HR ANALYTICS CASE STUDY

SUBMISSION

<u>Group Members</u>
Priya Raviraman
Muthuraman M PL
Raviraman Srinivasa Raghavan
Indira Srinivasa Raghavan

Abstract

Problem: A company, XYZ has a significant attrition rate (15%) which results in delayed deliverables in turn causing reputational loss .They also face recruitment overhead/challenges

Objective: Provide data-backed suggestions to Company XYZ, to identify key factors contributing to attrition enabling them to have a better business planning to retain employees.

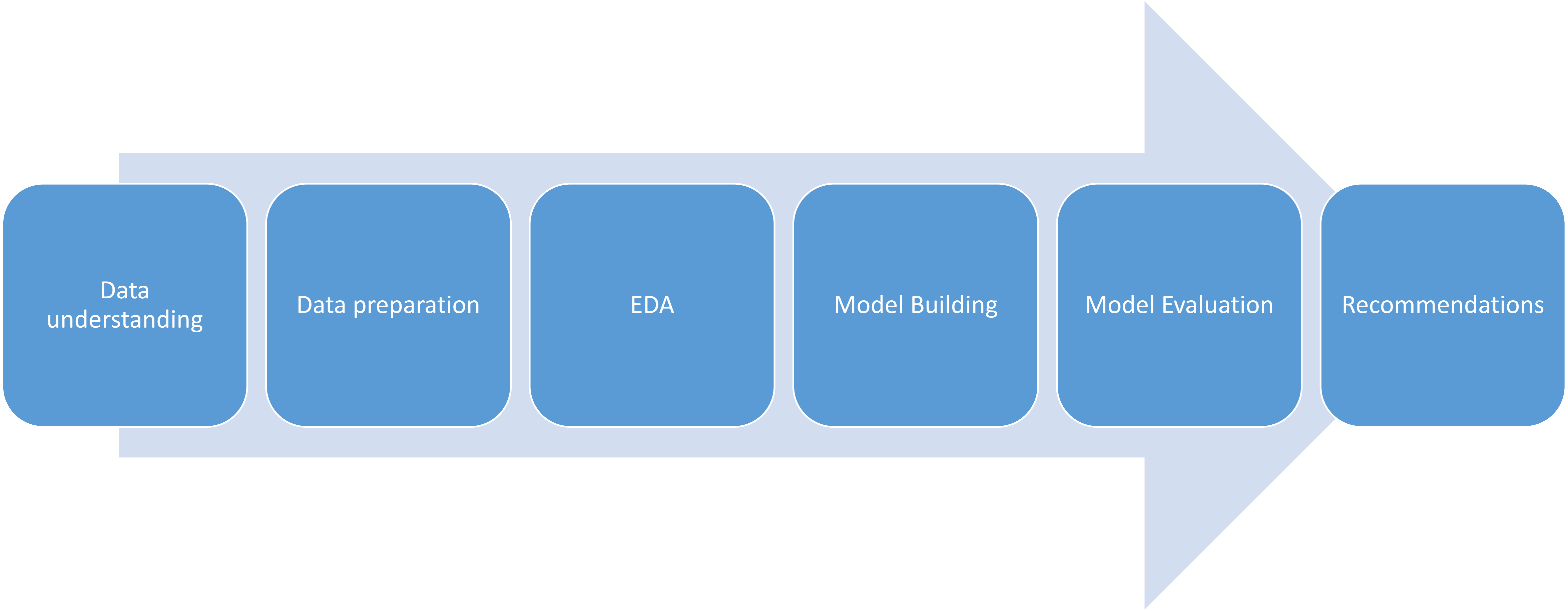
Constraints: To work around data quality issues and produce a reliable Logistic Regression model that would identify the key factors of attrition.

