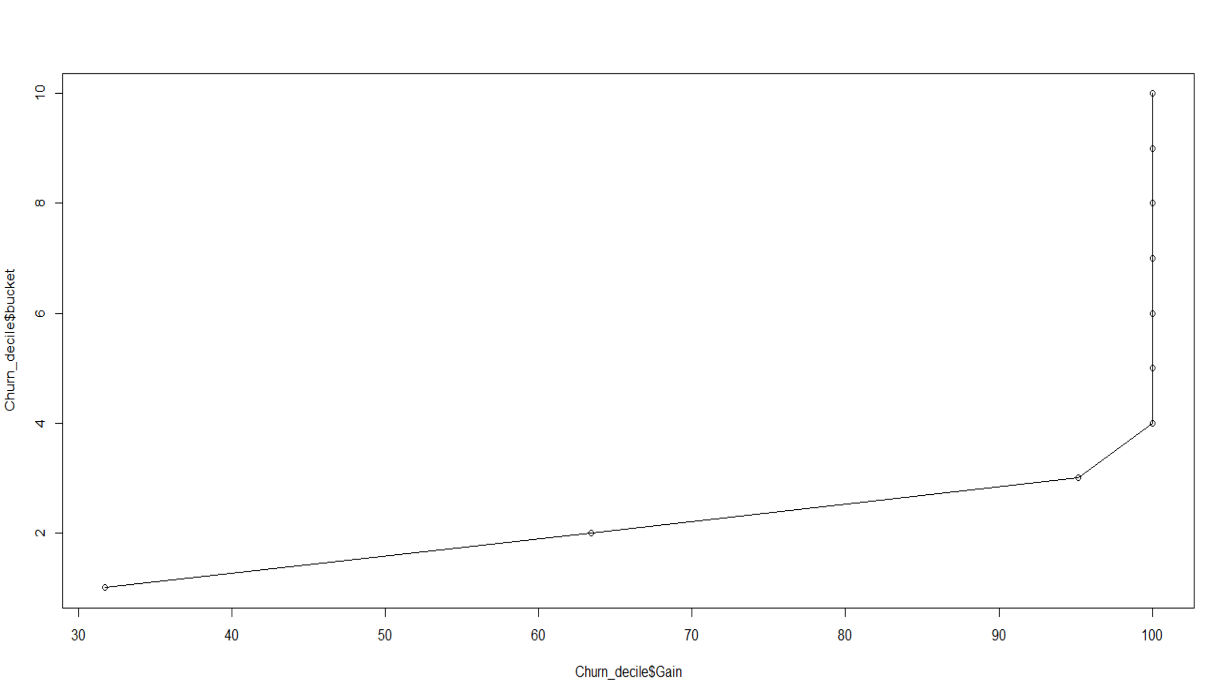
MODEL ASSESSMENT AND EVALUATION (CONTD..)

