HR Analytics Case Study

Group Members:

1. *Kabilan Karunakaran*
2. *Jaya Priya Radhakrishnan*
3. *Sivaprasath CM*
4. *John Manohar*

Abstract

A large 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** is bad for the company's reputation, completion of current live projects and resource management for recruiting new talent.

Business Objective:

To find the driving factors for employee churn and focus on the improvement of the driving factors to curb the employee churn rate, i.e. the variables which are strong indicators of employee churn.

Data

The dataset contains details of all employees of the company. The data is captured in five different files such as

Employee survey data	4410 obs. of 4 variables
General data	4410 obs. of 24 variables
In time	4410 obs. of 262 variables
Out time	4410 obs. of 262 variables
Manager survey data	4410 obs. of 3 variables

Analysis Steps

The analysis has following Steps:

- Data Sourcing
- EDA
  - Handling NA
  - Univariate Analysis
  - Bivariate Analysis
- Data Preparation
- Model building
- Model Evaluation

Data Sourcing

- The data collected in various files are merged to proceed for the data selection. As part of the data selection step, first manager survey data, employee survey data and general data are merged together.

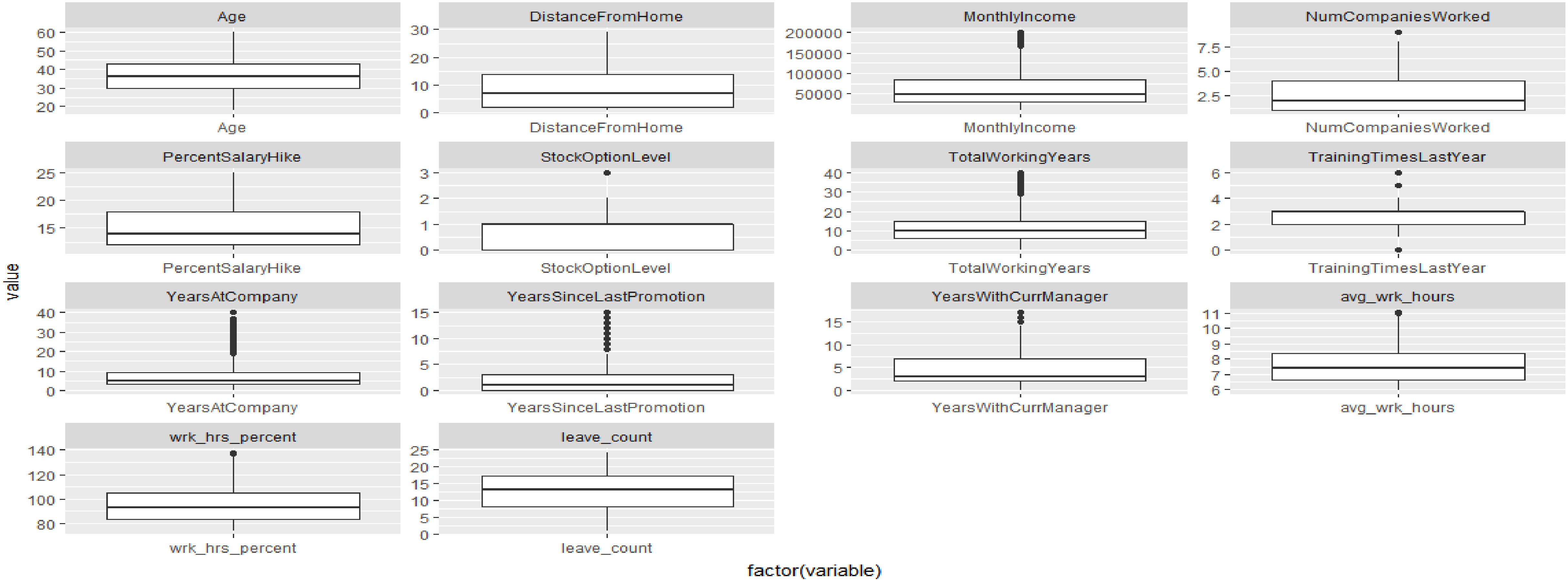
- From the “in_time” and “out_time” data, derived columns such as **average working hours, leave count and leaving from the office** are merged to the master data frame.

EDA

- **Handling NA :**
  - It was observed that certain columns in the master data frame were having NA values. As the percentage of NA values are significantly less(2.5%) , the rows having NA values are removed for the analysis .
- **Univariate and Bivariate analysis were also done as part of EDA and few insights were observed :**
  - Univariate analysis on categorical variables done using bar plot to understand the distribution and on numerical done using histogram and box plot.
  - Bivariate on categorical data was done using stacked bar plot to know the percentage of attrition in each level of categorical attributes.