Strategy: XYZ will revisit their business plan based on the feedback received from HR analytics Firm and make positive changes to improve employee retention

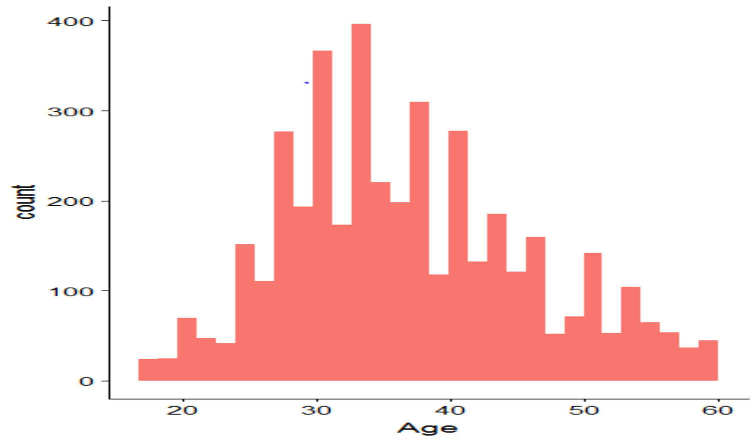
Goals:

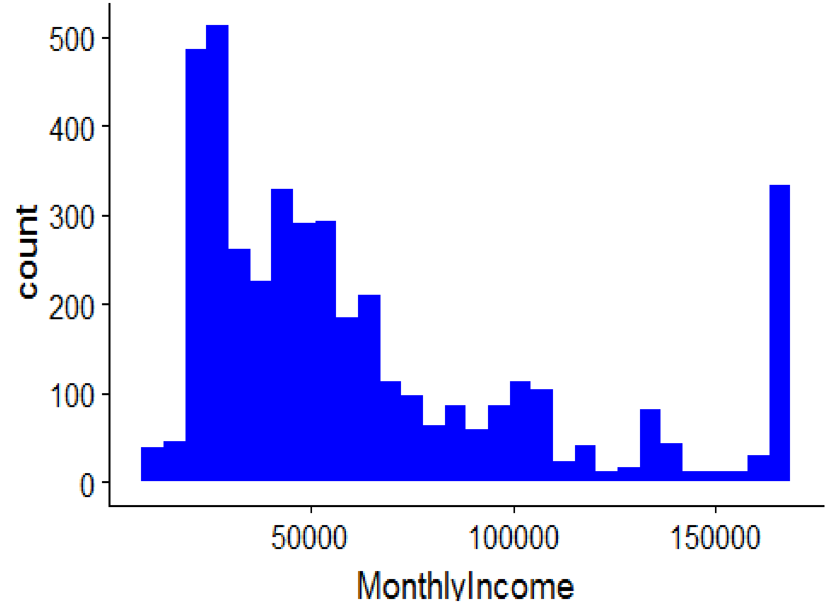
- Model the probability of attrition using a logistic regression
- Identify Key factors contributing to attrition
- List down suggestions to retain employees

