IBM Employee HR Attrition

3X Data Mining

Mohammed Topiwalla, Patricia Londono, Sanchita Kumari,
Valerio Trotta, Igor Pedevani, Millicent Ottchere

Problem Statement

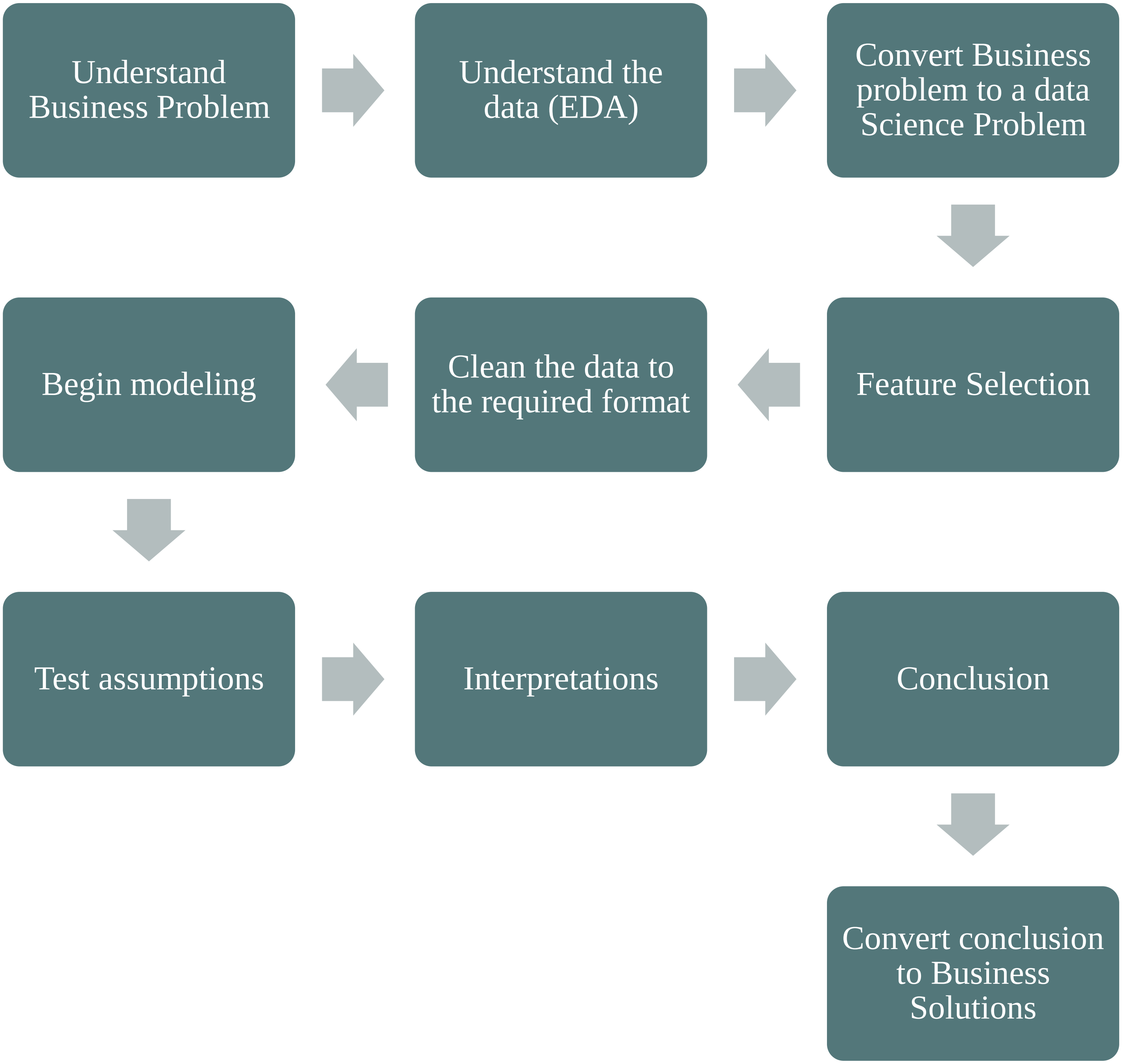
How can we reduce IBM company's attrition rate by predicting if a candidate will exit in India within the year?

- **S**pecific :- To Indian geography in IBM
- **M**easurable:- To reduce attrition rate(By at least 5%)
- **A**ction oriented:- Reduce employee attrition & suggest employee engagement & satisfaction programs
- **R**elevant:- Direct impact on company's top and bottom line
- **T**ime bound :- 12 months

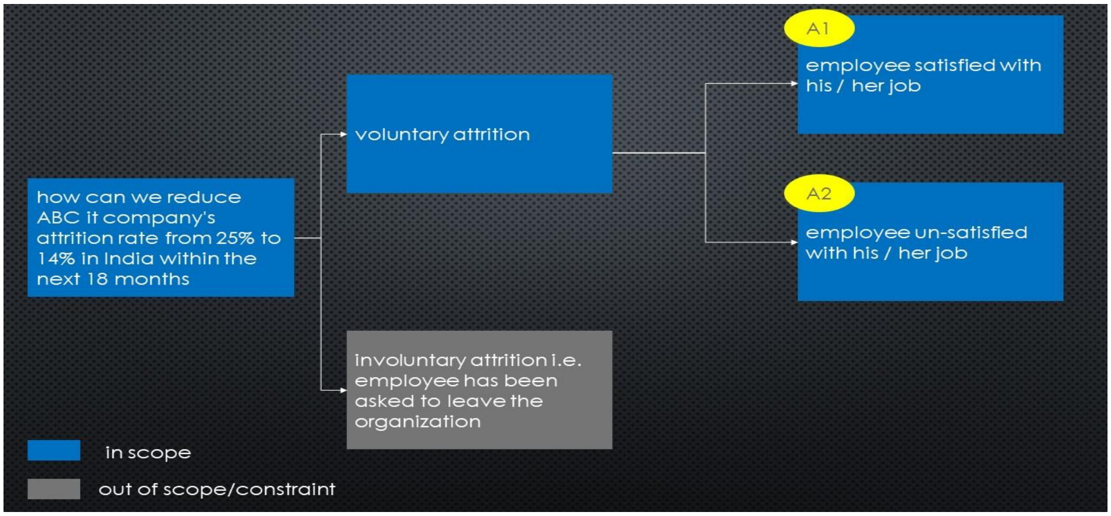
Our client is IBM a leading firm and in the IT sector. It is recently facing a steep increase in its employee attrition . Employee attrition has gone up from 14% to 25% in the last 1 year . We are asked to prepare a strategy to immediately tackle this issue such that the firm’s business is not hampered and also to propose an efficient employee satisfaction program for the long run. Currently, no such program is in place . Further salary hikes are not an option.

The attrition problem is not only unique to ibm but to other IT companies such as Infosys, India’s second largest IT services company, that is also battling high attrition, with a peak attrition of 20.4 % in the October-December quarter of FY15.

Methodology to solve the problem


Understand Business Problem
Understand the data (EDA)
Convert Business problem to a data Science Problem
Feature Selection
Clean the data to the required format
Begin modeling
Test assumptions
Interpretations
Conclusion
Convert conclusion to Business Solutions

Converting Business problem to Data Science problem



how can we reduce ABC it company's attrition rate from 25% to 14% in India within the next 18 months

voluntary attrition

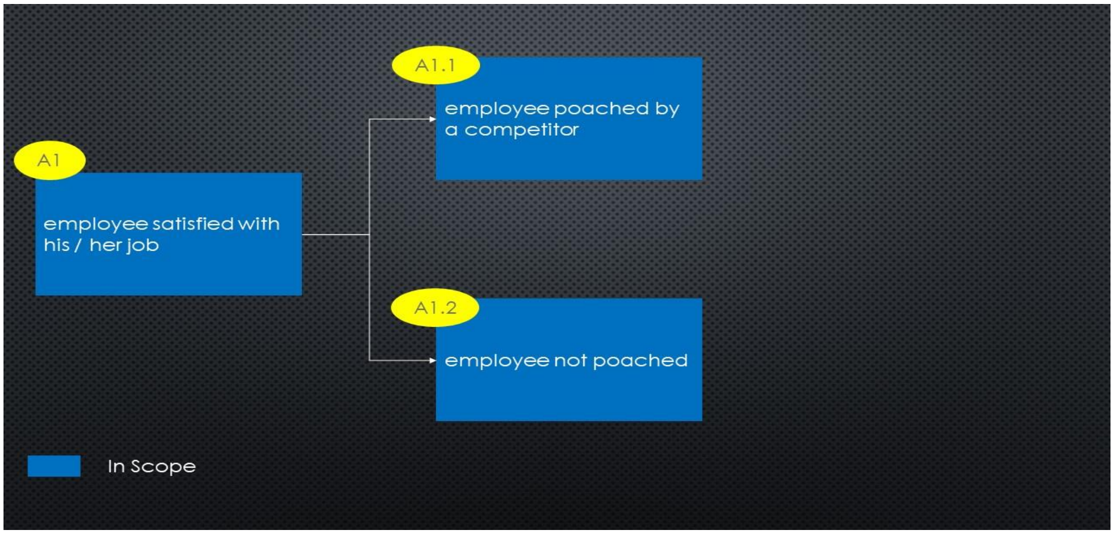
A1 employee satisfied with his / her job

