HR Analytics Case Study

26-Aug-2018

Manjula Raman

Srinivas Panganamala

Greenu Sharma

Vishnu Das

Problem Statement

- ❑XYZ Company needs to focus on curbing Attrition
- ❑ HR Analytics needs to
  - Model attrition based on all available data and surveys
  - Derive factors to focus on for curbing the attrition
  - Changes to workplace to make employees stay

Data Preparation

- ❑Checking for duplicates: There are no duplicates in the databases
- ❑The column name EmployeeID was missing from the in_time and out_time databases
- ❑Missing values imputation
  - gen_data: 19 NA values in the NumCompaniesWorked . We keep it as 1 as that's the default value
  - gen_data:9 NA values in TotalWorkingYears.We keep equal to YearsAtCompany
  - emp_survey: 25 NA values in EnvironmentSatisfaction. Assign the most common rating 3 to it
  - emp_survey: 20 NA values in JobSatisfaction. Assign the most common rating 4 to it
  - emp_survey:38 NA values in WorkLifeBalance Assign the most common rating 3 to it

Data Preparation

- Missing Value imputation in in_time and out_time
  - The databases have time stamps corresponding to in and out time for 261 days for 4410 employees
  - There are days where all entries are NA. These correspond to holidays; Removing 12 columns that are completely NA
- **Derived columns** to count the number of NAs in the in_time and out_time
  - The number of NAs in the in_time and out_time match
  - Only 1 columns required to count number of NAs – **leaves_taken**

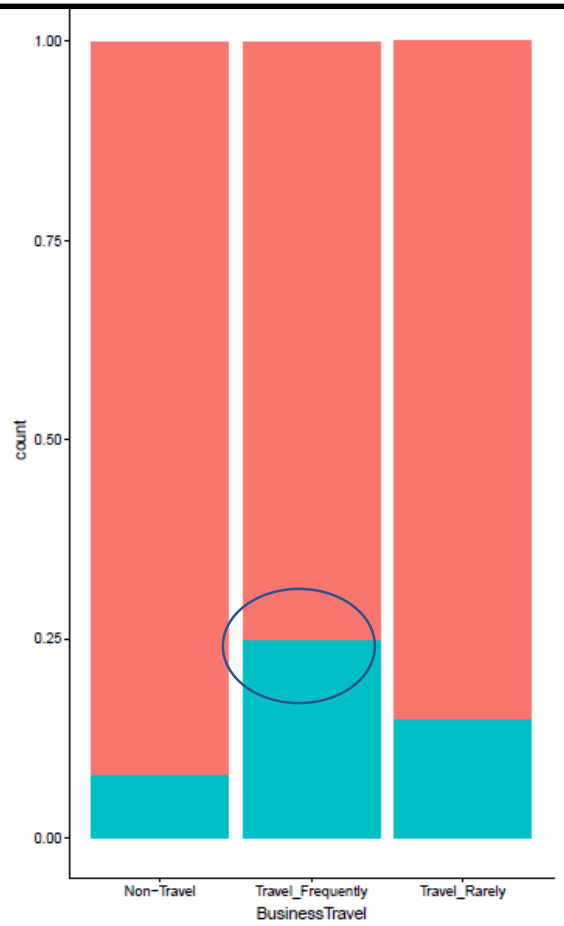
Data Preparation

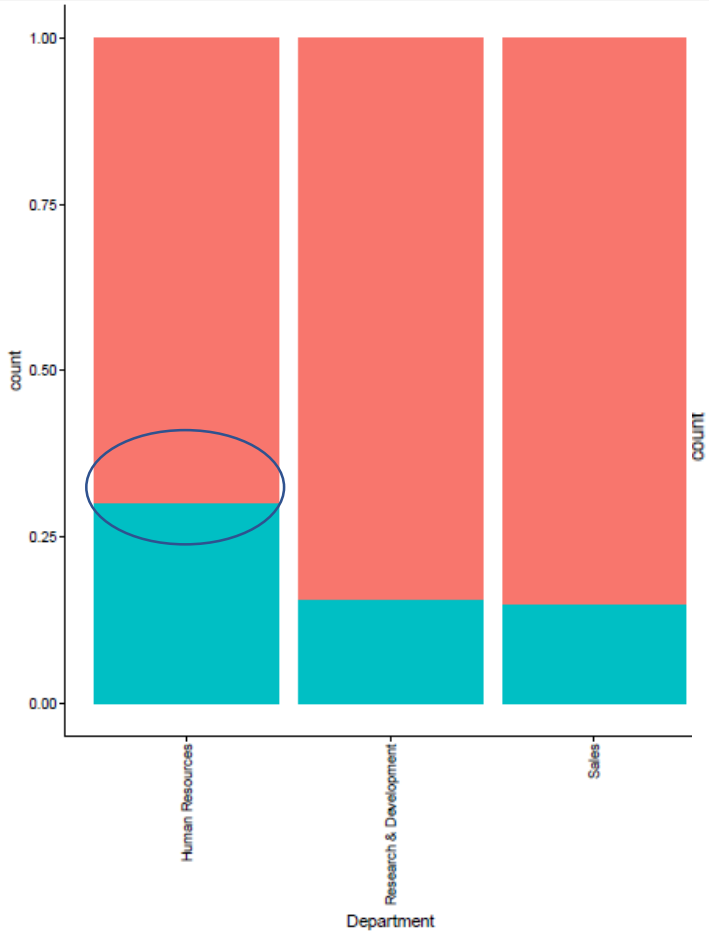
- Derived Metric : mean_timeatwork
  - After all missing values in the timestamp are imputed, the time stamp is converted to time in %H:%M:%Y format
  - For each employee, the difference in the in_time and out_time is calculated
  - The mean across all the working days is calculated for each employee
- Merging all the data into master_frame using EmployeeID as the common field
- Redundant Column Removal: All entries are Same or Column not needed for modelling
  - Over18- all Yes;StandardHours-al l8;EmployeeCount-all 1
  - EmployeeID: not required for modelling

