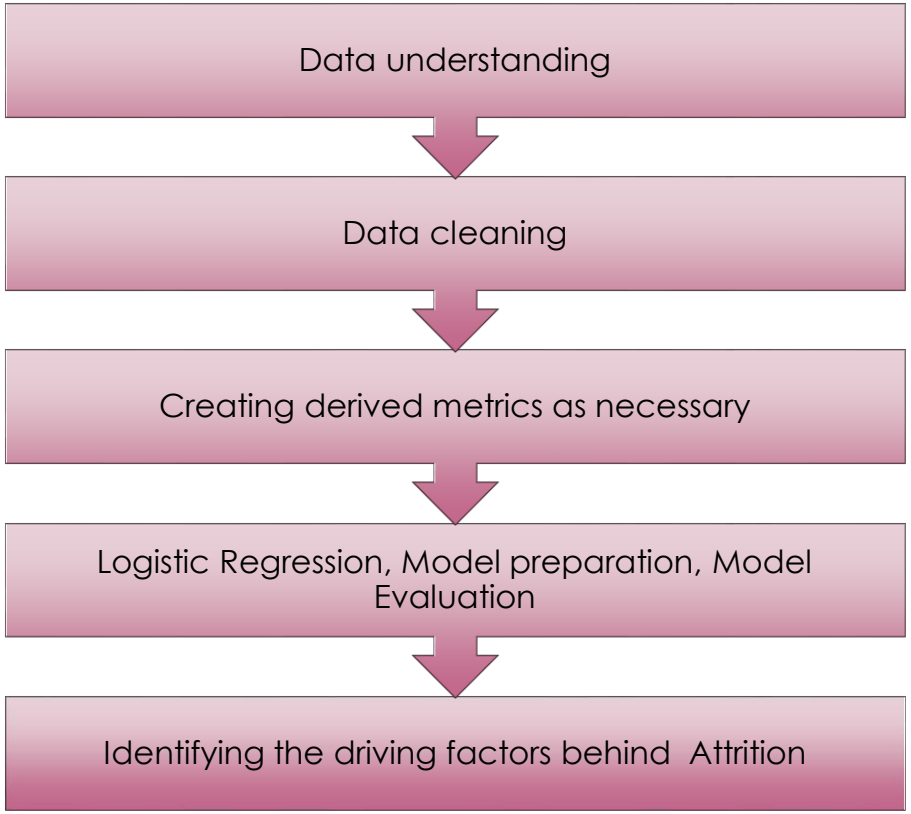
HR Analytics Case study –
Probability of attrition using Logistic Regression

PRADNYA PAITHANKAR
DR. RAJENDRA WARKE
VAIBHAV PALKAR
SAMEER SINHA

INTRODUCTION

- Preamble: A company named XYZ faces challenge of
  - Employees leaving, either on their own or because they got fired
  - 15% of employee attrition every year and need to replace this gap with talent pool available in open market
- Case study objective
  - To curb attrition
  - understand the probability of attrition using a logistic regression Inputs available
  - XYZ has employees around 4410 employees

Methodology followed



Input data – Preliminary understanding

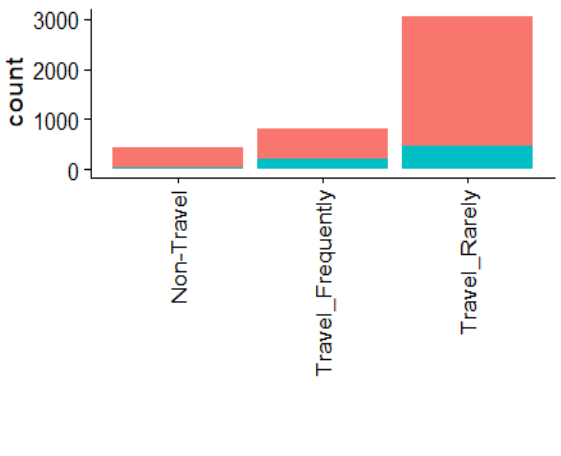
We have data of more than 4410 employees of a company XYZ