A2 employee un-satisfied with his / her job

involuntary attrition i.e. employee has been asked to leave the organization

🟦 in scope

⬜ out of scope/constraint

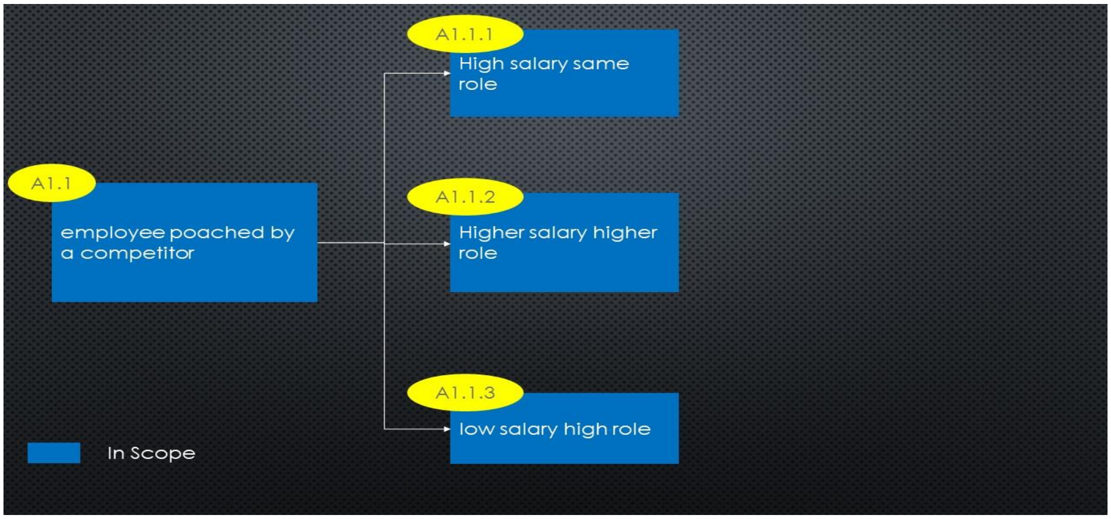


A1 employee satisfied with his / her job

A1.1 employee poached by a competitor

A1.2 employee not poached

🟦 In Scope



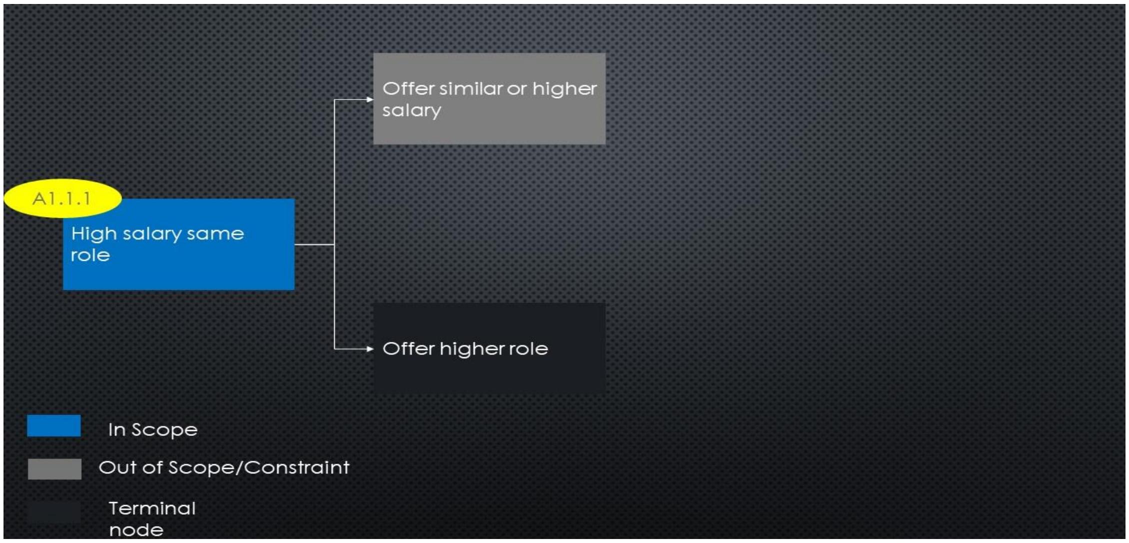
A1.1 employee poached by a competitor

A1.1.1 High salary same role

A1.1.2 Higher salary higher role

A1.1.3 low salary high role

🟦 In Scope



A1.1.1

High salary same role

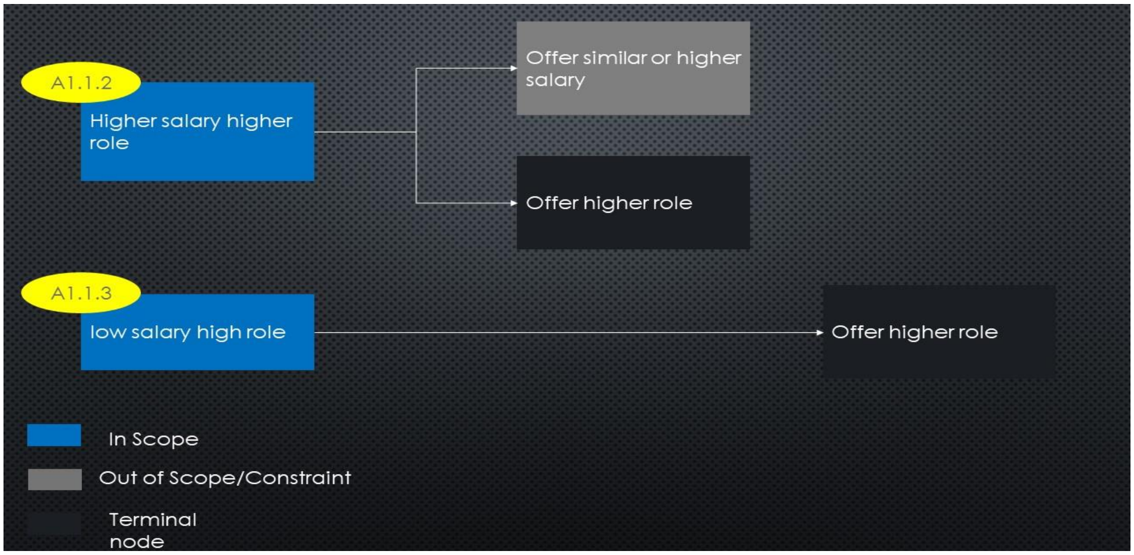
Offer similar or higher salary

Offer higher role

🟦 In Scope

▨ Out of Scope/Constraint

⬛ Terminal node



A1.1.2

Higher salary higher role

Offer similar or higher salary

Offer higher role

A1.1.3

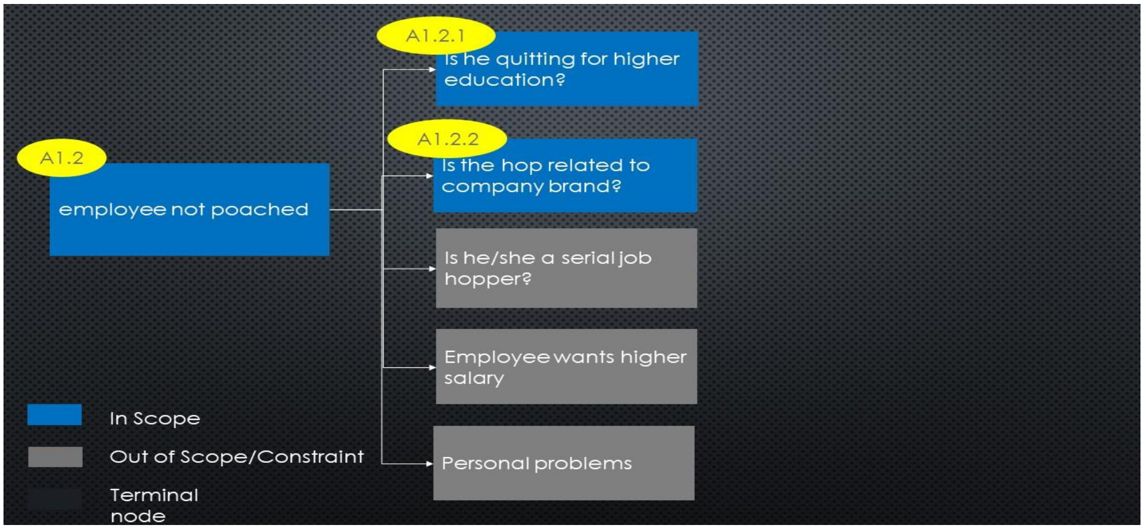
low salary high role

Offer higher role

🟦 In Scope

▨ Out of Scope/Constraint

⬛ Terminal node



A1.2

employee not poached

A1.2.1

Is he quitting for higher education?