EDA : Univariate Analysis

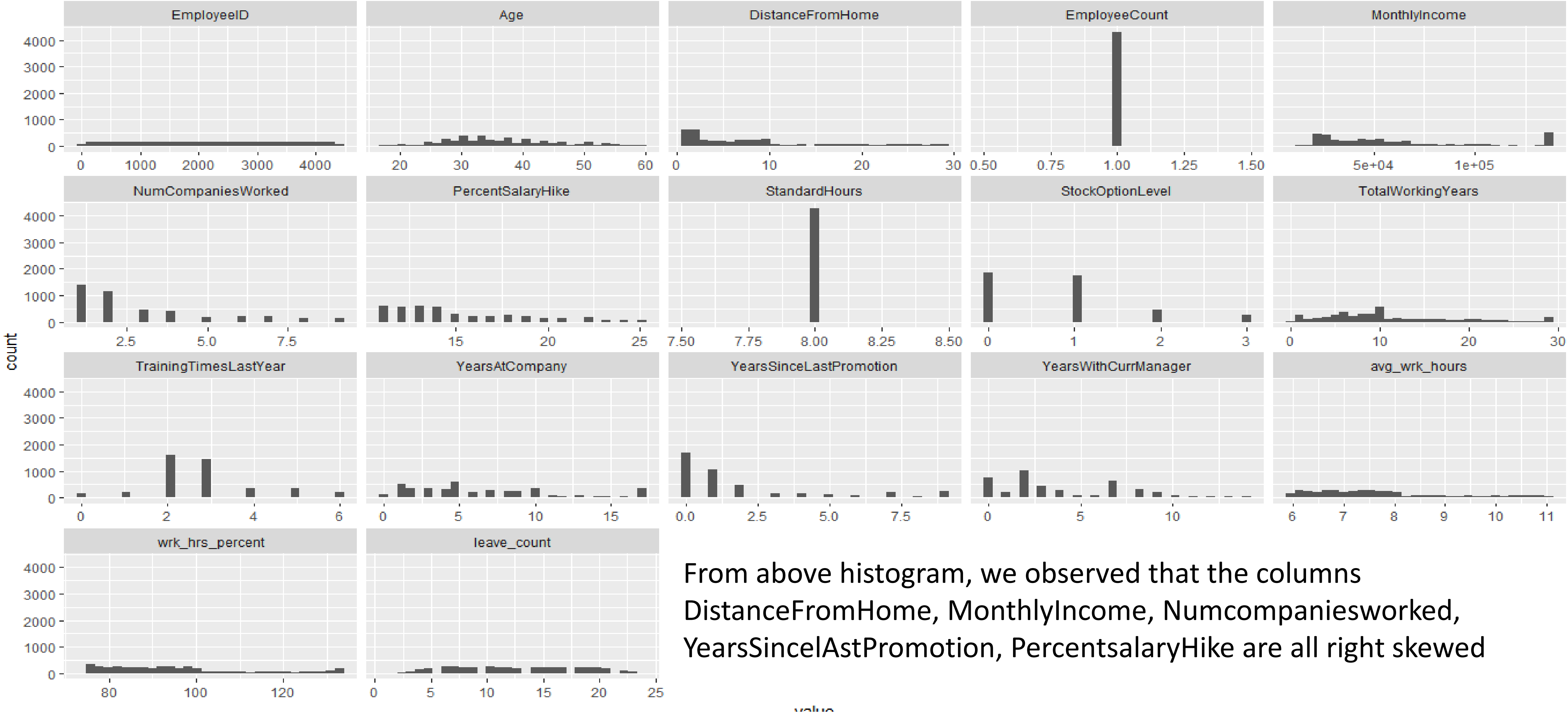
The univariate analysis were performed for all categorical and numerical columns which we considered for analysis . And few insights which are found are listed below :



From above plots, we observed that, there exists outliers in columns YearsAtcompany, YearsSincelastpromotion, avg_wrk_hours,trainingTimesLastYear, StockOptionLevel, TotalWorkingyears.

EDA : Univariate Analysis

And few more insights which are found are listed below :



From above histogram, we observed that the columns DistanceFromHome, MonthlyIncome, Numcompaniesworked, YearsSincelAstPromotion, PercentsalaryHike are all right skewed

EDA : Bivariate Analysis

The Bivariate analysis was performed for the variables and few insights were found that are listed below :