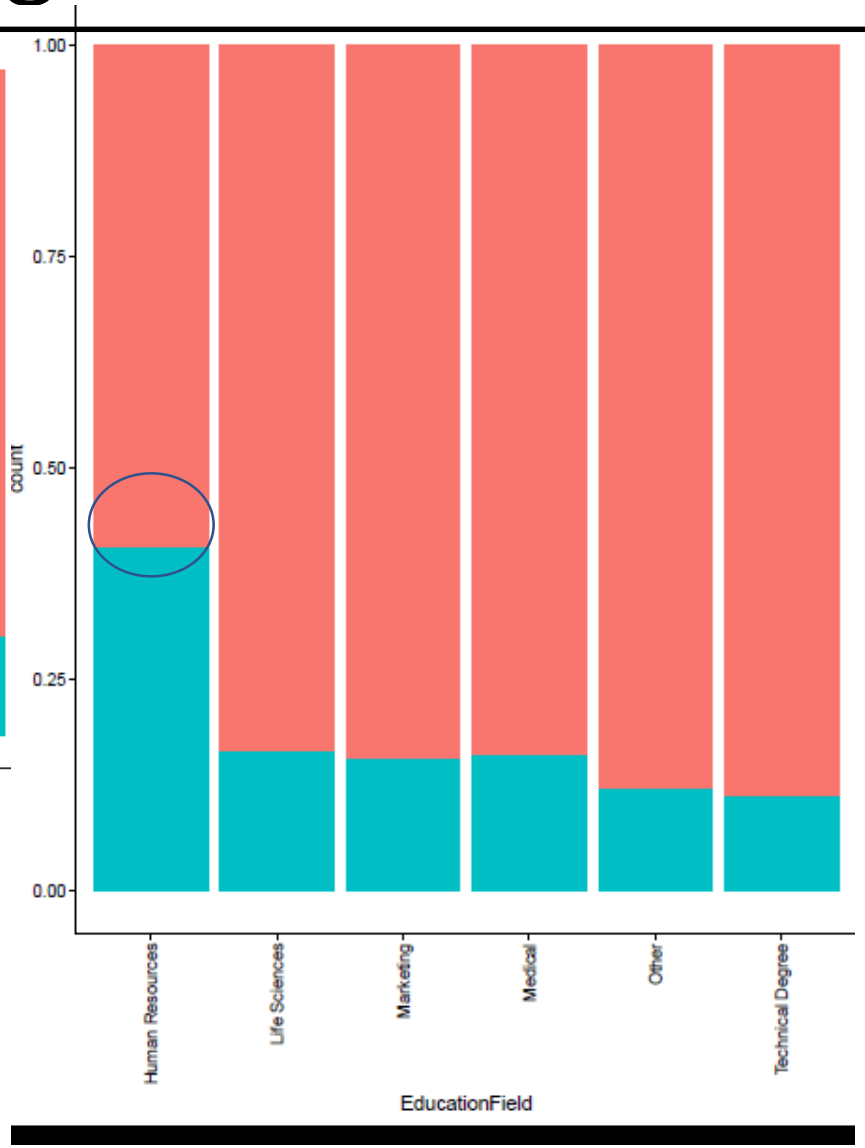
Exploratory Data Analysis(EDA)

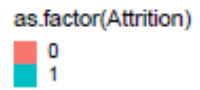
- Numeric Variables : 18 variables; All variables that have an inherent order in them like Ratings, Education etc are taken as numeric
  - Age,DistanceFromHome,Education,MonthlyIncome,NumCompaniesWorked,PercentSalaryHike,TotalWorkingYears,TrainingTimesLastYear,YearsAtCompany,YearsSinceLastPromotion,YearsWithCurrManager,mean_timeatwork,JobInvolvement,PerformanceRating,EnvironmentSatisfaction,JobSatisfaction,WorkLifeBalance,leaves_taken
- Categorical Variables: 8 variables
  - BusinessTravel,Department,EducationField,Gender,JobLevel,JobRole,MaritalStatus,StockOptionLevel