A1.2.2

Is the hop related to company brand?

Is he/she a serial job hopper?

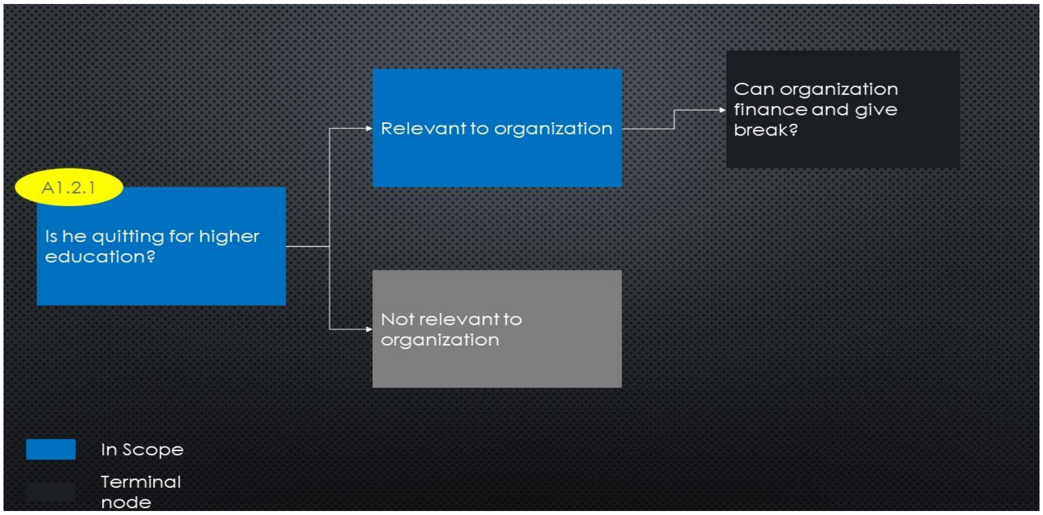
Employee wants higher salary

Personal problems

🟦 In Scope

▨ Out of Scope/Constraint

⬛ Terminal node



A1.2.1

Is he quitting for higher education?

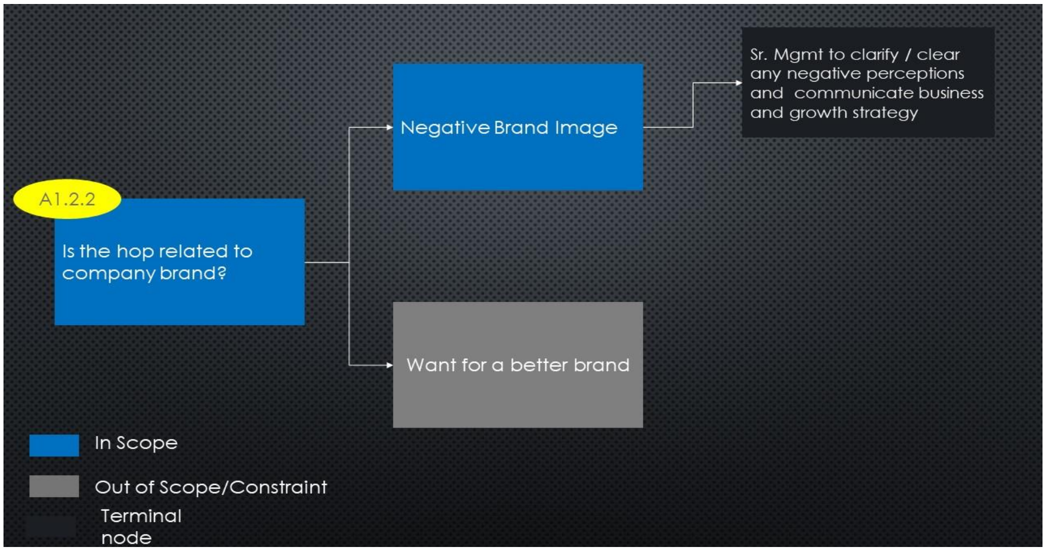
Relevant to organization

Can organization finance and give break?

Not relevant to organization

🟦 In Scope

⬛ Terminal node



A1.2.2

Is the hop related to company brand?

Negative Brand Image

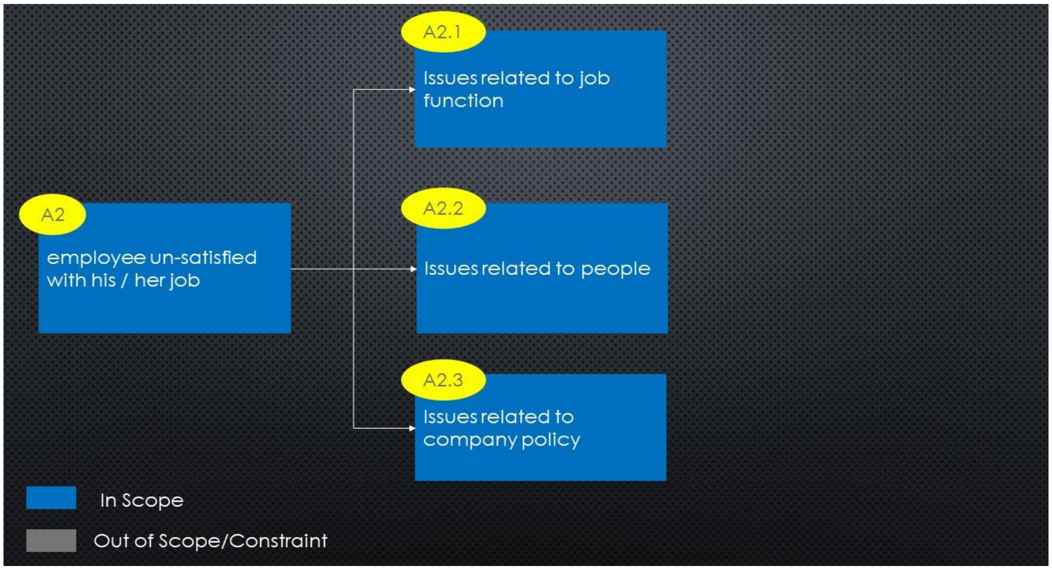
Sr. Mgmt to clarify / clear any negative perceptions and communicate business and growth strategy