Problem-solving methodology



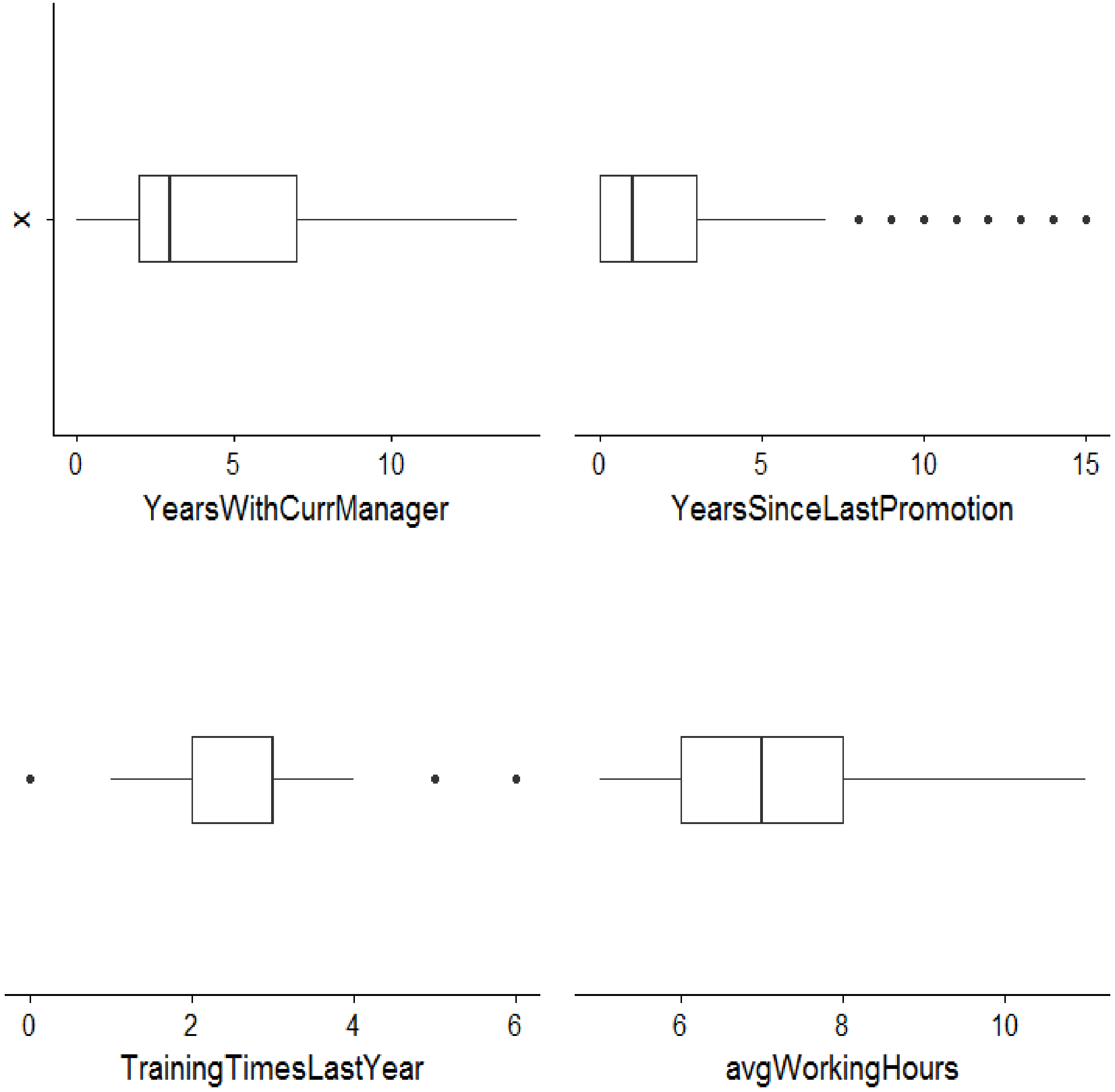
Data Understanding





- Given 5 data sets containing records of 4410 employees. EmployeeID is an auto-generated unique key variable that is common in all the 5 data sets.
- The data comprises of ,
  - General HR details(age, marital status, gender, workExp.,companies, monthly income, attrition) (24 columns). Attrition is the response variable
  - Swipe history for 2015 (260 columns in each IN and OUT time)
  - Manager Feedback (2 Columns)
  - Employee Engagement Survey Results (3 Columns)
- 16% employee attrition found in 2015.
- The variables like age, distance travelled, NumberofCompaniesWorked, hike, TotalYearsWorked qualify for bucketing to improve explainability.
- Outliers were identified
- Key data anomalies found
  - When general, empl_survey, manager_survey data are merged and duplicates identified without considering EmployeeID , there were 2837 duplicates (i.e., 29 rows matched exactly)
  - All 9 JobRoles are spread across all departments and levels .
  - Average Monthly Income for each Joblevel is very similar