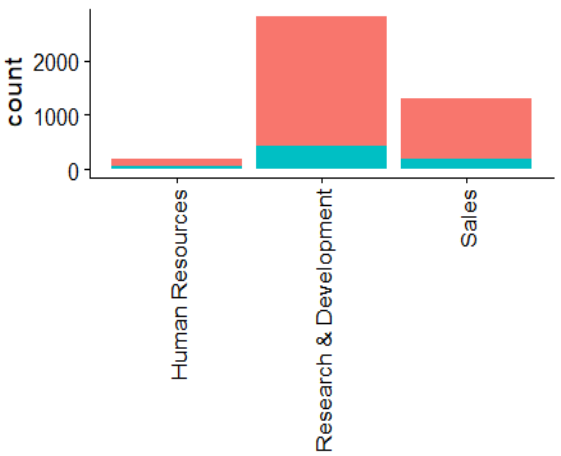
- General Data
- Employee In time & Out time
- Manager Survey Data
- Employee Survey Data

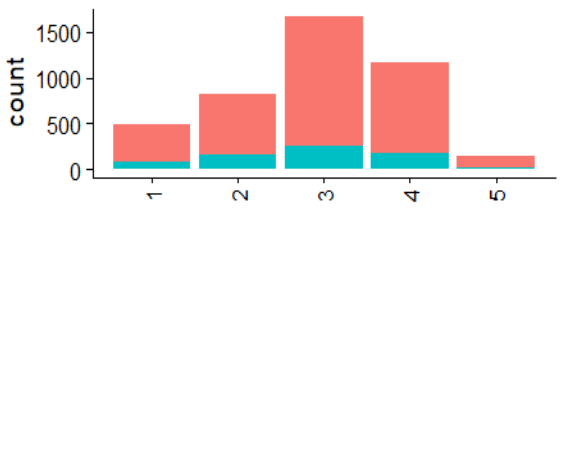
Data Cleaning & EDA

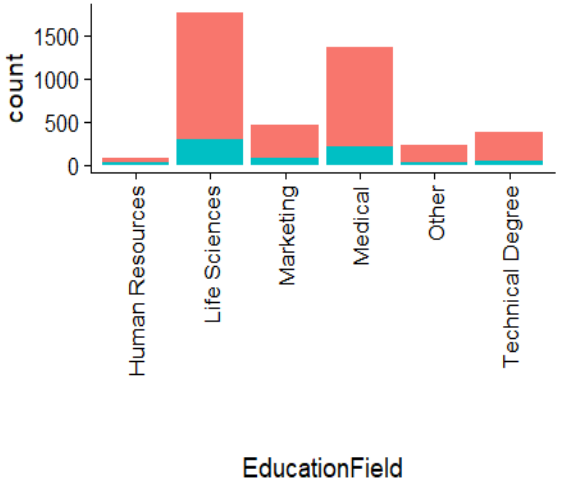
- Checked duplicated rows – No duplicate records found
- Removed all those columns where
  - All values are NA(2.5 %) - This will also not impact the analysis
- Column naming
- Removed unnecessary columns
- Dummy variables creation, Standardized Numeric values
- Merging data set to create final data frame
- Exploratory Data Analysis
  - Derived metrics – Average working hours, Employee leaves
  - Graphical Plots
- **This reduced data set to 4300 observation and 70 variables**