Want for a better brand

🟦 In Scope

⬜ Out of Scope/Constraint

⬛ Terminal node



A2

employee un-satisfied with his / her job

A2.1

Issues related to job function

A2.2

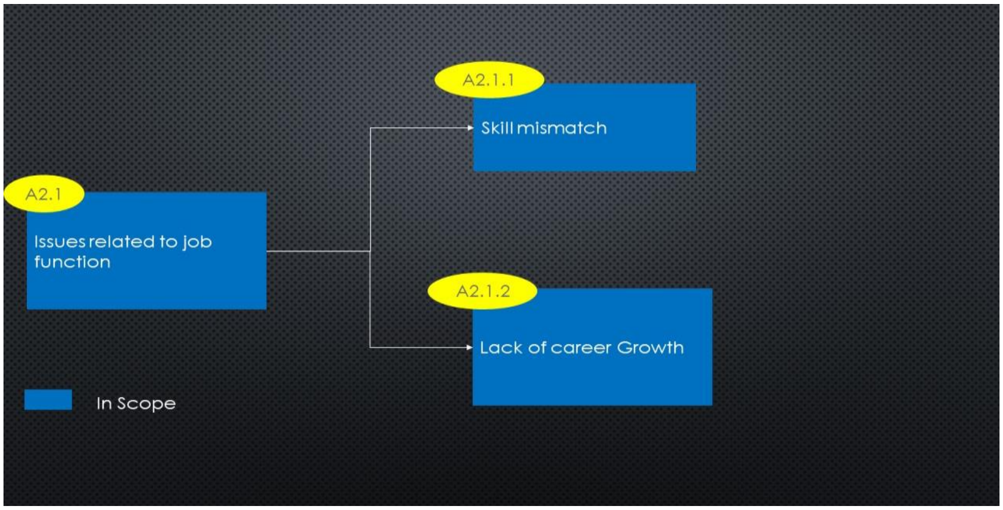
Issues related to people

A2.3

Issues related to company policy

🟦 In Scope

⬜ Out of Scope/Constraint



A2.1

Issues related to job function

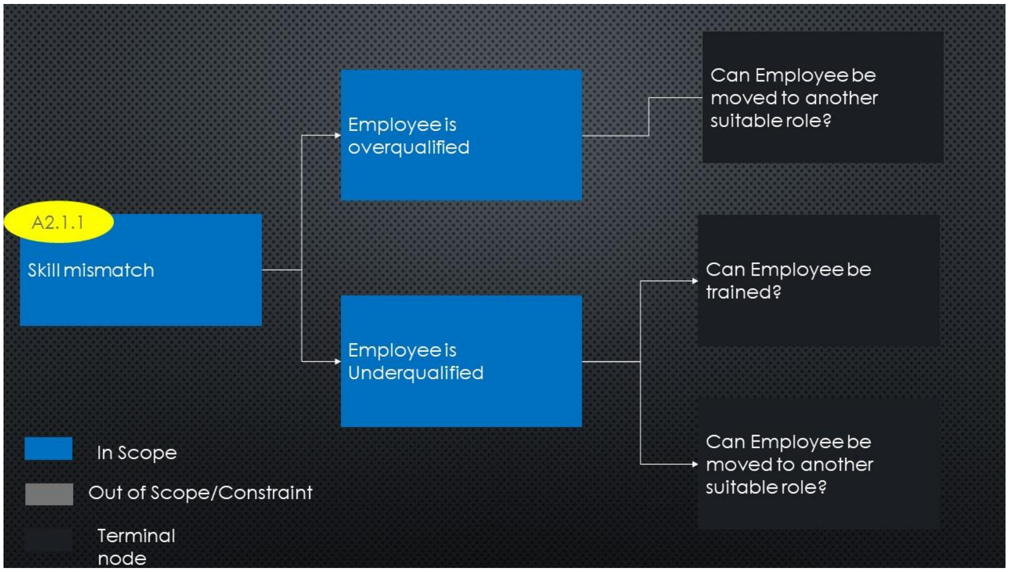
A2.1.1

Skill mismatch

A2.1.2

Lack of career Growth

🟦 In Scope



A2.1.1

Skill mismatch

Employee is overqualified

Can Employee be moved to another suitable role?

Employee is Underqualified

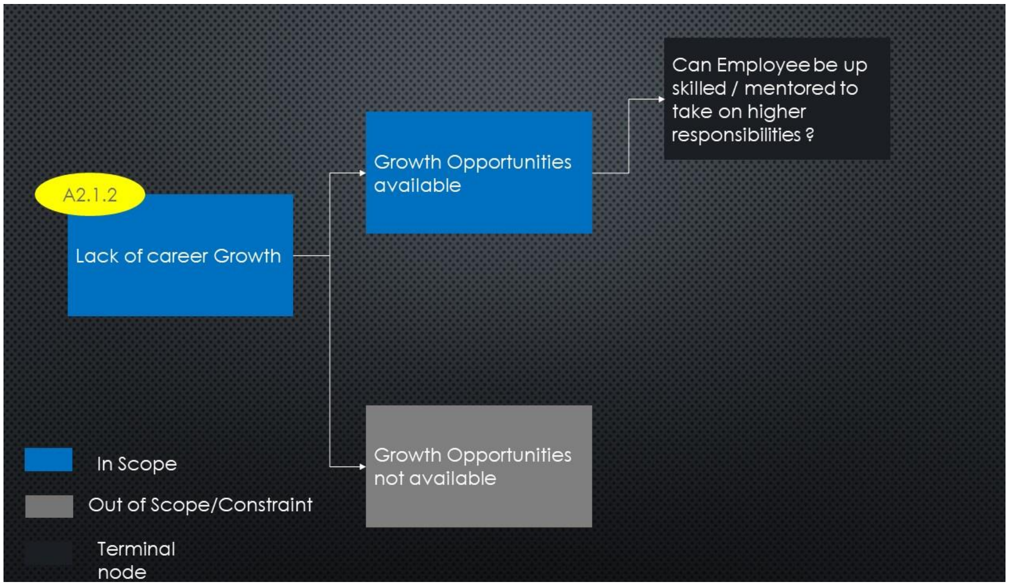
Can Employee be trained?

Can Employee be moved to another suitable role?

🟦 In Scope

▨ Out of Scope/Constraint

⬛ Terminal node



A2.1.2

Lack of career Growth

Growth Opportunities available

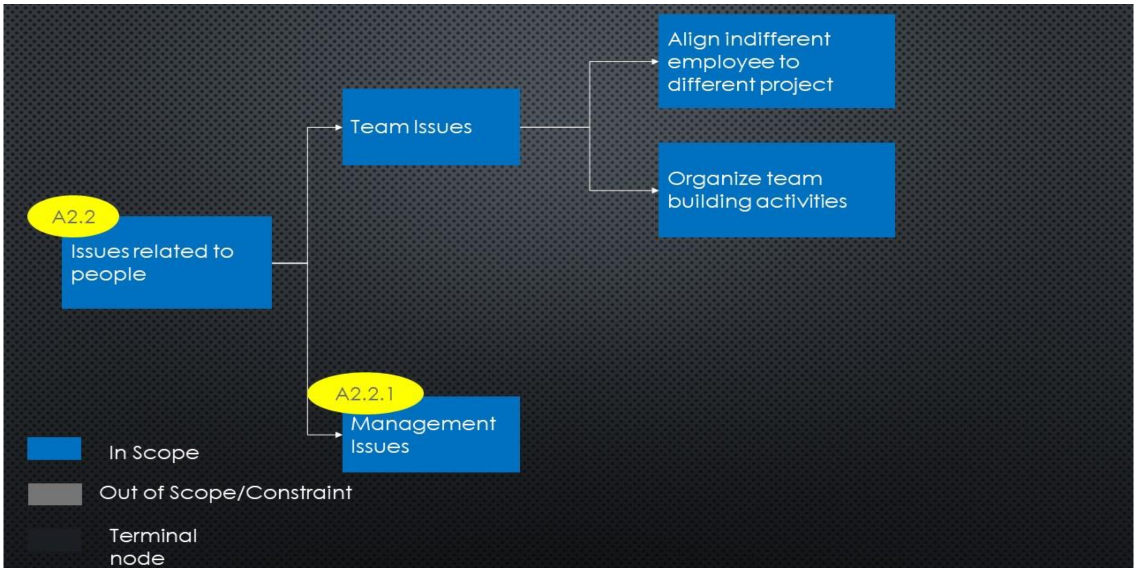
Can Employee be up skilled / mentored to take on higher responsibilities ?

Growth Opportunities not available

🟦 In Scope

▨ Out of Scope/Constraint

⬛ Terminal node



A2.2

Issues related to people

Team Issues

Align indifferent employee to different project

Organize team building activities

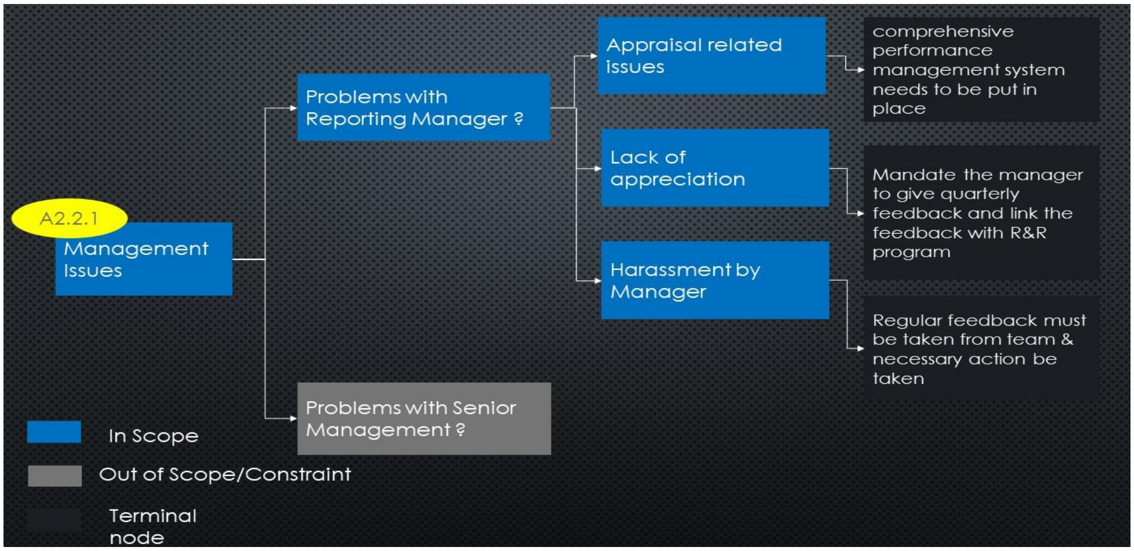
A2.2.1

Management Issues

🟦 In Scope

▨ Out of Scope/Constraint

■ Terminal node



A2.2.1

Management Issues

Problems with Reporting Manager ?