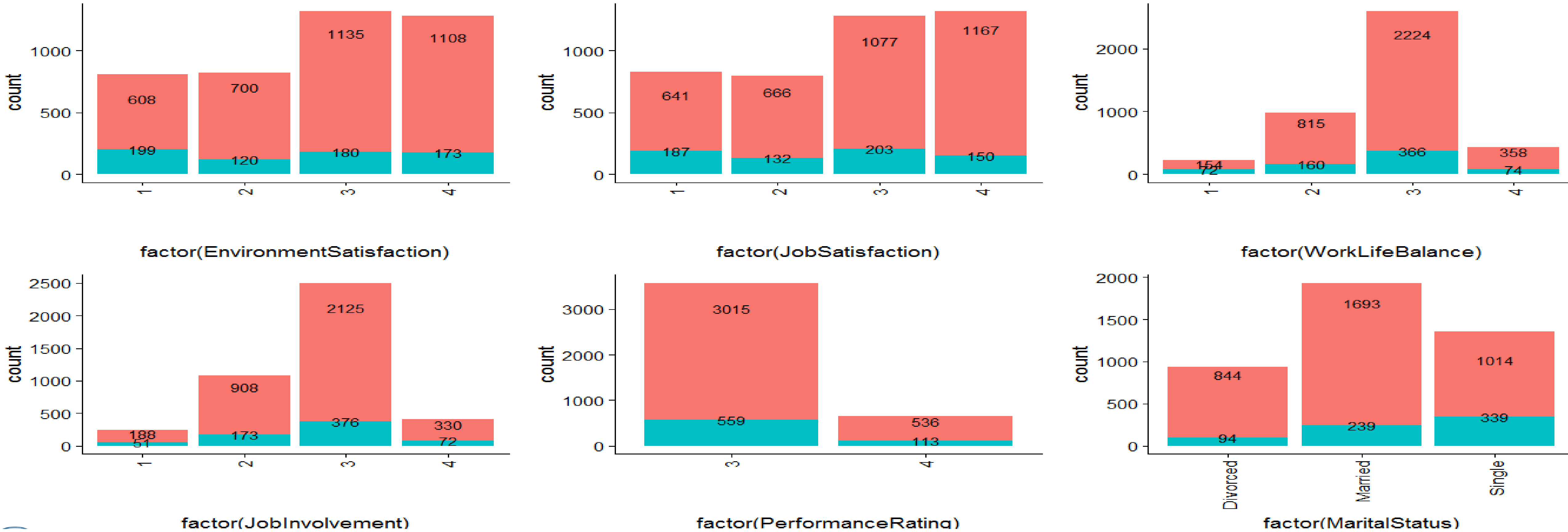
Data Preparation



- Bucketing and outlier handling was done as per data understanding
- Data errors found in NoCompaniesworked, TotalWorkingYears and YearsAtCompany were addressed for NA's and 0 values
- The working hours per employee is calculated from swipe report and average working hours determined. No. of leaves taken by each employee calculated. Holidays were removed.
- Scaling of numerical variables performed
- Dummy variables were created for categorical ones.
- De duplicated the data set to get the total of 4223 records.