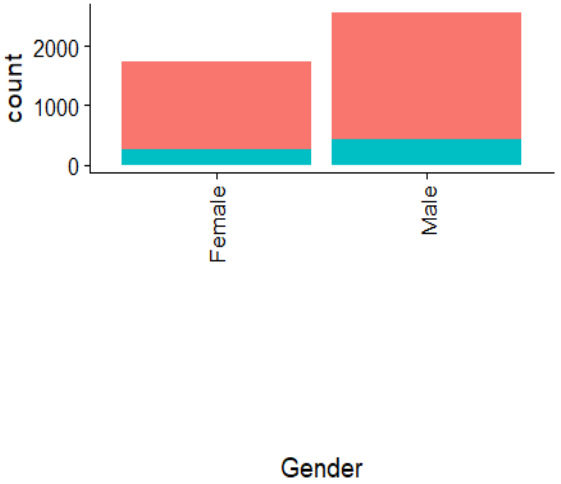
Preliminary Data Analysis





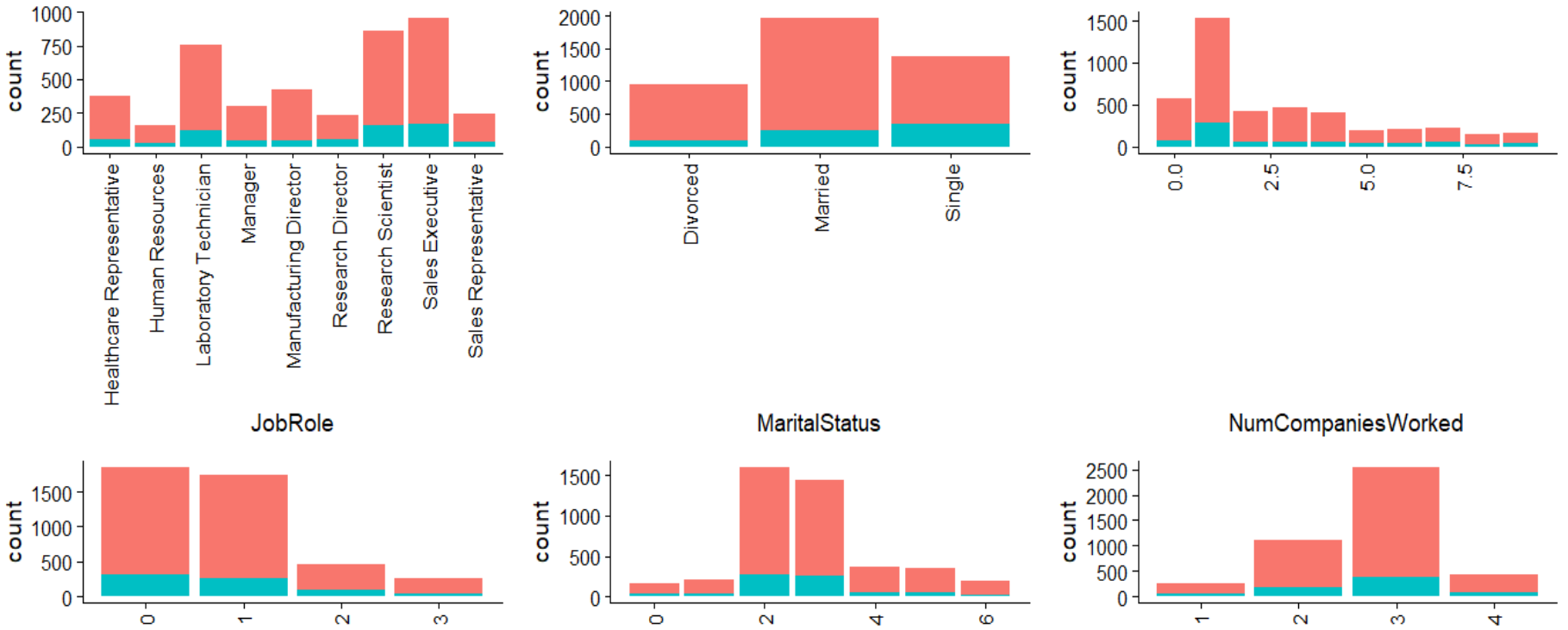






Plots showing Attrition for
- Business Travel
- Department
- Education
- Education Field
- Gender
- Job Level