EDA: Bar Charts for Categoricals





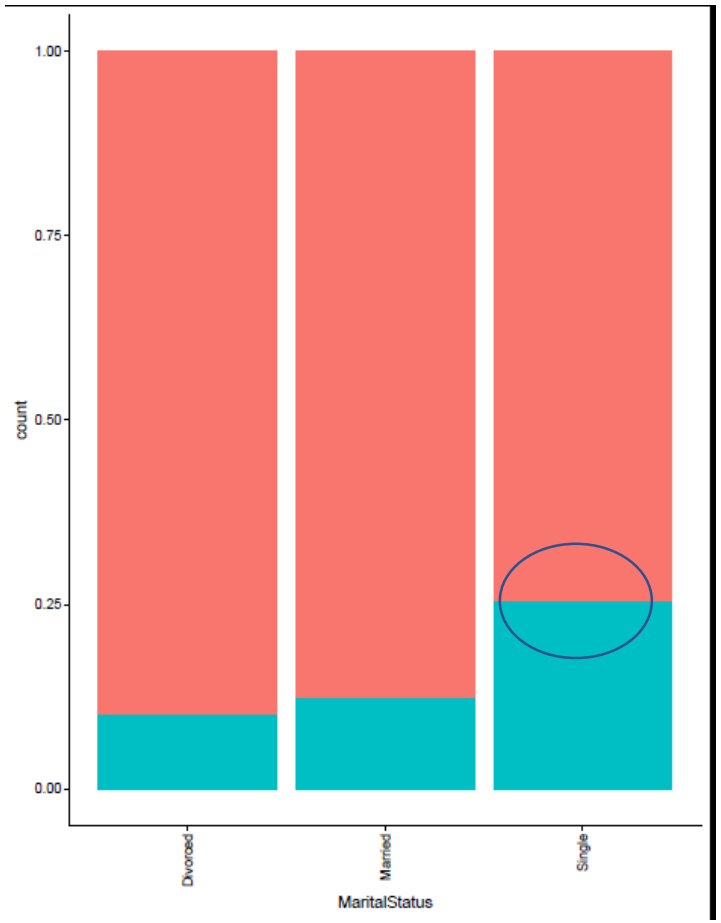


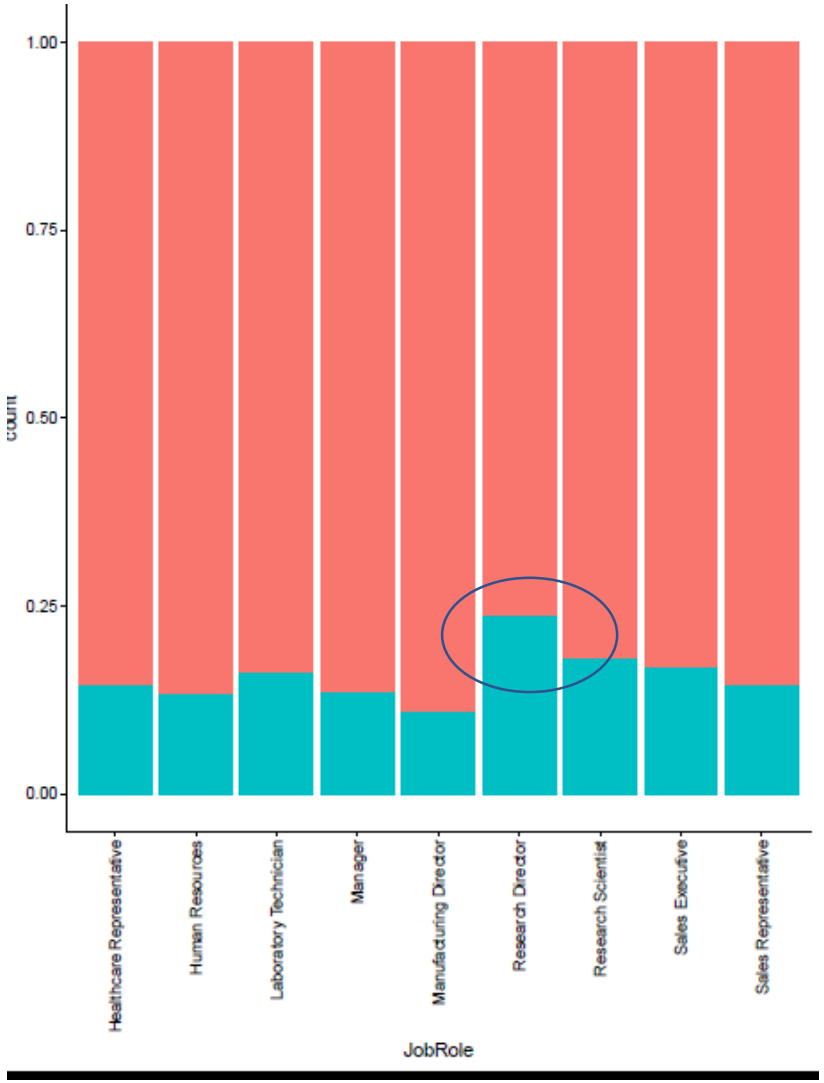


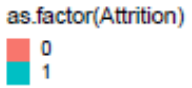
0- No Attrition
1- Yes Attrition

✓ The highlighted bars have more significant Attrition

EDA: Bar Charts for Categorical