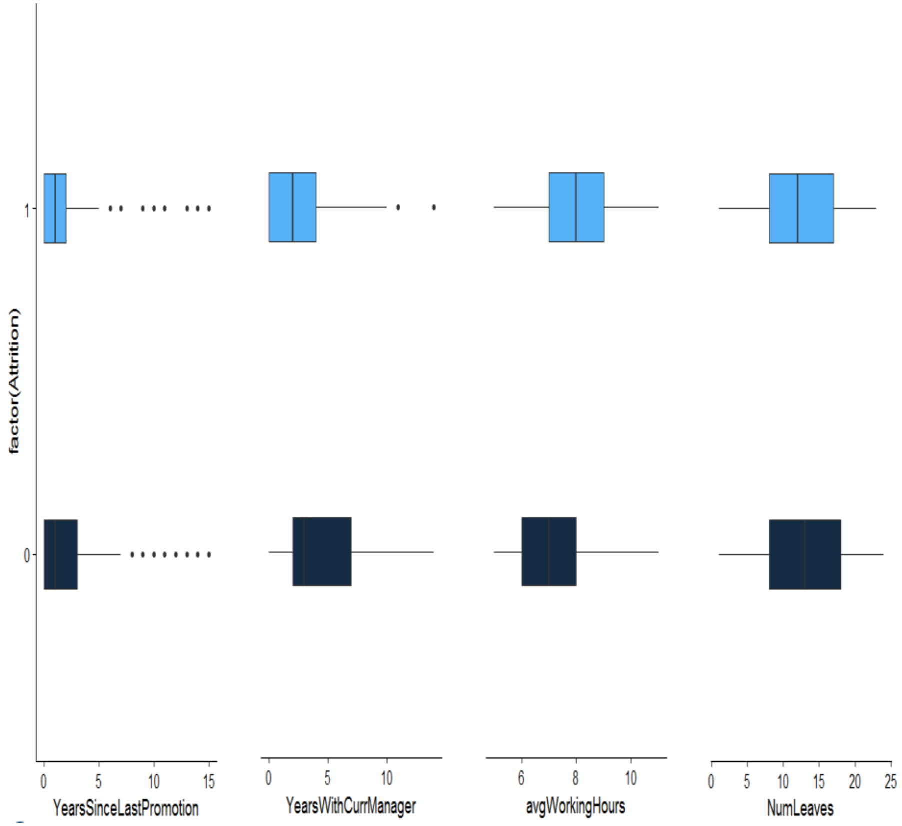
Outliers for YearsSinceLastPromotion and TrainingTimesLastYear were retained as they may be key factors for determining attrition.