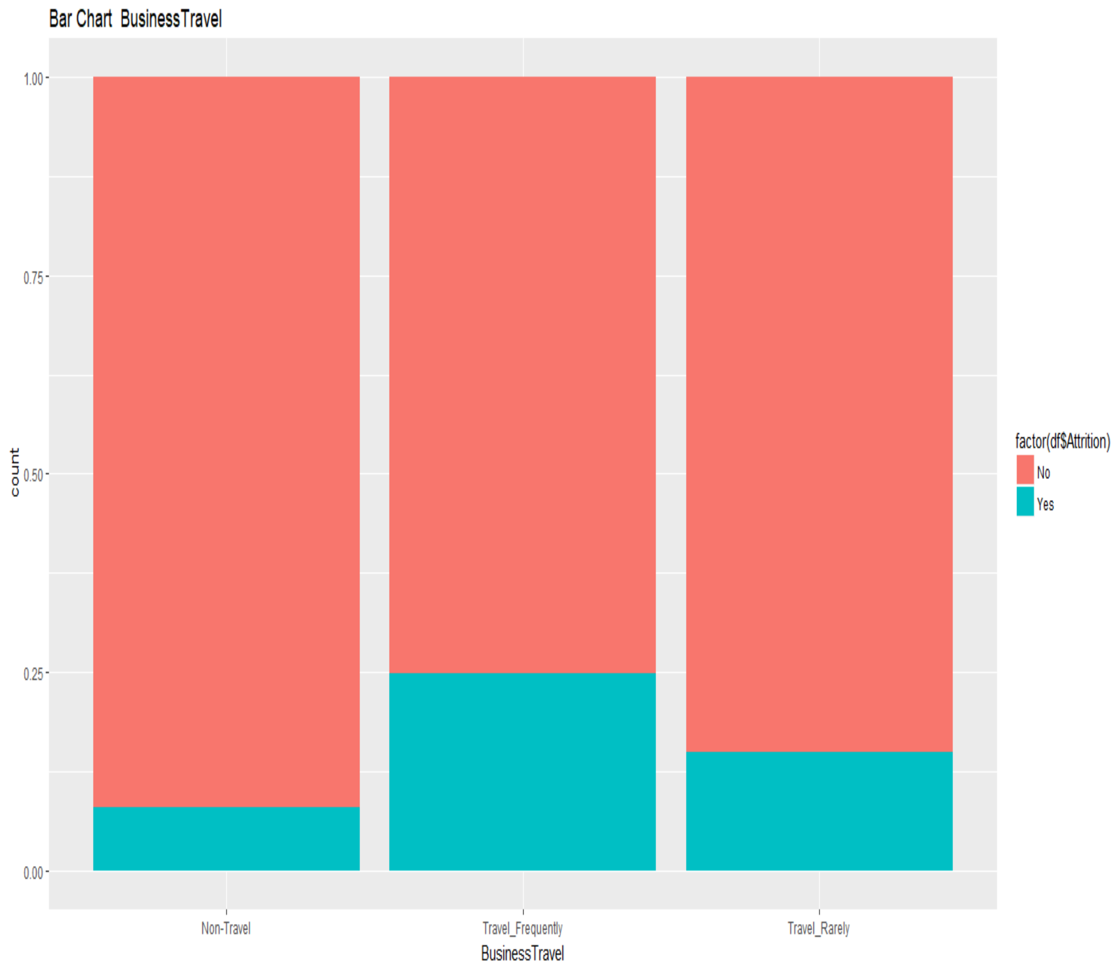
- Attrition rate is 16%
- Max attrition percent on category travel_frequently on BusinessTravel
- Max attrition percent occurs in HR department
- Max attrition percent occurs in HR Education Field
- Max attrition percent occurs in Research Director JobRole
- Max attrition percent occurs in Single Marital Status
- Max attrition percent occurs in employee group who gave Evironmentsstisfaction score 1
- Max attrition percent occurs in employee group who gave Jobsatisfaction score 1
- Max attrition percent occurs in employee group who gave Worklifebalance score 1
- Max attrition percent occurs in employee group who gave Jobinvolvement score 1
- Max attrition percent occurs in employee group who got performanceRating as 4

The supporting plots are shown in further slides.