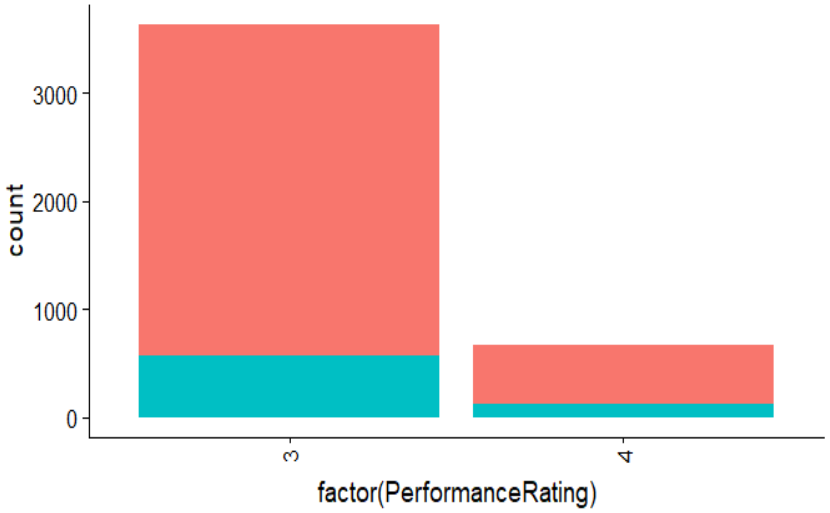
Preliminary Data Analysis

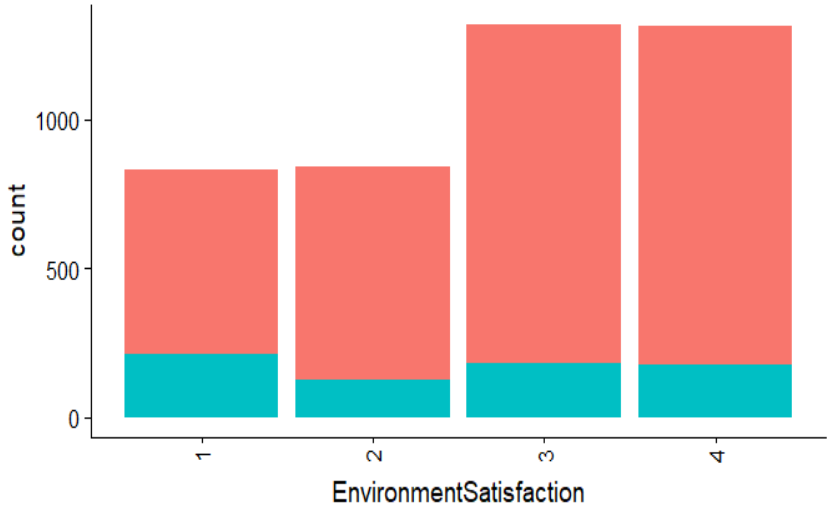


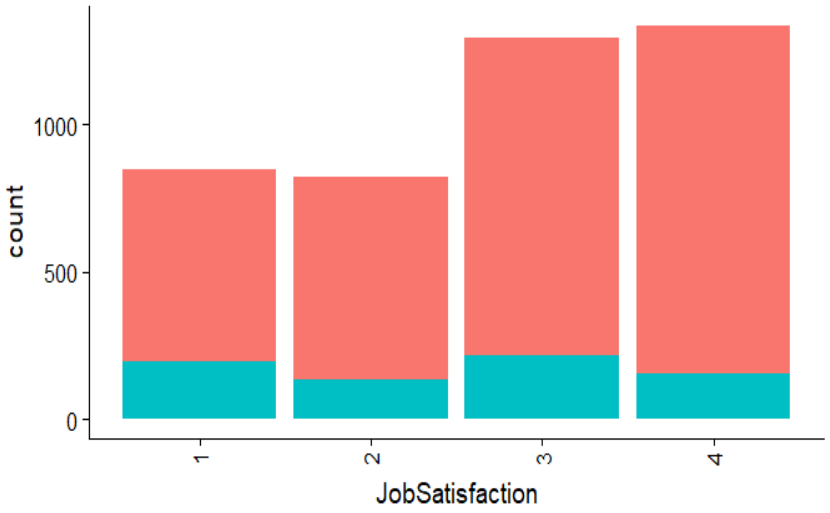
Plots showing Attrition for

- Job Role
- Marital Status
- NumCompaniesWorked
- StockOptionLevel
- TrainingTImesLastYear
- Job Involvement Level