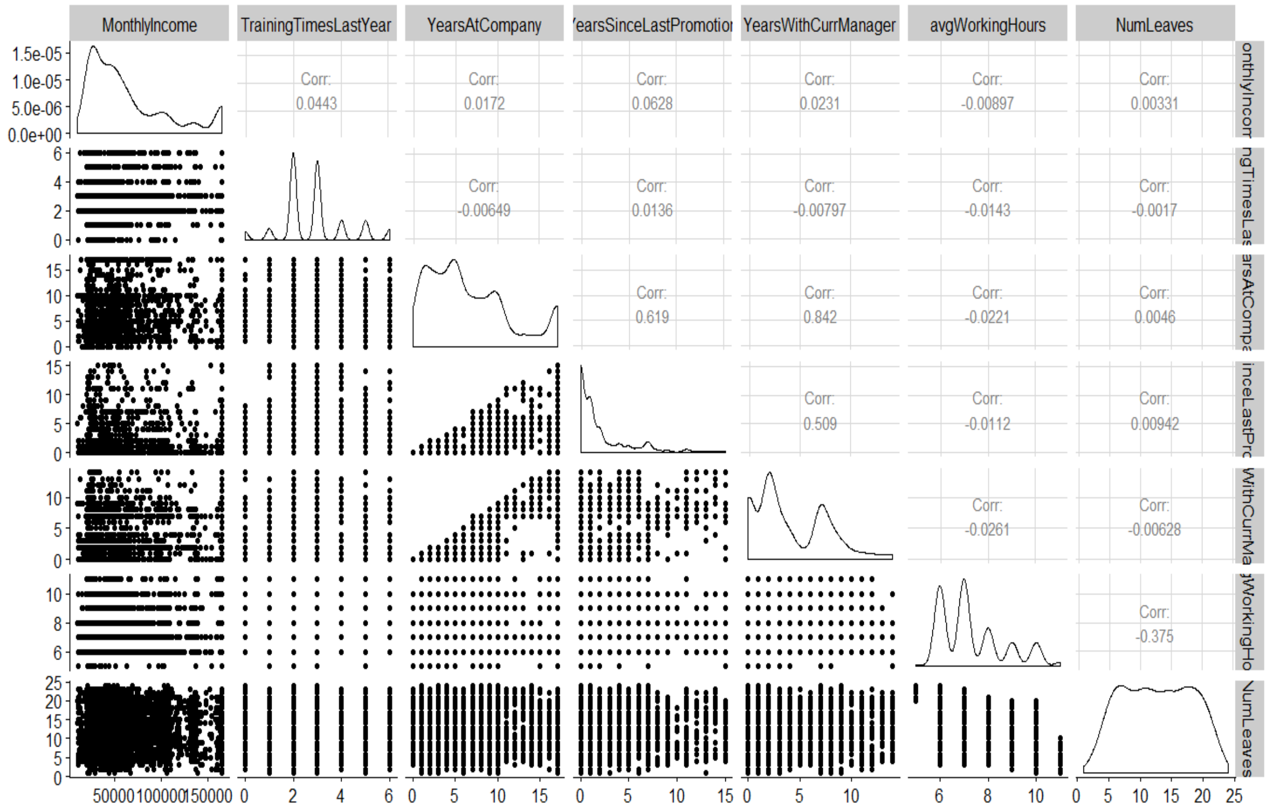
EDA – Categorical factors



Class	Attrition Rate	Class	Attrition Rate
Education Field - HR	41%	Age Under 30	27%
HR Department	31%	Marital Status - Single	25%
Work life Balance - Bad	31%	Business travel-Frequent	24%
Experience Under 5 Years	31%	Job Role-Research director	24%

EDA – Continous factors





Data distribution varies between attrition and no attrition for the following variables:
Monthly income, YearsAtCompany, YearsWithCurrManager, AvgWorkingHours
The following factors are positively correlated: YearsAtCompany & YearsSinceLastPromotion, YearsAtCompany & YearsWithCurrManager, YearsWithCurrManager & YearsSinceLastPromotion
NumLeaves & avgWorkingHours are highly negatively correlated.

Model Building

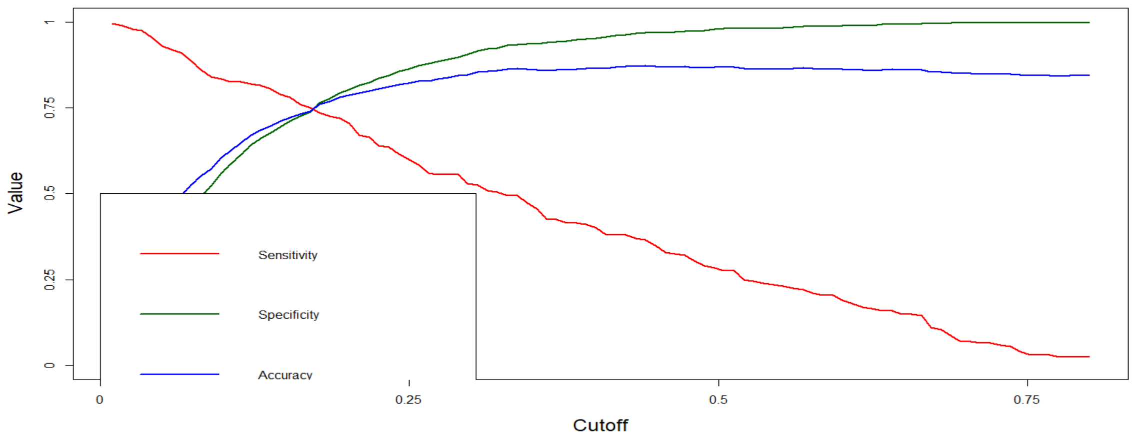
- Performed variable selection using AIC, VIF and p-value considerations, on the prepared data set with 4223 employee records and 57 columns
- Split the prepared data set in 70:30 proportion for training and test respectively
- Step 1: Invoked GLM function for binomial family on the training data set and got a AIC of 2058.3
- Step 2: Invoked step-wise AIC to iterate and finally obtained a model with AIC = 2028.7 and 35 independent variables
- Iterated on the model from step 2, 20 times removing one variable at a time with less significance (> 0.001 p-value) / high VIF
- Obtained a final model with 15 variables as follows (with all *** p-value and less than 2 VIF)
- This model is also in sync with EDA observations