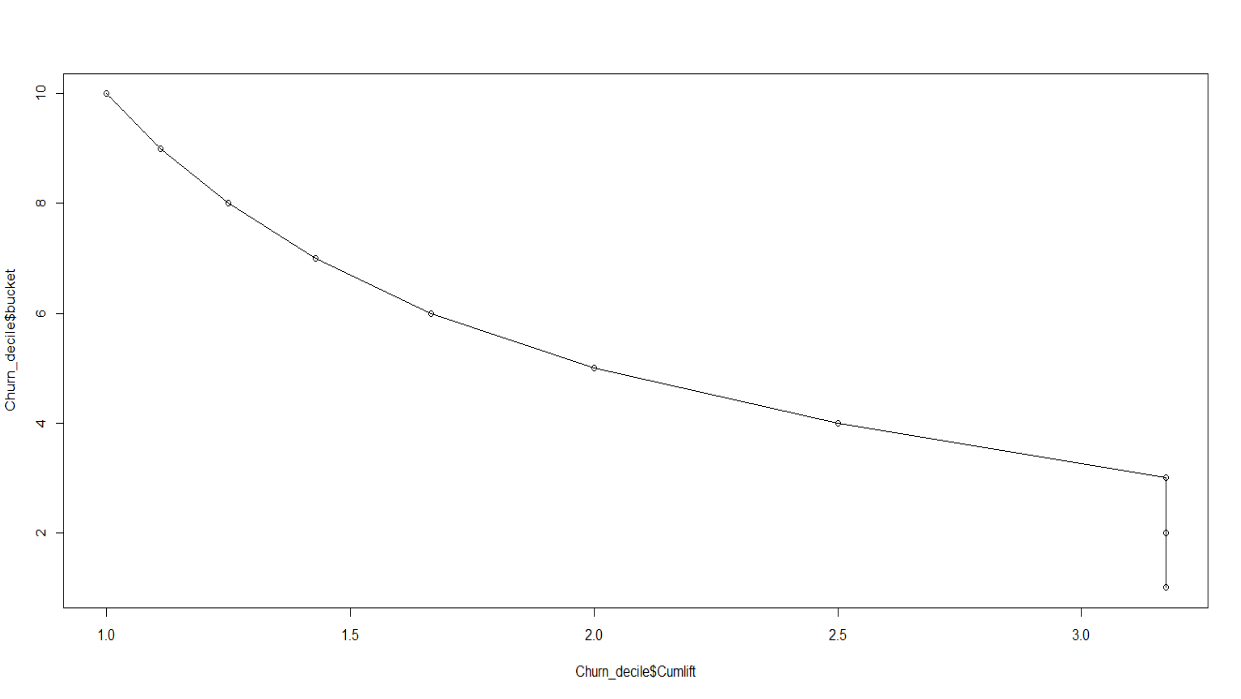
- ROC Curve



MODEL ASSESSMENT AND EVALUATION (CONTD..)