Preliminary Data Analysis



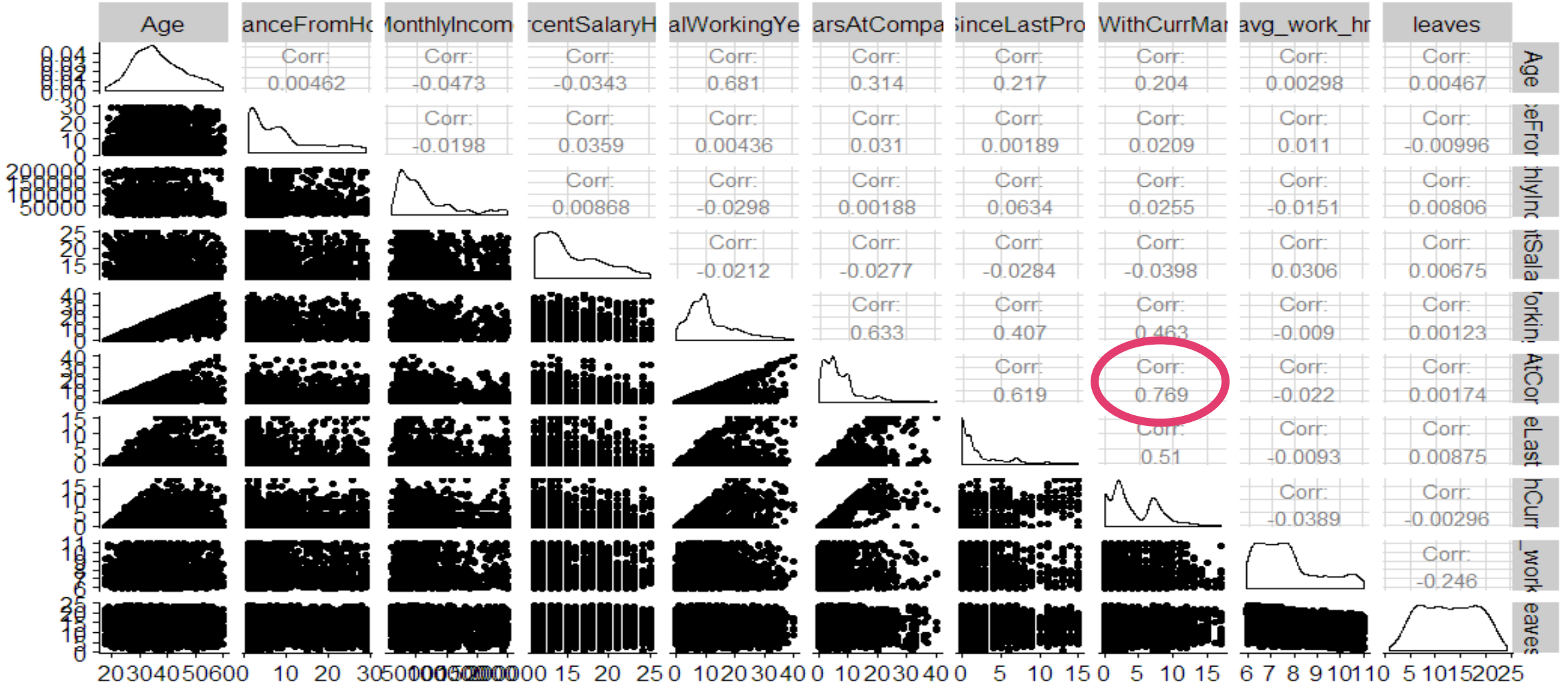




Plots showing Attrition for
- Performance Rating
- Environment Satisfaction
- Job Satisfaction
- Work Life Balance

Preliminary Data Analysis

Graph showing Correlation between Numeric Variables



Model Preparation

- **Ran model using all variable's on Training Data set (70%)**
- **Removed insignificant variables using StepAIC**
- **Improved model iteratively by**
  - **Removing multi-collinearity using VIF values**
  - **Choosing most significant variables using P-values**
- **Selected final model with all significant variables**
  - Age, YearsSinceLastPromotion , YearsWithCurrManager , avg_work_hr , NumCompaniesWorked.x5 , NumCompaniesWorked.x7, NumCompaniesWorked.x9, EnvironmentSatisfaction.x2 , EnvironmentSatisfaction.x3 , EnvironmentSatisfaction.x4 , MaritalStatus.xSingle