ln(odds) = 0.59 - 0.2 * TrainingTimesLastYear + 0.46 * YearsSinceLastPromotion - 0.53 * YearsWithCurrManager + 0.56 * avgWorkingHours - 0.32 * age_bucket + 0.23 * NumCompaniesWorkedBucket - 0.52 * TotalWorkingYearsBucket + 0.84 * BusinessTravel.xTravel_Frequently + 0.93 * MaritalStatus.xSingle - 0.78 * EnvironmentSatisfaction.x2 - 0.85 * EnvironmentSatisfaction.x3 - 0.91 * EnvironmentSatisfaction.x4 - 0.61 * JobSatisfaction.x2 - 0.68 * JobSatisfaction.x3 - 1.13 * JobSatisfaction.x4 - 0.47 * WorkLifeBalance.x3

Model Evaluation

- Predicted the probabilities of attrition on the test data set (1267 records)
- Identified the optimal cut off for the probability by plotting accuracy, sensitivity and specificity at 100 points between 0.01 and 0.8
- Chose 0.17 which is close to the point where accuracy, sensitivity and specificity converge (as seen in the graph)
- From the gain chart, the model with 0.17 probability cut off is high performing with 100% attrition captured in the top 40% of the predicted probability.

With probability Cut off 0.17,
Accuracy - 73.9%
Sensitivity - 75%
Specificity - 73.8%
KS statistic – 48.8%



	bucket	total	totalresp	Cumresp	Gain	Cumlift
1	1	127	127	127	29.67290	2.967290
2	2	126	126	253	59.11215	2.955607
3	3	126	126	379	88.55140	2.951713
4	4	126	49	428	100.00000	2.500000
5	5	126	0	428	100.00000	2.000000
6	6	126	0	428	100.00000	1.666667
7	7	126	0	428	100.00000	1.428571
8	8	126	0	428	100.00000	1.250000
9	9	126	0	428	100.00000	1.111111
10	10	126	0	428	100.00000	1.000000
11	NA	6	0	428	100.00000	NA

Model Interpretation

- Key factors impacting attrition are as follows:

  - Positive impact (that is, additive increase in these factors produce a multiplicative increase in odds of attrition)

    - **MaritalStatus.xSingle**
    - **BusinessTravel.xTravel_Frequently**
    - **avgWorkingHours**
    - **YearsSinceLastPromotion**
    - **NumCompaniesWorkedBucket**

  - Negative impact (that, additive increase in these factors produce a multiplicative decrease in odds of attrition)

    - **JobSatisfaction.x4**
    - **EnvironmentSatisfaction.x4**
    - **EnvironmentSatisfaction.x3**
    - **EnvironmentSatisfaction.x2**
    - **JobSatisfaction.x3**
    - **JobSatisfaction.x2**
    - **YearsWithCurrManager**
    - **TotalWorkingYearsBucket**
    - **WorkLifeBalance.x3**
    - **age_bucket**
    - **TrainingTimesLastYear**

Recommendations

- Based on the coefficient strength in the final prediction model, below recommendations are put forth:
  - **Single employees have very high attrition rates (25%)**
    - Maintain positive work environment (through food subsidies, dorm facilities, free transport, “FunAtWork” activities, updated job infrastructure and facilities)
    - Institute/review mentoring program
    - Institute periodic team building activities
    - Reward longevity
  - **Frequent business travelers face travel fatigue and hence spike attrition rates (24%)**
    - Specialized leave policy to support family needs
    - Flexible work hours
    - Travel perks including increased per-diem, increased travel kit package, improved stay and transport at destinations
    - Consider periodic role shifts to less-travelling roles
  - **16% of employees putting in over 5 hours have quit**
    - Institute a part-time work program. This may be suitable for retaining such employees
    - Understand efficiency of such employees to determine if there is a deficit in training or distribution of work load
  - **47% of those with under 5 years experience have quit immediately after promotion**
    - Revisit policies for the first promotion of entry level employees to ensure industry standardization and greater stringency