Appraisal related issues

comprehensive performance management system needs to be put in place

Lack of appreciation

Mandate the manager to give quarterly feedback and link the feedback with R&R program

Harassment by Manager

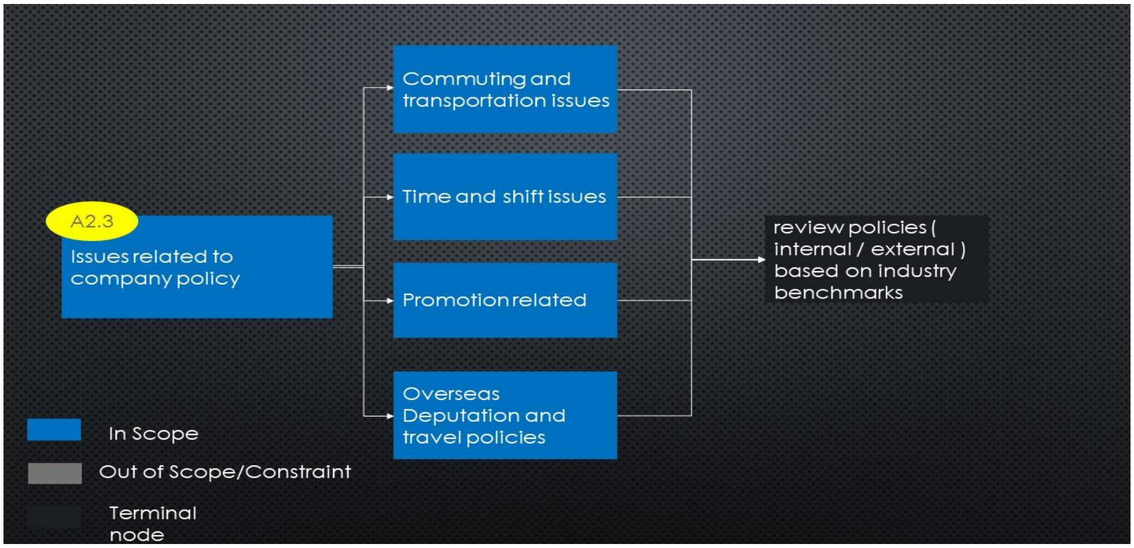
Regular feedback must be taken from team & necessary action be taken

Problems with Senior Management ?

🟦 In Scope

▨ Out of Scope/Constraint

■ Terminal node



A2.3

Issues related to company policy

Commuting and transportation issues

Time and shift issues

Promotion related

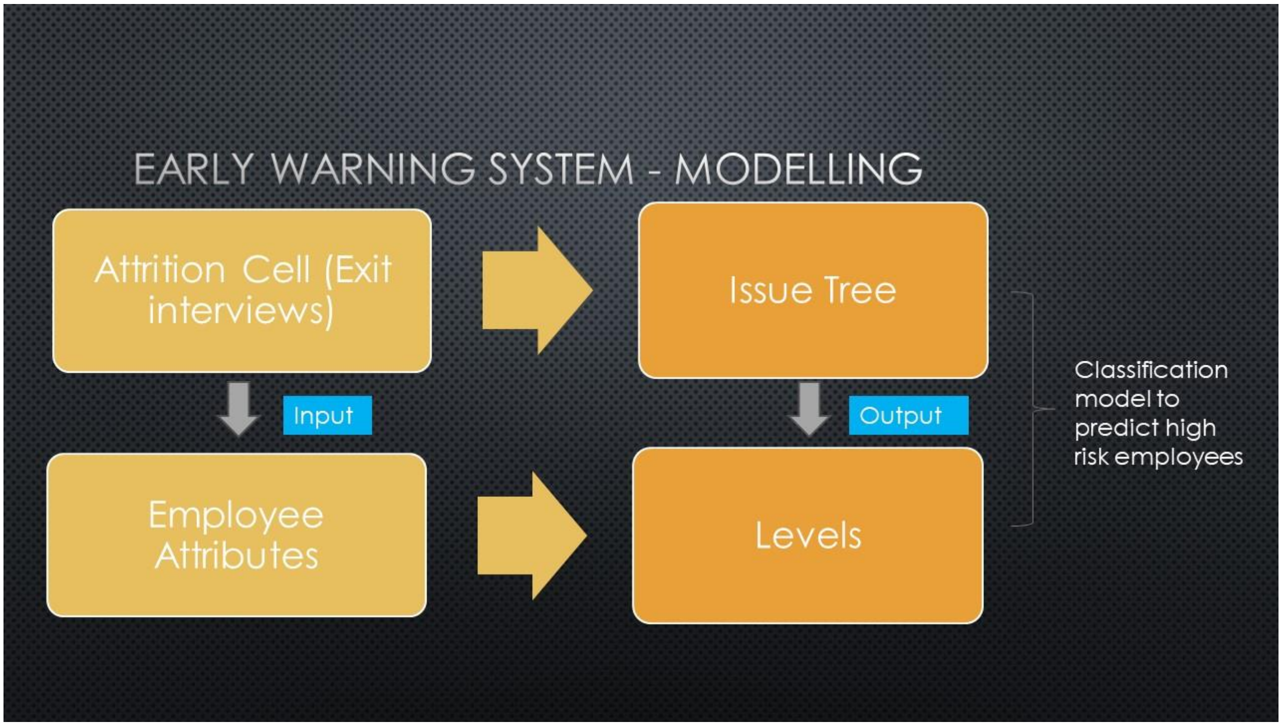
Overseas Deputation and travel policies

review policies (internal / external) based on industry benchmarks

🟦 In Scope

▨ Out of Scope/Constraint

■ Terminal node



EARLY WARNING SYSTEM - MODELLING

Attrition Cell (Exit interviews)

Input

Employee Attributes

Issue Tree

Output

Levels

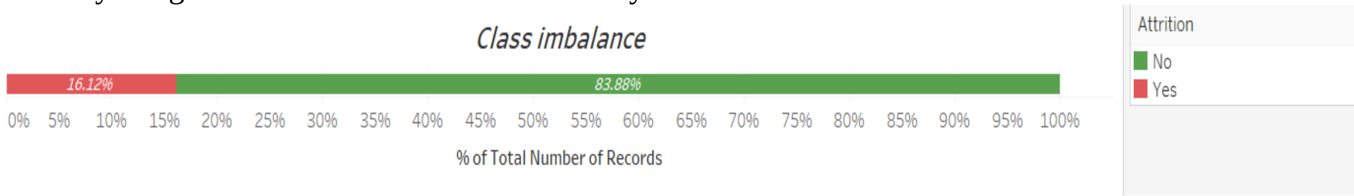
Classification model to predict high risk employees

Who made this – Mohammed, Millicent

Who reviewed this – Sanchita, Igor

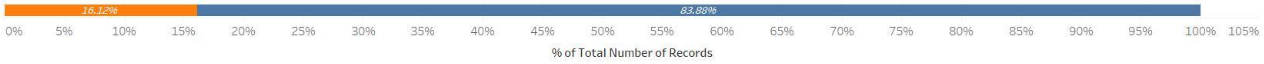
Exploratory Data Analysis

- We were blessed with an excellent data set, the source of data is here
- There were no nulls or blanks in our data set
- The only thing which was visible from miles away was the class imbalance



- The following insights were drawn from the EDA performed in Tableau
  - The age group of IBM employees in this data set is concentrated between 25-45 years
  - Attrition is more common in the younger age groups and it is more likely with females As Expected it is more common amongst single Employees
  - People who leave the company get lower opportunities to travel the company
  - People having very high education tend to have lower attrition
  - The correlation plot was as expected
  - Link to eda workbook in python is here
  - From the Tableau plots we can conclude that below mentioned category are having higher attrition rate:
    - Sales department among all the departments
    - Human Resources and Technical Degree in Education
    - Single's in Marital status (Will not use this due to GDPR)
    - Male in comparison to females in Gender (Will not use this due to GDPR)
    - Employee with job satisfaction value 1
    - Job level 1 in job level
    - Life balance having value 1
    - Employee staying at distant place
    - Environment Satisfaction value 1
  - Link to Tableau workbook for EDA is here