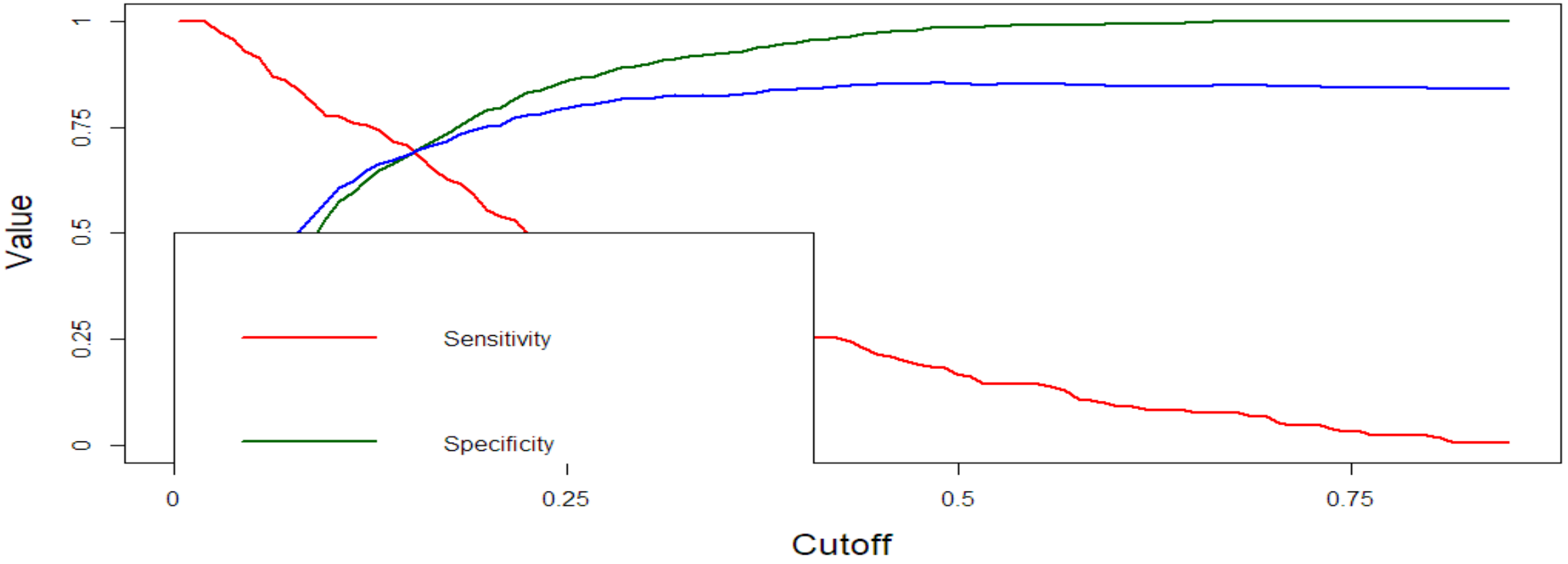
Model Evaluation:

- Created Prediction based on final model on Test Data Set
- Checked Accuracy, Sensitivity and Specificity using different values of cutoff
- Selected final cutoff having difference of sensitivity and specificity as 0.025

ACTUAL(Attrition)	PREDICTED(Attrition)	
	NO	YES
NO	758	323
YES	67	142

Accuracy	70%
Attrition(Sensitivity)	67.94%
Non-Attrition(Specificity)	70.12%

Model Evaluation : Cutoff graph



Summary of observations

- Observations and Proposed changes based on the model observations
  - Younger and single employees, being prone to attrition, can be monitored for possible attrition
  - Change of manager, in case of long association , can be considered
  - Delayed promotion is contributing to attrition. Same needs to be investigated.
  - Average working hours should be monitored as more working hours is contributing to attrition
  - A person having multiple job changes (more than 5) are prone to attrition. Hence same needs to be considered during recruitment
  - People with lowest level of environment satisfaction are prone to attrition. Measures may be taken to improve work environment