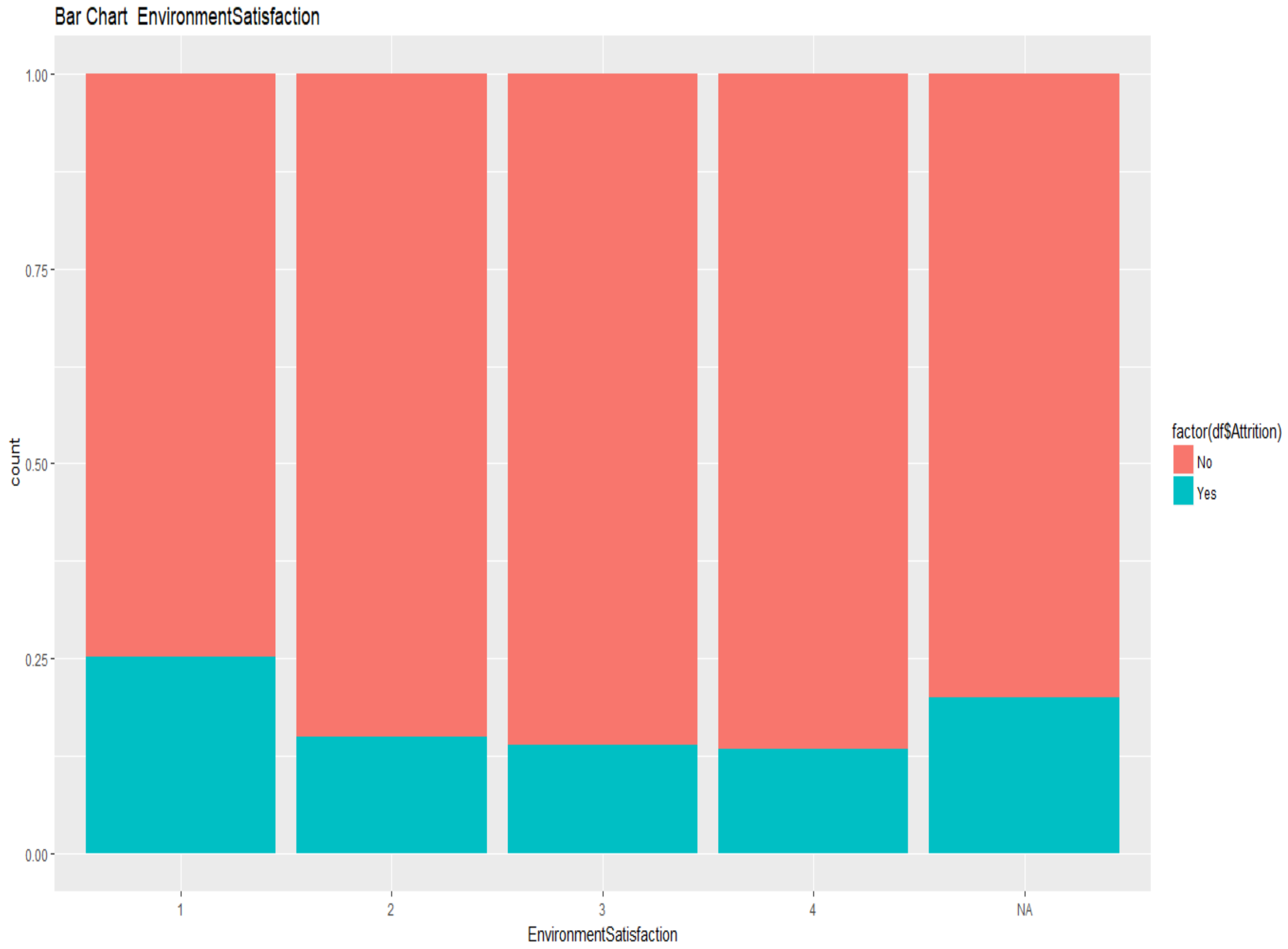
EDA : Bivariate Analysis

The Bivariate analysis was performed for the variables and supporting plots for few variables are listed below:

Business Travel



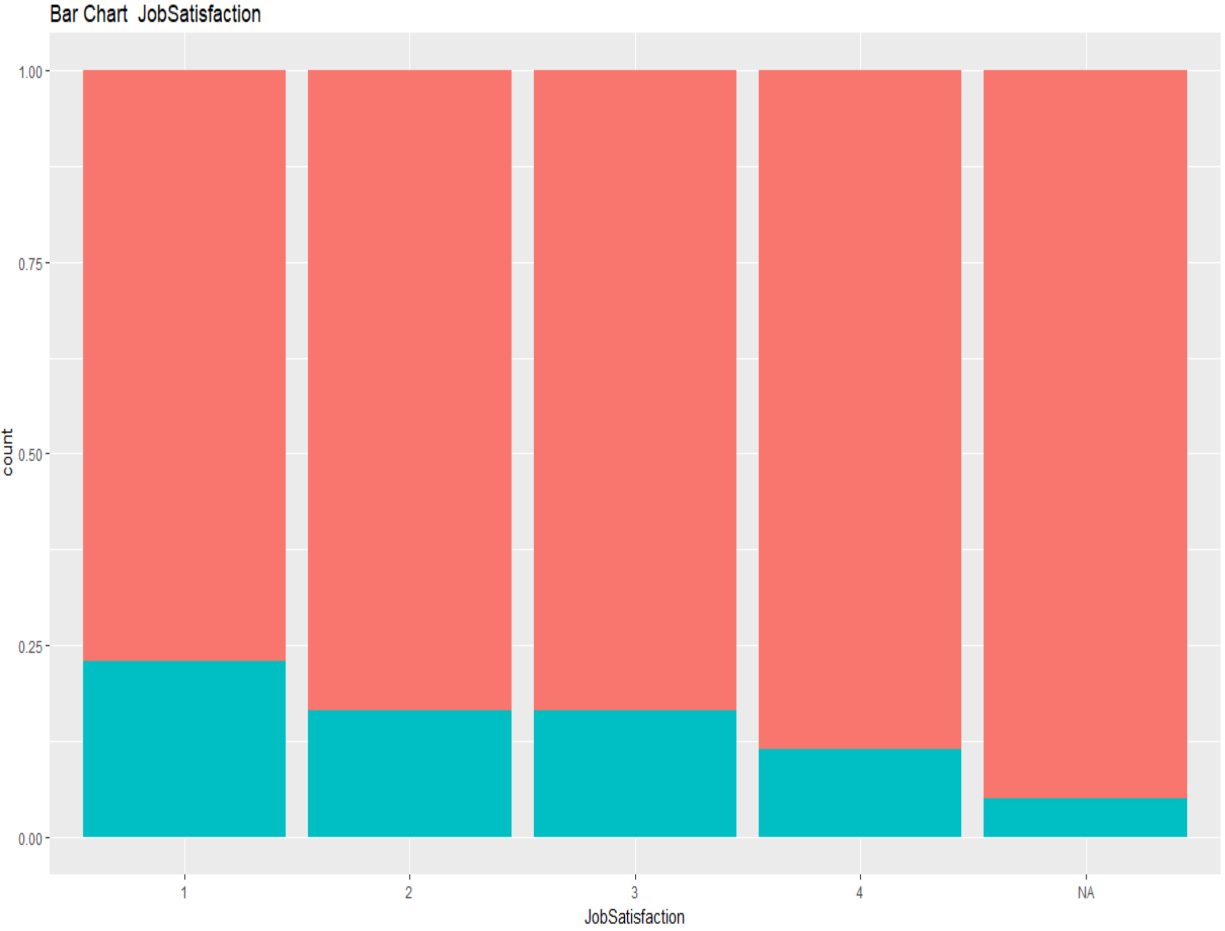
Environment Satisfaction



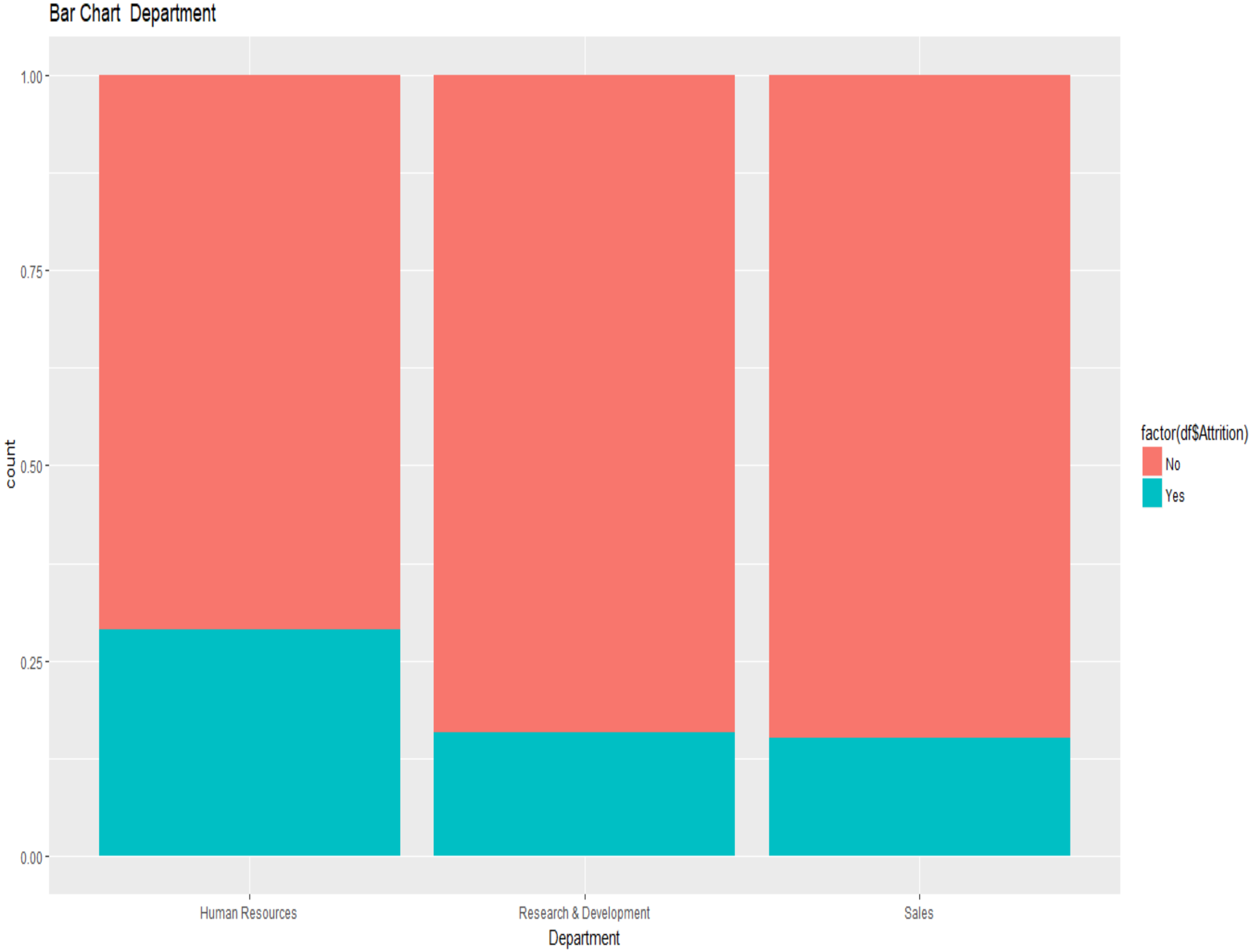
EDA : Bivariate Analysis

It cane observed from below plots, that higher percent attrition happen in job satisfaction level 1 category and in case of department , it happens in HR department