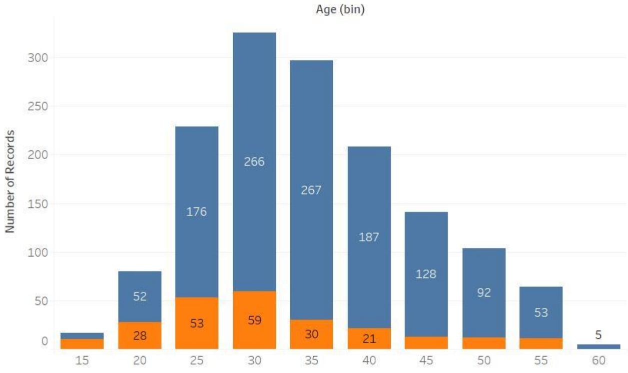
Class imbalance



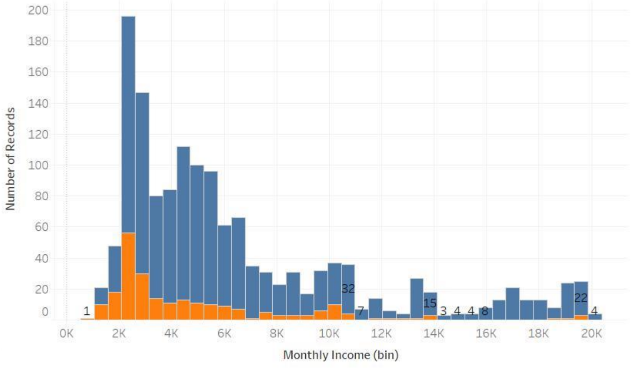
Attrition
🟦 No
🟧 Yes

Number of people who left vs who did not		*Average Age and Monthly incomes are*	
Yes	No	Age	Monthly Income
237	*1,233*	*54,278*	*9,559,309*

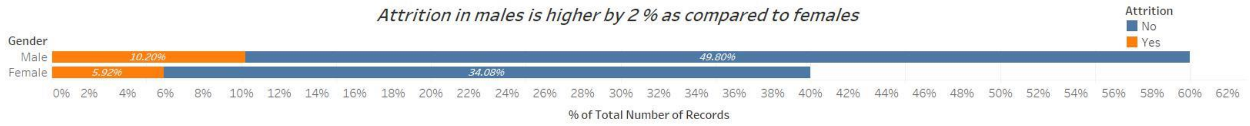
Histogram of age where color denotes attrition



Histogram of monthly income where color denotes attrition



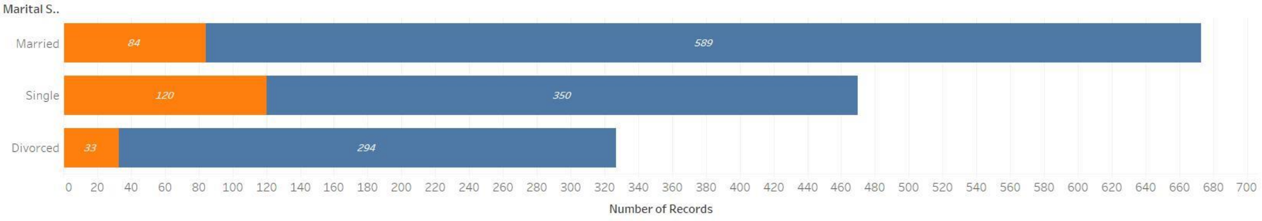
Attrition in males is higher by 2 % as compared to females

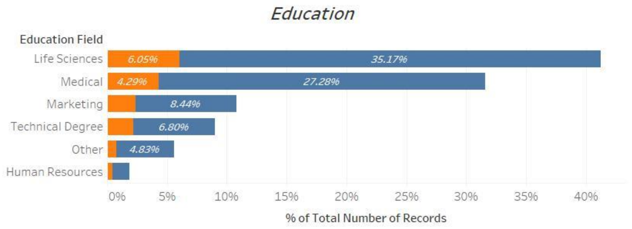


The employees dont seem to have a very positive view of the company and it seems to be common amongst both gender groups

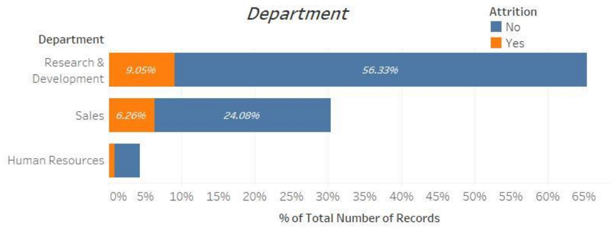
Attrition	Gender	Avg. Job Involvement	Avg. Job Satisfaction	Avg. Performance Rating	Avg. Work Life Balance
No	Female	*2.7465*	*2.7285*	*3.1577*	*2.7605*
	Male	*2.7869*	*2.8128*	*3.1503*	*2.7951*
Yes	Female	*2.5287*	*2.4253*	*3.1724*	*2.7816*
	Male	*2.5133*	*2.4933*	*3.1467*	*2.5867*

Does Marital status influence attrition?
Apparently it does - A single employee has a higher probability of exiting the company