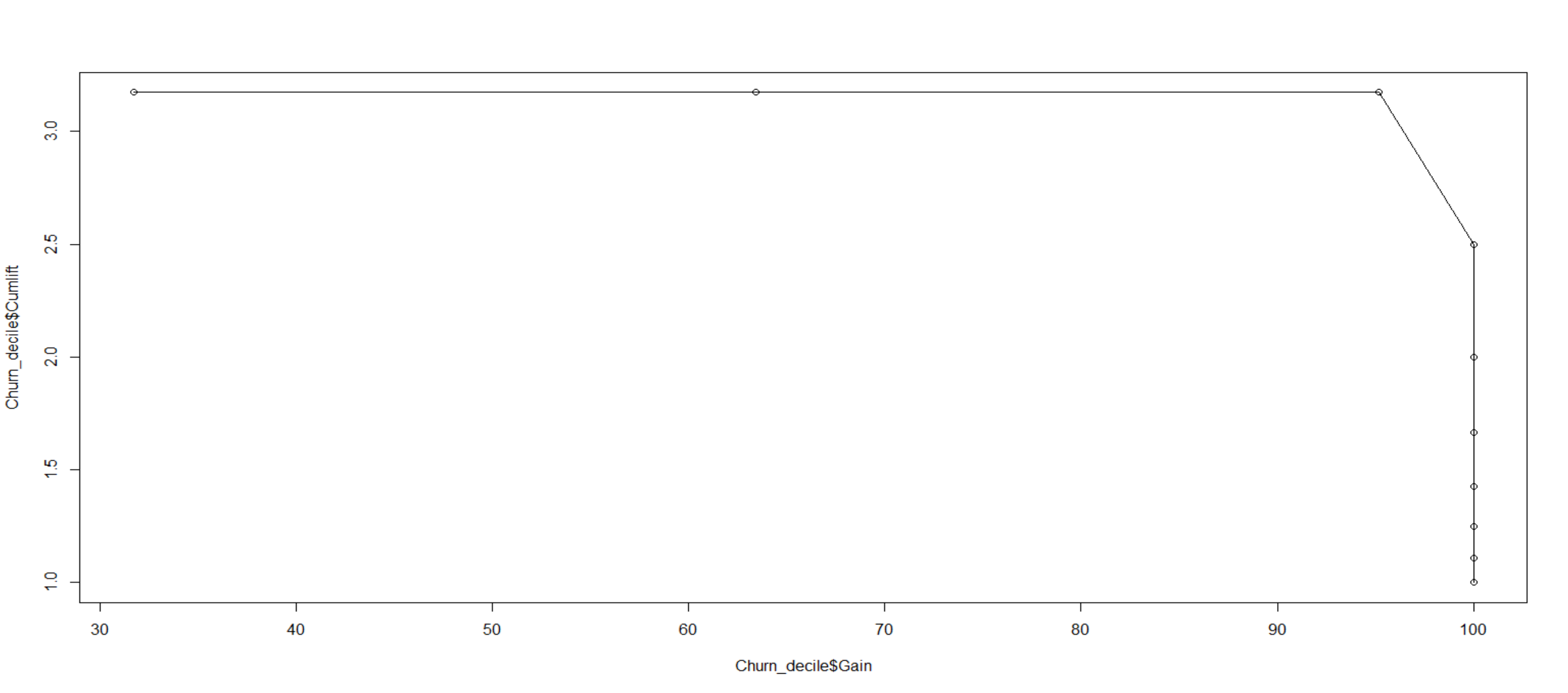
- Gain and Lift charts





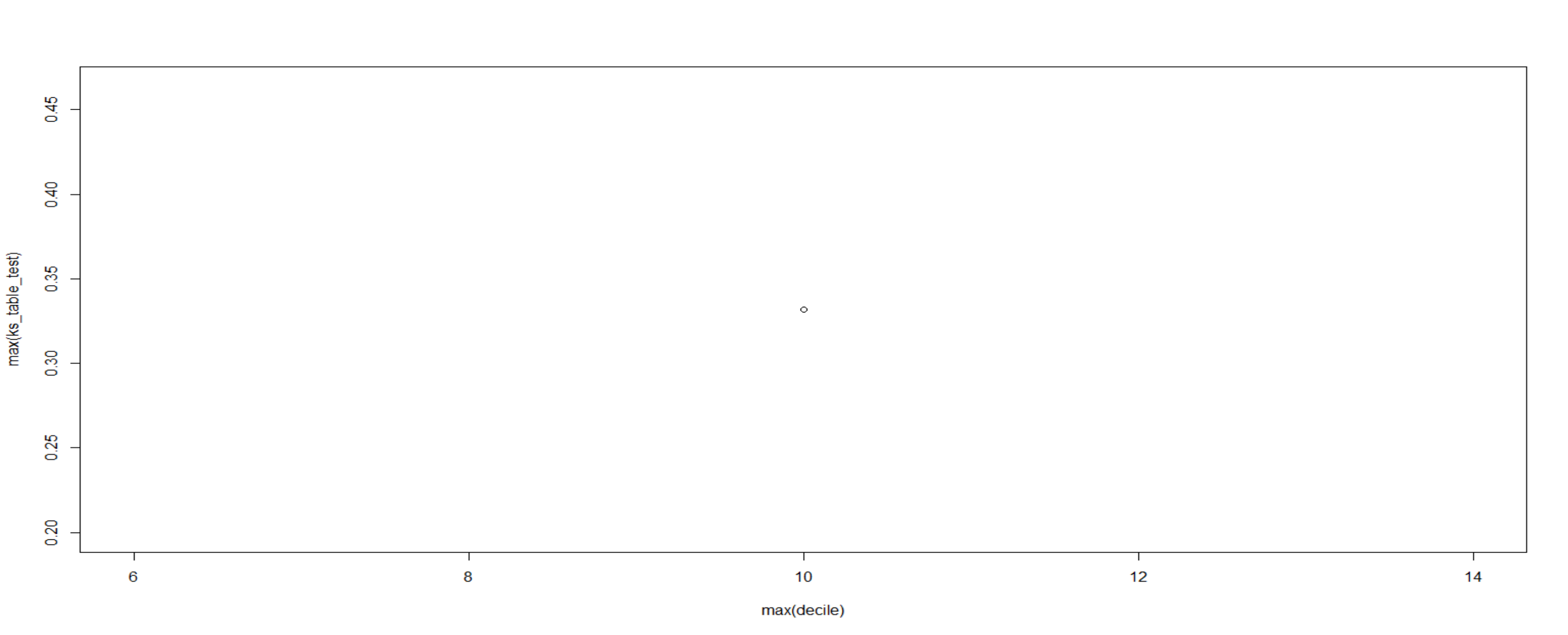
MODEL ASSESSMENT AND EVALUATION (CONTD..)

- Gain vs Lift chart



MODEL ASSESSMENT AND EVALUATION (CONTD..)

- KS Statistic Analysis



MODEL ASSESSMENT AND EVALUATION SUMMARY

- The accuracy, sensitivity, specificity and KSS values of the final model are 0.7684, 0.7595, 0.7701and 0.3319, respectively.
- From the optimum probability cut-off (0.1616) and the ROC curve obtained, it is clear that the model is a good fit when it comes to accuracy.
- From the gain and lift charts obtained, it is clear that the model has increasing gain and decreasing lift.
- It is clear that the model predicts attrition more accurately as most of the values are accurately predicted by the 4th decile.
- Thus, from the above conclusions regarding model assessment and evaluation, it is clear that the final model accurately predicts the rate of attrition and is the model suitable and appropriate model for analysis.