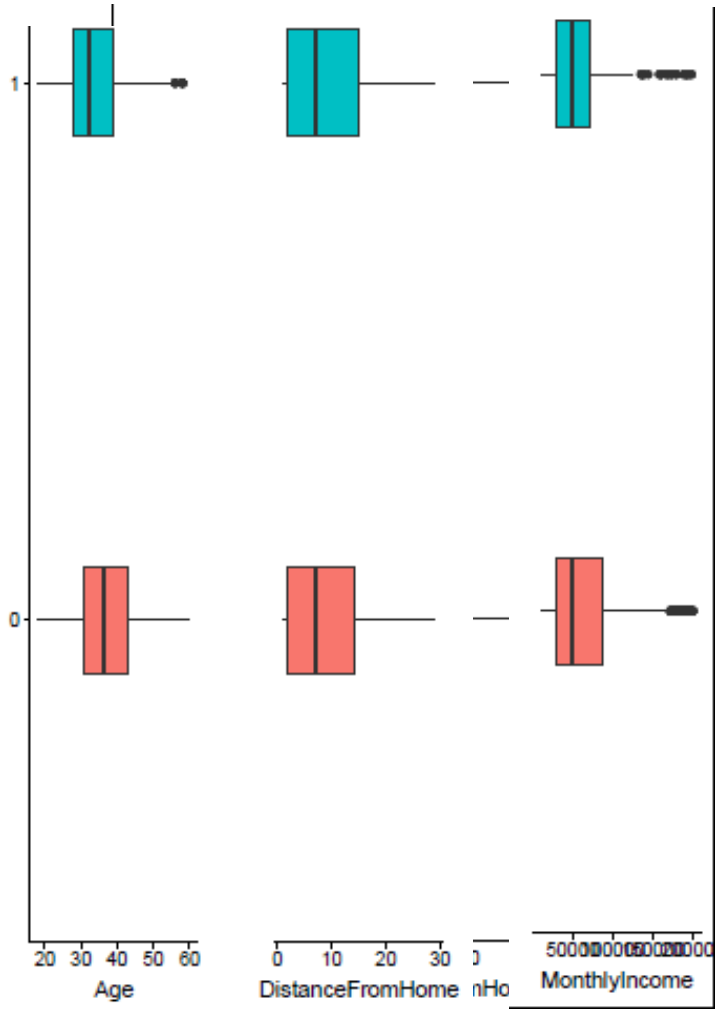




0- No Attrition
1- Yes Attrition

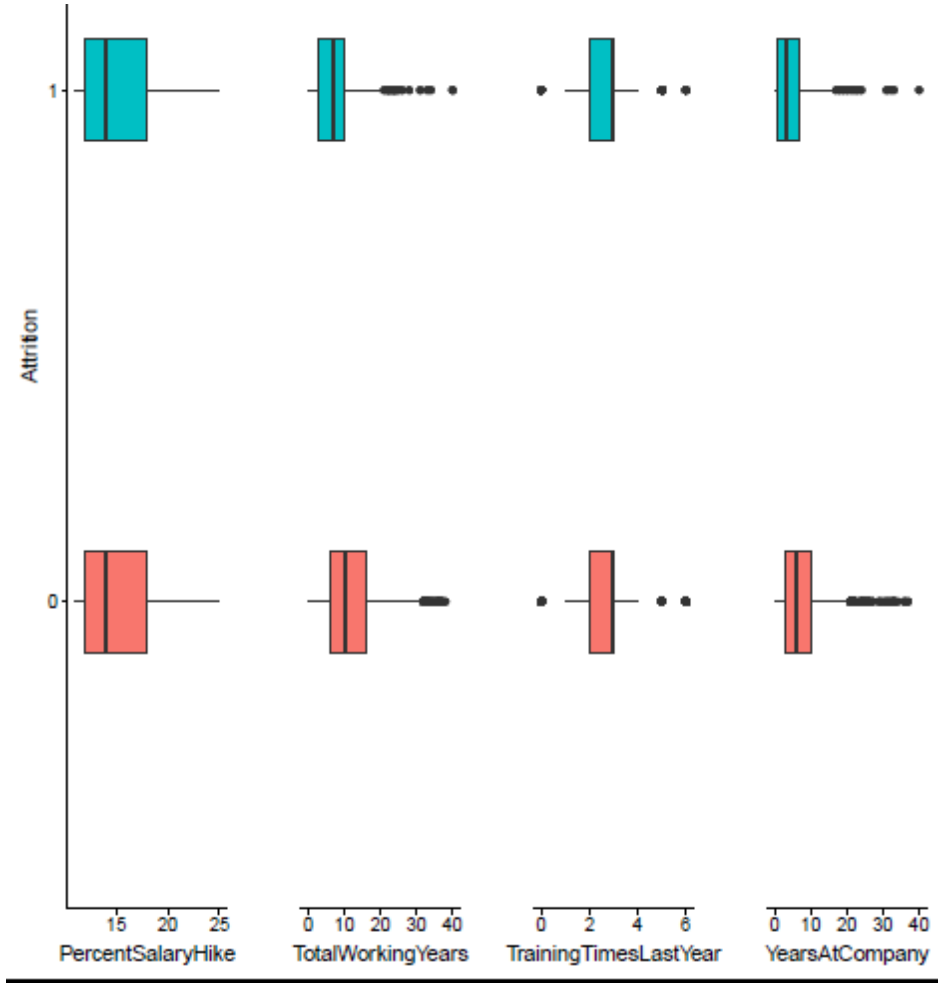
- ✓ The highlighted bars have more significant Attrition
- ✓ StockOptionLevel, JobLevel and Gender do not show any significant variations with respect to Attrition