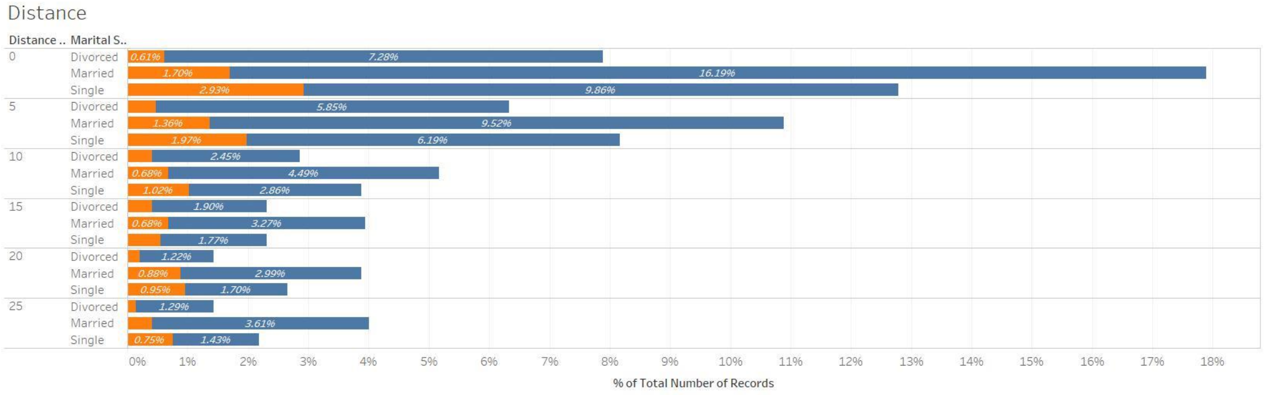




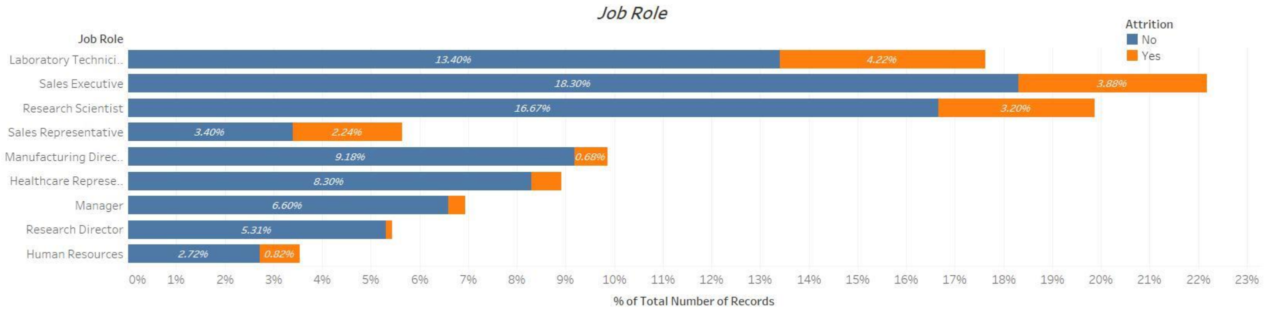
Education



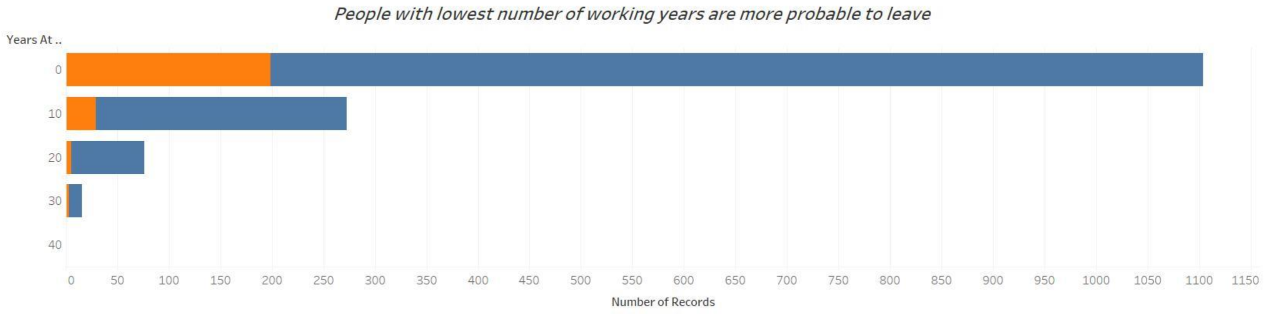
Department



Distance



Job Role



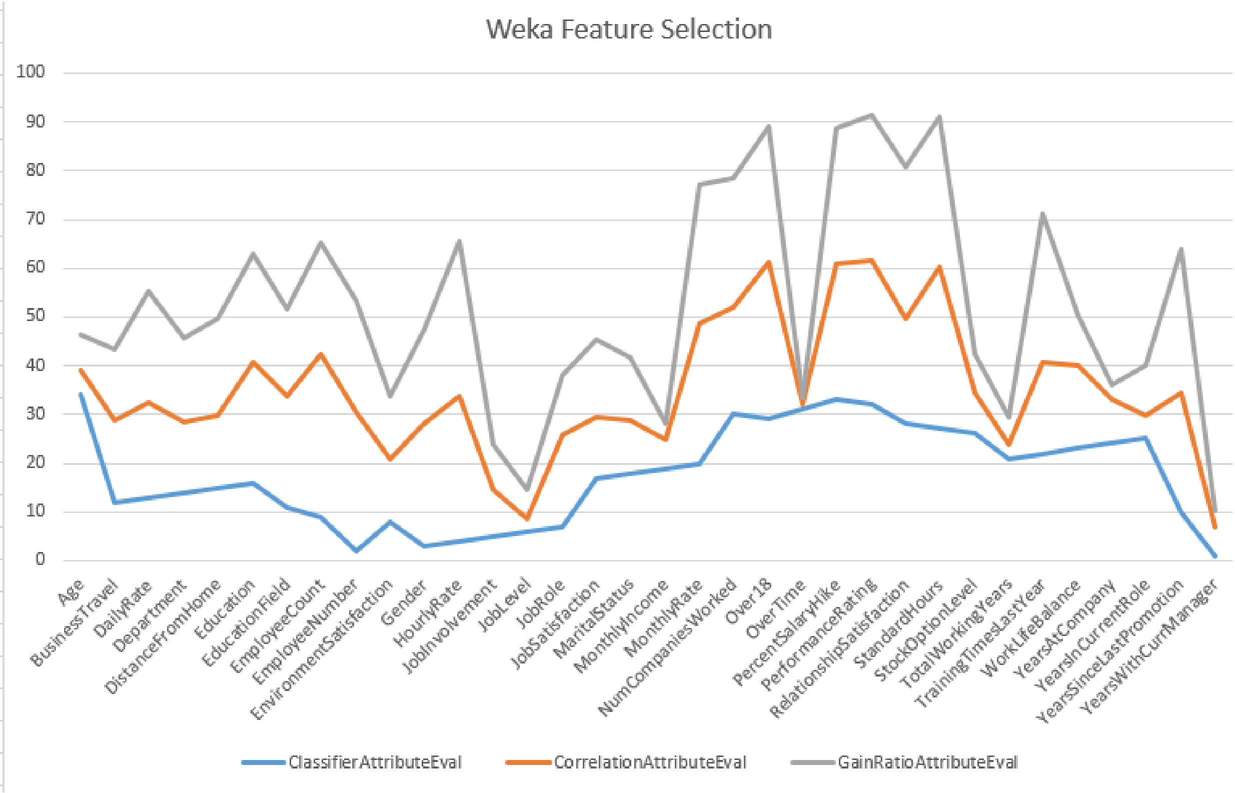
People with lowest number of working years are more probable to leave

Feature Selection

The following programs were used for Feature Selection:

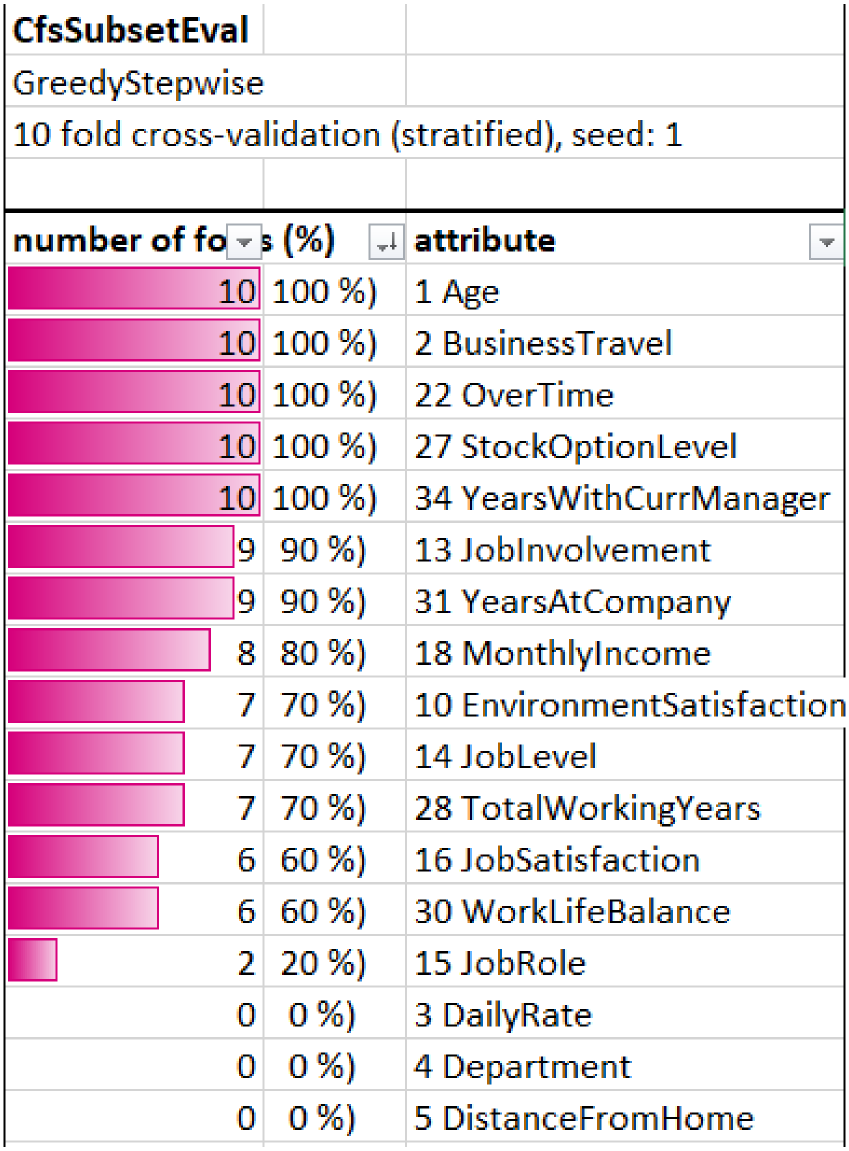
Weka:

Classifier Attribute Evaluation, Correlation Attribute Evaluation and Gain Ratio Attribute Evaluation. The chart below summarizes the results that these models found (higher peaks show most important features).



A Correlation based feature selection Subset Evaluation method was also run which proved to be the most useful in improving the accuracy of our classifier model. You can read about this feature selection method https://www.cs.waikato.ac.nz/~mhall/thesis.pdf .

See results below (Pink bars show the most important features):



CfsSubsetEval

GreedyStepwise

10 fold cross-validation (stratified), seed: 1

number of fo	s (%)	attribute
10	100 %)	1 Age
10	100 %)	2 BusinessTravel
10	100 %)	22 OverTime
10	100 %)	27 StockOptionLevel
10	100 %)	34 YearsWithCurrManager
9	90 %)	13 JobInvolvement
9	90 %)	31 YearsAtCompany
8	80 %)	18 MonthlyIncome
7	70 %)	10 EnvironmentSatisfaction
7	70 %)	14 JobLevel
7	70 %)	28 TotalWorkingYears
6	60 %)	16 JobSatisfaction
6	60 %)	30 WorkLifeBalance
2	20 %)	15 JobRole
0	0 %)	3 DailyRate
0	0 %)	4 Department
0	0 %)	5 DistanceFromHome

See excel file attached containing full reports for each of the methods explained above here.

SAS:

Logistic Regression

The LOGISTIC procedure fits linear logistic regression models for discrete response data by the method of maximum likelihood.