Job Satisfaction



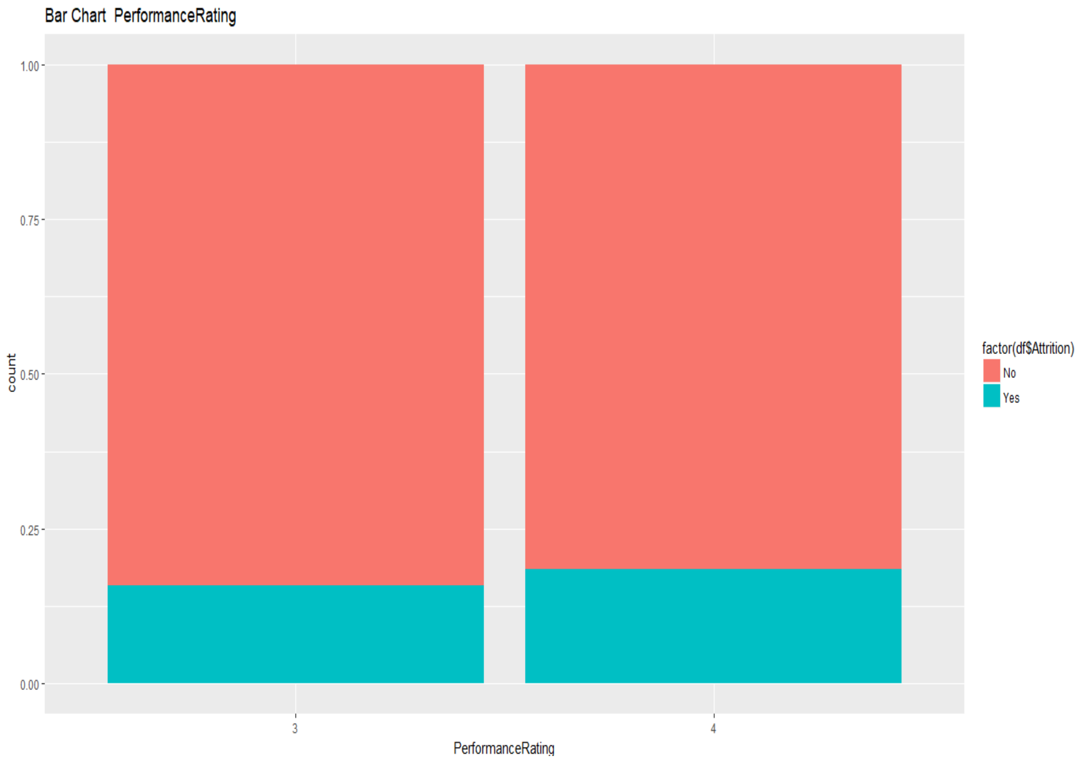
Department



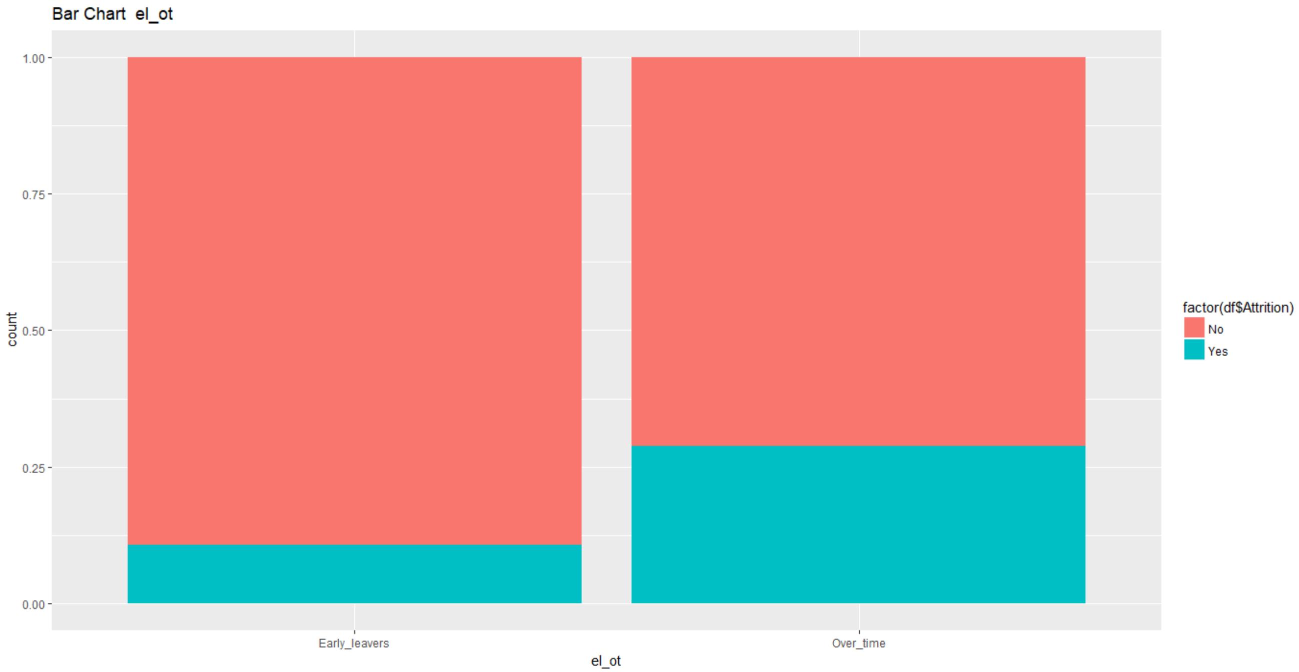
EDA : Bivariate Analysis

From below plots, its evident that higher percent of attrition happens in employee group who stays longer hours in office and in case of performance rating, higher attrition percent is with employee who received rating 4

Performance Rating



el_ot (Early_leavers/Over_Time)



Data Preparation

- Outliers identified using Boxplots were removed with help of quantile function and cupping the vales where the sudden jump occurs in the data .
- Feature standardisation were performed using scale method on numeric values
- Dummies were created for categorical values
- Finally all the required attributes were merged into a single data frame named “master_final ”

Model Building

- **Data Split:** Dataset was split into two : train and test , where train set hold 70 percent od data and test holds 30 percent of data
- **Initial Model :** First model named model_1 was created by including all independent variables as predictors
- **StepAIC :** Using stepAIC, insignificant variables were excluded and the model_2 is obtained

- Based on **VIF factors**, the attributes EducationFieldLife.Sciences , BusinessTravelTravel_Rarely YearsAtCompany were removed as VIF was high.