EDA Box Plots For Numericals

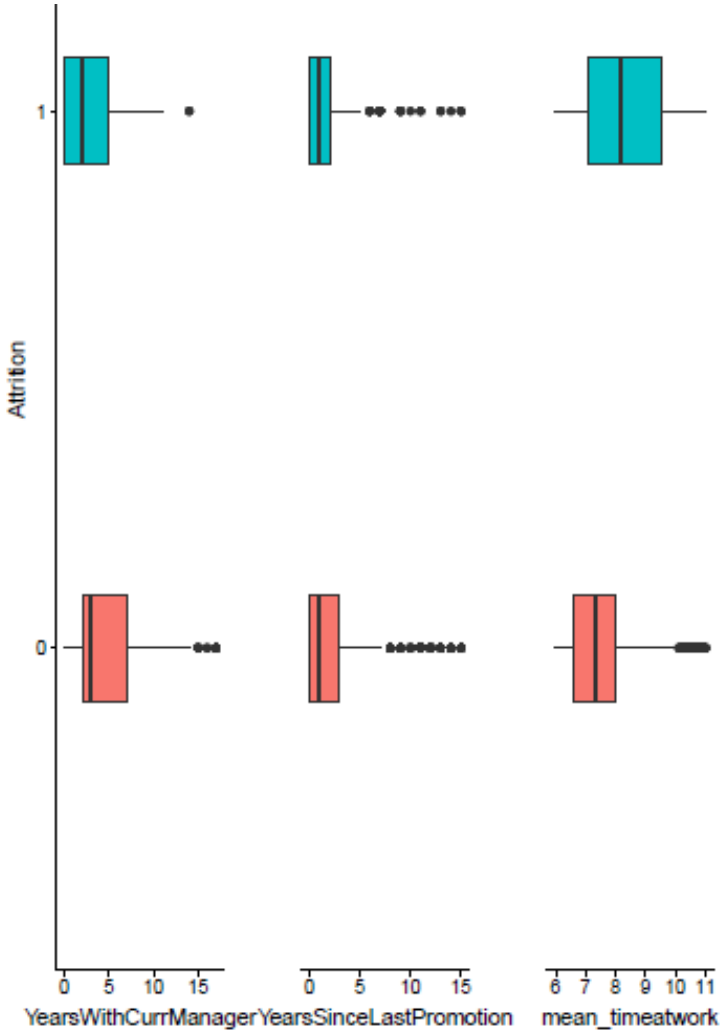


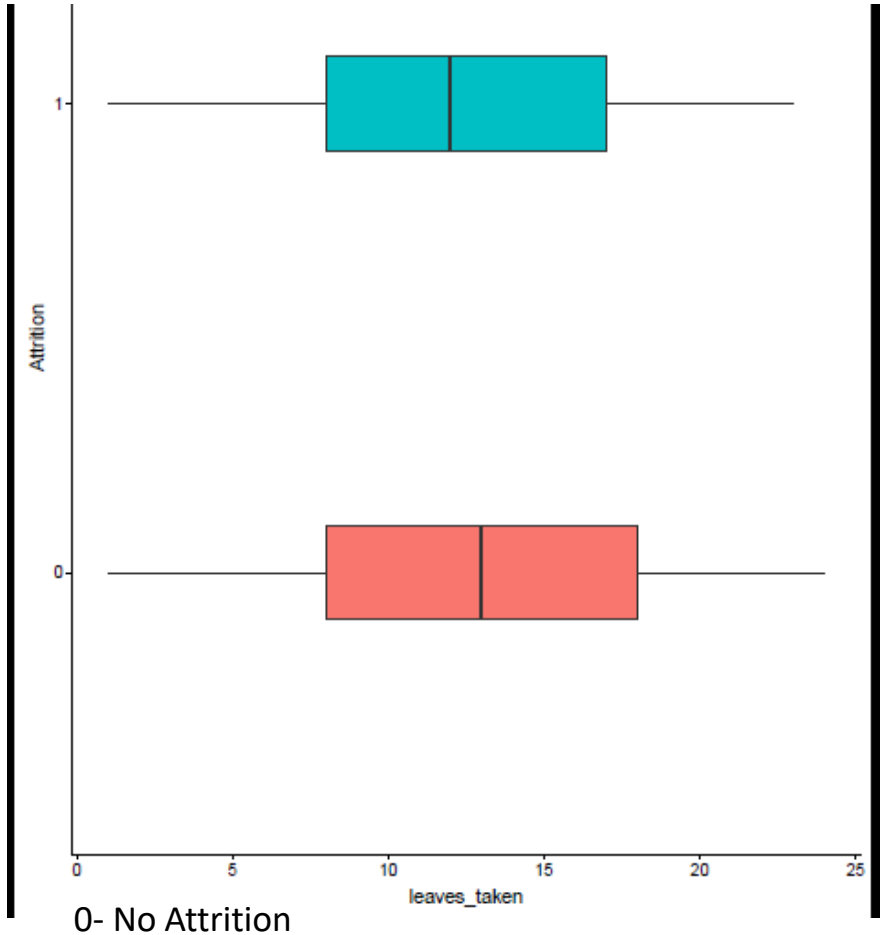
0- No Attrition
1- Yes Attrition

- ✓ Age , shows a higher median value for 0 attrition; Older people tend to stay
- ✓ Monthly Income upper quartile higher for 0 Attrition; people with higher income stay
- ✓ Total Working years spread is higher; indicating that more experienced people have lower attrition



EDA Boxplots For Numericals





0- No Attrition
1- Yes Attrition

- ✓ YearsWithCurrManager has higher spread and upper quartile; longer with current manager tend to stay
- ✓ mean_timeatwork has a higher median(greater than StandardWorkHours) and spread is higher; Attrition high
- ✓ Leaves taken: People who are leaving are taking fewer leaves; similar to mean_timeatwork trend

Data Preparation For Modelling

- Converted target variable Attrition to 0 and 1 for No and Yes
- Outlier Analysis and scaling done for Numeric
- Dummy Variables created for Categoricals
- Data Divided randomly into ‘test’ and ‘train’:30% test;70% train

Logistic Regression

- ❑Initial Model was created using ‘glm’ function
  - 45 varaibles, AIC:2165.2
- ❑stepAIC was used to create reduced model
  - 30 variables,AIC:2144.6
- ❑Multicollinearity removal: YearsAtCompany, MaritalStatusMarried removed
- ❑Less significant variables removed iteratively
- ❑model_17 with 15 variables with all (***) significance generated