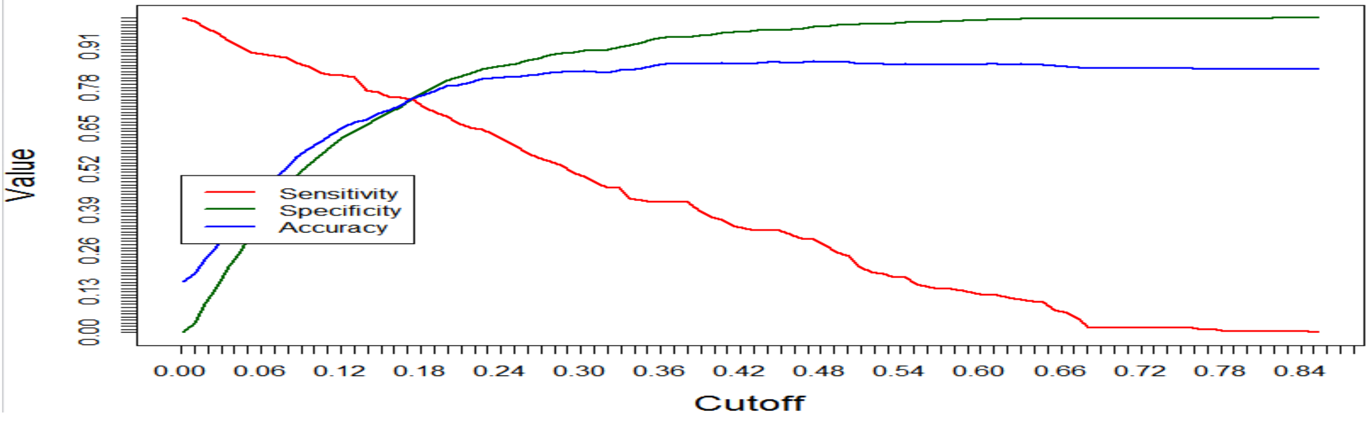
- And the other attributes which were found to be insignificant **based on p-values** were removed at the end of several iterations and obtained model named model_19 as final model

- The final_model includes totally **10 unique predictors** and **AIC value is 2207.2**

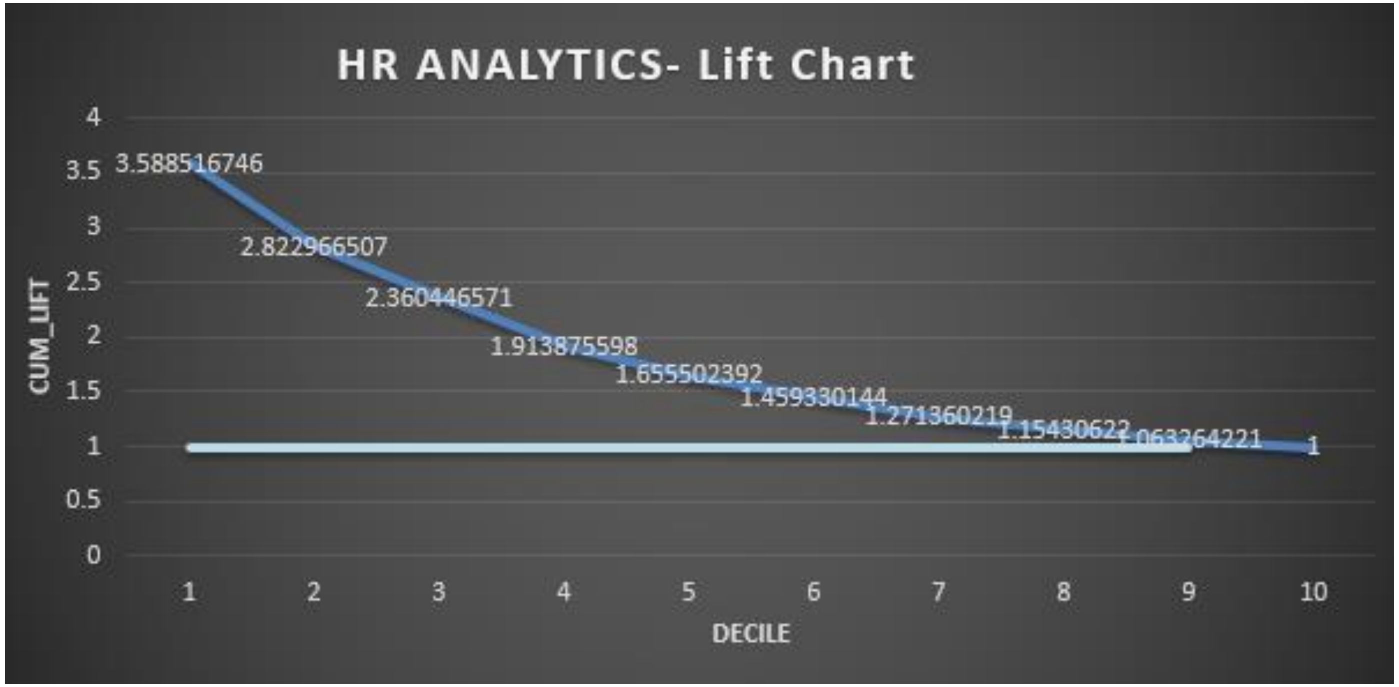
- The model accuracy achieved is **86.04 %** by assuming **cut-off as 0.5**