Final Model

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.43280    0.33274  -7.311 2.64e-13 ***
Age                             -0.34773    0.07950  -4.374 1.22e-05 ***
NumCompaniesWorked               0.40123    0.05702   7.036 1.98e-12 ***
TotalWorkingYears               -0.51300    0.10795  -4.752 2.01e-06 ***
TrainingTimesLastYear           -0.20471    0.05680  -3.604 0.000313 ***
YearsSinceLastPromotion          0.54423    0.07684   7.083 1.41e-12 ***
YearsWithCurrManager            -0.49882    0.08500  -5.869 4.40e-09 ***
mean_timeatwork                  0.52888    0.05334   9.914  < 2e-16 ***
EnvironmentSatisfaction         -0.41945    0.05536  -7.577 3.54e-14 ***
JobSatisfaction                 -0.36015    0.05534  -6.508 7.61e-11 ***
BusinessTravelTravel_Frequently  1.74877    0.27505   6.358 2.04e-10 ***
BusinessTravelTravel_Rarely      1.00218    0.25951   3.862 0.000113 ***
DepartmentResearch...Development -1.09366   0.22142  -4.939 7.84e-07 ***
DepartmentSales                 -1.22087    0.23411  -5.215 1.84e-07 ***
JobRoleManufacturing.Director   -0.72749    0.21653  -3.360 0.000780 ***
MaritalStatusSingle              1.00066    0.11308   8.849  < 2e-16 ***
---
```

15 variables

AICC=2172.7

There are 4 variables in the final model that have high vif (>2.5) BusinessTravelTravel_Frequently,BusinessTravelTravel_Rarely,DepartmentResearch…Development,DepartmentSales . The 4 variables have high significance. We attempted model removing them and found that AICC increased by 30. We decided to keep the 4 variables as removing them would drop the accuracy of final model

Model Evaluation

- Model was tested using the ‘test’ data
- First , a probability cut-off of 50% was used; table for predicted attrition and actual attrition was plotted

	Test Pred Attrition	
Test Actual Attrition	NO	YES
NO	1088	22
YES	162	51

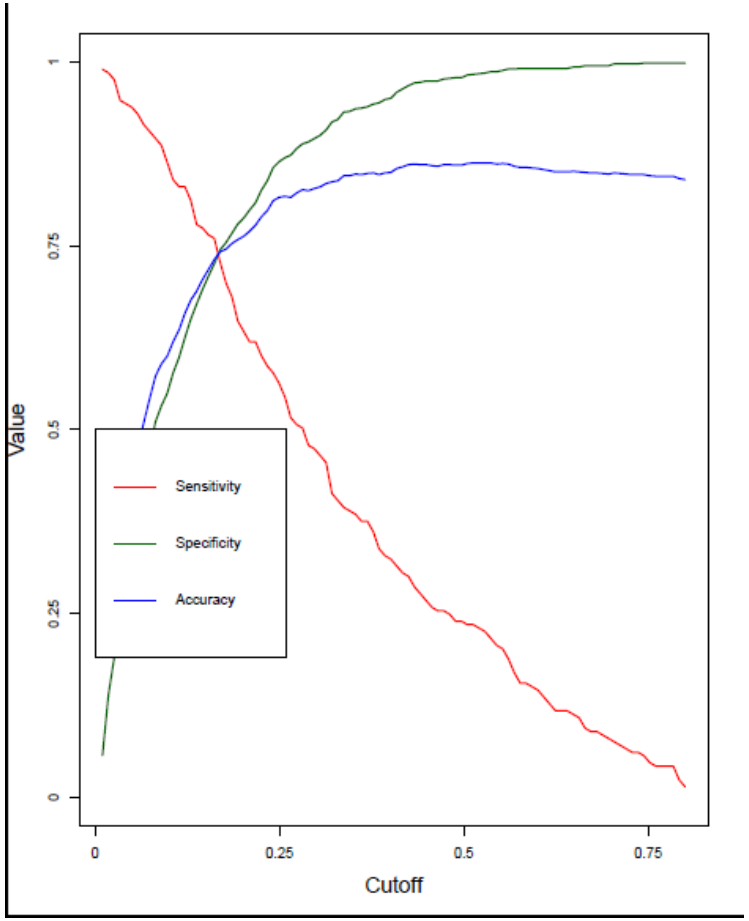
The 162 predicted 'No' that are actually 'Yes' needs to be reduced

Model cut-off needs to be lowered

Model Evaluation

- Cut-off lowered to 0.4 and ‘confusionMatrix’ calculated
  - Sensitivity = 0.323
- Cut-off lowered to 0.3 and ‘confusionMatrix’ calculated
  - Sensitivity = 0.47418
- To find optimal cut-off probability value, a matrix of cut-off values were used and the value of probability was chosen such that sensitivity and specificity difference was < 0.005(nearly equal sensitivity and specificity)
- The cut-off probability was found to be 0.169596

Model Evaluation



Final Confusion matrix with the Probability of 0.16959 was generated

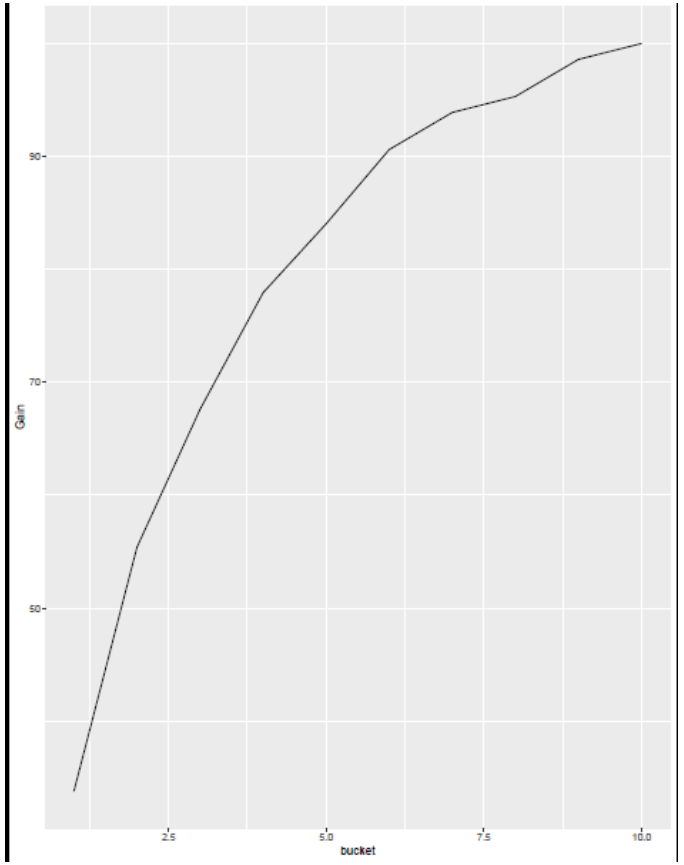
	Reference	
Prediction	NO	YES
No	827	57
Yes	283	156

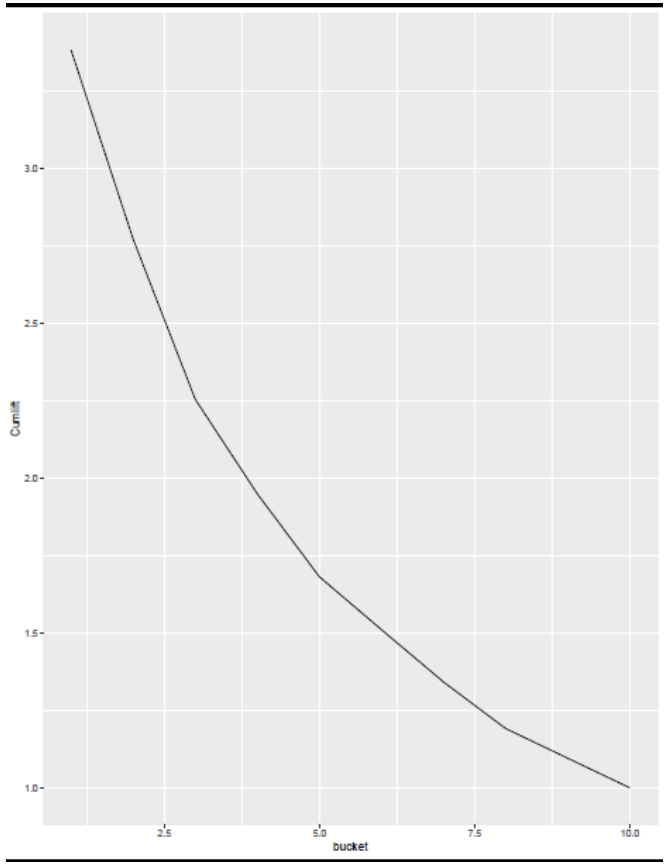
Accuracy : 0.7430
Sensitivity : 0.7324
Specificity : 0.7450

KS statistic using the ‘ks_table_test’ function was calculated
KS statistic is 47.74%

Model Evaluation

- Gain and lift charts plotted





bucket	total	totalresp	Cumresp	Gain	Cumlift
<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	133	72	72	33.8	3.38
2	132	46	118	55.4	2.77
3	132	26	144	67.6	2.25
4	133	22	166	77.9	1.95
5	132	13	179	84.0	1.68
6	132	14	193	90.6	1.51
7	133	7	200	93.9	1.34
8	132	3	203	95.3	1.19
9	132	7	210	98.6	1.10
10	132	3	213	100	1

- ✓ In the Gain, plot 77.9% attrition will be detected by the 4th decile
- ✓ In the lift plot, we find that by the 3rd decile, the model catches attrition 2.25 times better than the random model

Conclusions