Model Evaluation

- **Optimal Cut-off :** As the randomly picked cut-off 0.5 may not be the optimal one. The optimal cut-off is found from intersection point of line plots of sensitivity, specificity and accuracy. It is found to be 0.17 in our case . And on using this optimal cut-off, the accuracy achieved is 74%

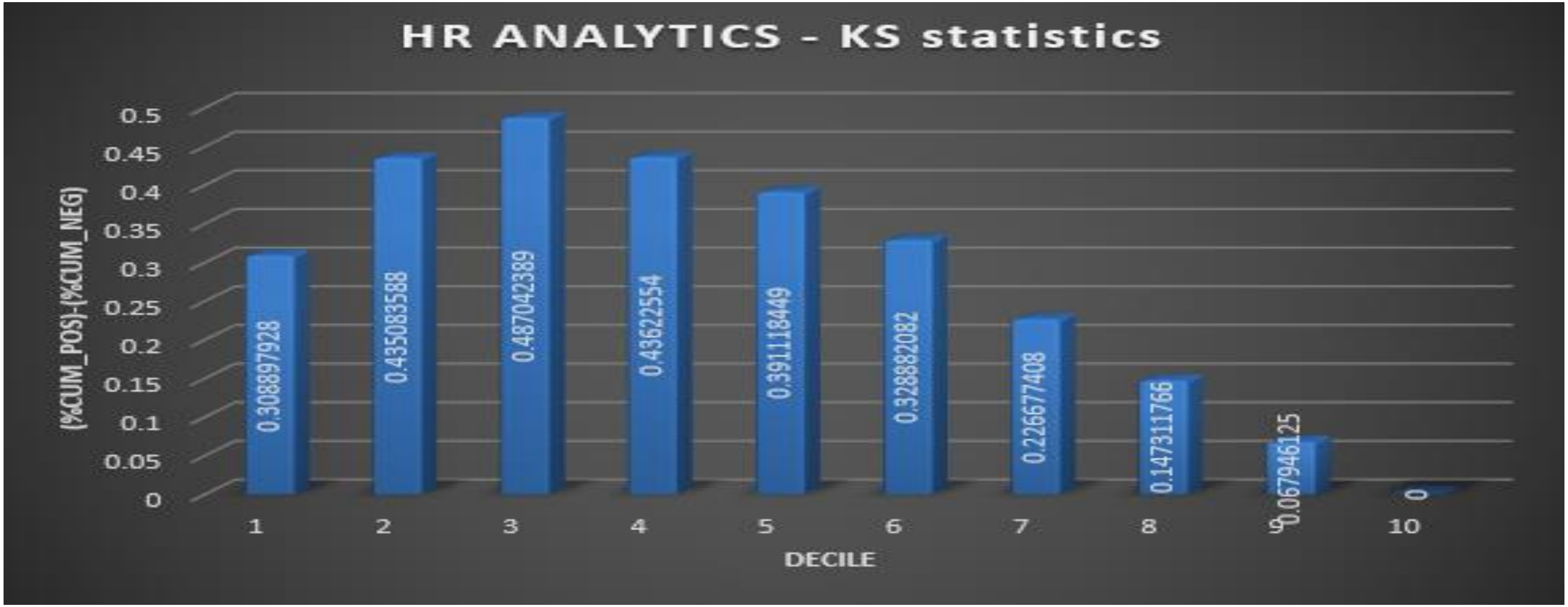


- **Lift Chart :** The plot represents the cumulative lift chart obtained for the model .



Model Evaluation

- **KS-Statistics :** The Ks-statistics of the model is found to be 48% and it is greater than 40% and it also resides in 3rd decile as observed in the below plot .



Conclusion

From the final model, its is found that the key driving factors of Attrition of employee of The XYZ company are :

- Age
- NumCompaniesWorked
- TotalWorkingYears
- YearsSinceLastPromotion
- YearsWithCurrManager
- el_ot (Early_leavers/Over_Time)
- BusinessTravelTravel
- EnvironmentSatisfaction
- JobRole
- JobSatisfaction

The final model has accuracy of ~ 74% on using the optimal cut-off value as 0.17 . And the KS-statistics is found to be 48% which lies in the 3rd decile .

Thank You