- Based on the coefficients of the final model, we see that the variables can be divided into 2 groups

Factors that increase attrition (positive coeff)	Factors that decrease attrition (negative coeff)
BusinessTravelTravel_Frequently(1.7487)	DepartmentSales(-1.22087)
BusinessTravelTravel_Rarely(1.00225)	DepartmentResearch…Development (-1.09366)
MaritalStatusSingle(1.00059)	JobRoleManufacturing.Director (-0.72749)
YearsSinceLastPromotion(0.5442)	TotalWorkingYears(-0.51300)
mean_timeatwork(0.5288)	YearsWithCurrManager(-0.49882)
NumCompaniesWorked(0.40123)	EnvironmentSatisfaction(-0.41945)
	JobSatisfaction(-0.36015)
	Age(-0.34773)
	TrainingTimesLastYear(-0.20471)

Conclusions

- ✓Marital Status Single group have higher attrition
- ✓People who have spent more years since last promotion tend to leave. Indicating that people should be recognized well by the organization by giving promotions on time
- ✓Those who spend more and more time at work; greater than standard working hours (8) tend to leave. An indicator of poor work life balance
- ✓People who have worked in more companies(NumCompaniesWorked) will have a greater tendency to leave
- ✓Business Travel both frequently and rarely can rise attrition. The NoTravel group has less attrition

Conclusions

- ✓As the Total Work Experience increases, lesser attrition. Senior people will stay
- ✓As YearsWithCurrManager increases, attrition decreases. People like to work under the same manager once they have stayed for a while with the manager.
- ✓Environment and Job satisfaction ratings given by employee are good indicators . Higher Rating means employee will stay
- ✓Higher Age, lower attrition
- ✓More Trainings, more employee engagement, less attrition
- ✓People in Sales and Research and Development stay in company