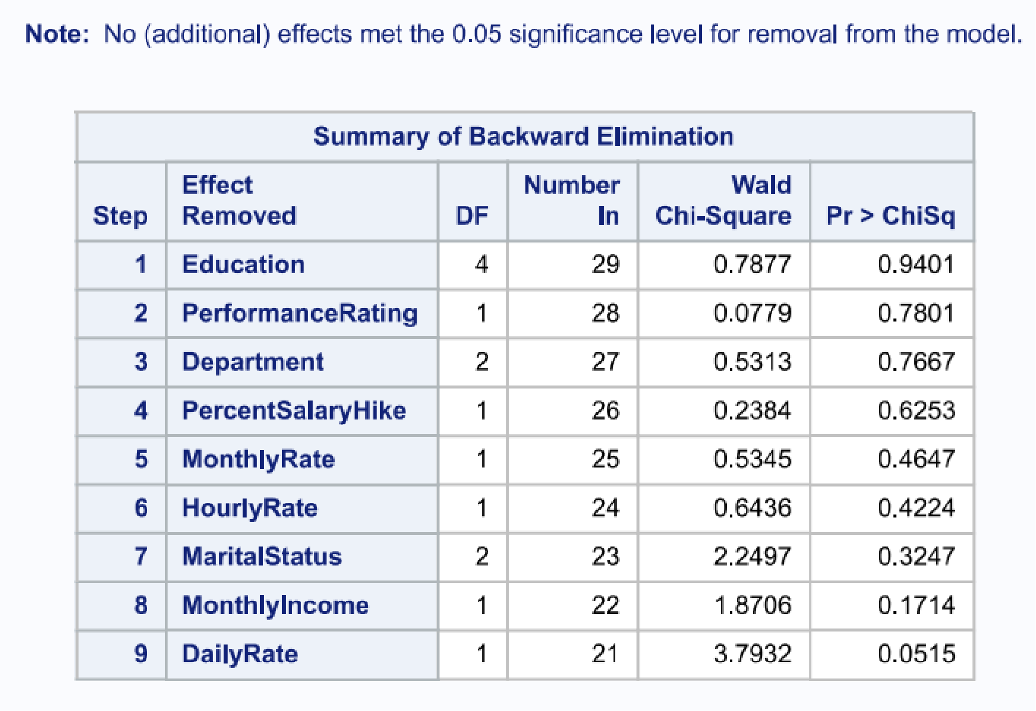
The backward elimination technique starts from the full model including all independent effects. Then effects are deleted one by one until a stopping condition is satisfied. At each step, the effect showing the smallest contribution to the model is deleted. In traditional implementations of backward elimination, the contribution of an effect to the model is assessed by using an F statistic. At any step, the predictor producing the least significant F statistic is dropped and the process continues until all effects remaining in the model have F statistics significant at a stay significance level (SLS).

More precisely, if the current model has p parameters excluding the intercept, and if you denote its residual sum of squares by RSS_p and you drop an effect with k degrees of freedom and denote the residual sum of squares of the resulting model by RSS_{p-k}, then the F statistic for removal with k numerator degrees of freedom and $n-p-k$ denominator degrees of freedom is given by

$$F = \frac{(RSS_{p-k} - RSS_p)/k}{RSS_p/(n-p-k)}$$

where n is number of observations used in the analysis.

Using the Backward elimination method, we were able to identify the following 9 variables that were not useful in predicting attrition in our classification model:



Note: No (additional) effects met the 0.05 significance level for removal from the model.

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	Education	4	29	0.7877	0.9401
2	PerformanceRating	1	28	0.0779	0.7801
3	Department	2	27	0.5313	0.7667
4	PercentSalaryHike	1	26	0.2384	0.6253
5	MonthlyRate	1	25	0.5345	0.4647
6	HourlyRate	1	24	0.6436	0.4224
7	MaritalStatus	2	23	2.2497	0.3247
8	MonthlyIncome	1	22	1.8706	0.1714
9	DailyRate	1	21	3.7932	0.0515

See full report attached here.

Who made this – Patricia

Who reviewed this – Millicent, Igor

Data cleansing and Smote

Our data set had no missing values therefore no special treatment was required, however since we are running a classification algorithm

A. We had many categorical variables which we needed to convert into dummy variables or ordered integer variables
B. There was a massive class imbalance to the proportion of 83:16 , in order to tackle this we understood that even without any algorithm if we predicted that the employee wouldn't leave we would still be right 83% of the time. So any model we make should be able to provide a better accuracy than this or else it would not be worth it.
C. The second solution was to balance the classes (A paper by Chawla led us the way - > https://www.jair.org/media/953/live-953-2037-jair.pdf)
D. With the help of smote and code help from stack overflow we achieved Smote
    a. What smote does is simple, it first of all takes the class with low proportion and artificially boosts its values to increase the number of records, it tries to generate new rows by replicating them in the bases of a range of values already present making various random combinations
    b. It takes the overpopulated proportion and performs sampling to pick up those rows which are the most representative of the population and then it uses them to reduce this proportion
    c. This is repeated till the classes are nearly balanced
E. Two techniques of smote from 2 different packages were used therefore we have 2 smote data sets

Who made this – Mohammed

Who reviewed this – Sanchita , Igor

Classification

In order to create an early warning system, it was essential to perform a supervised classification.

We had 3 sets of data

1. The original data (With one hot encoding)
2. Data generated from SMOTE (Set1)
3. Data generated from SMOTE (Set 2)

Before running any model it was ensured that the best parameters were selected on the basis of validation set or cross validation set. The splits were always on the basis of 50:25:25 (Training:Validation:Testing). All the sampling is stratified sampling in order to ensure perfect class balance

The approach was to run simple models like

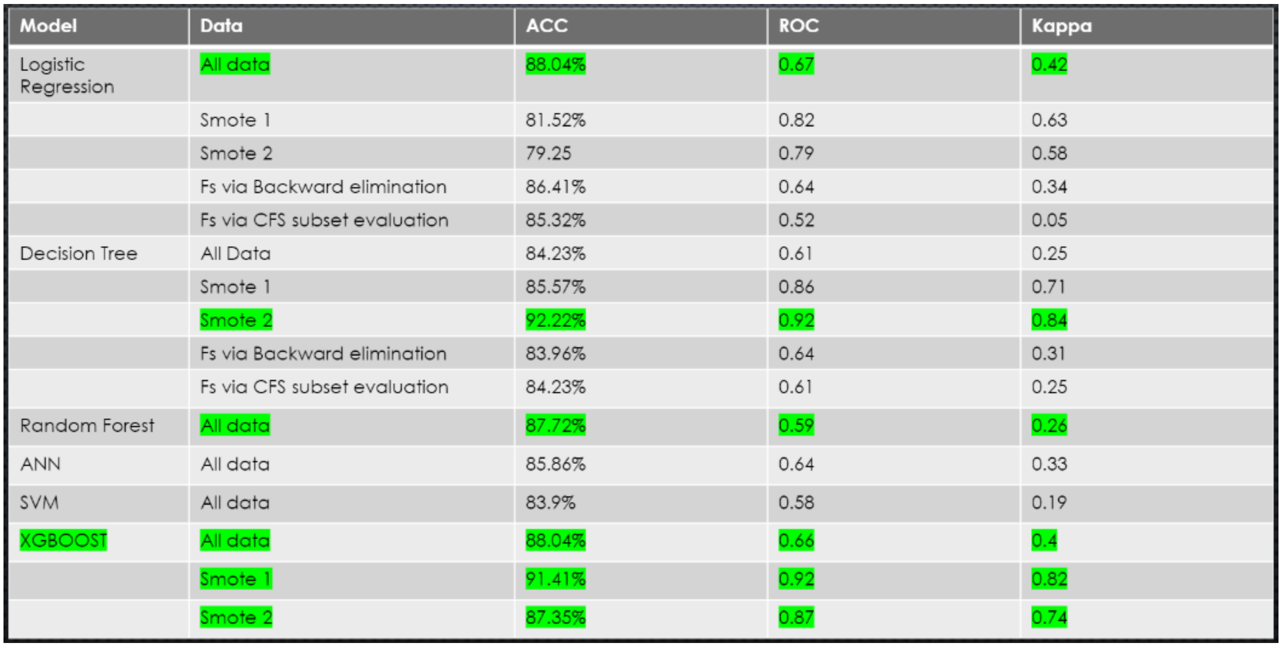
A. Decision tree
B. Logistic Regression

On both these models we would train each of the 3 data sets

After the process was completed we trained data on complex models live

A. Support vector machines
B. Artificial Neural Network
C. Random forest
D. Extreme Gradient Boosting

The output for running all those models are



Model	Data	ACC	ROC	Kappa
Logistic Regression	All data	88.04%	0.67	0.42
	Smote 1	81.52%	0.82	0.63
	Smote 2	79.25	0.79	0.58
	Fs via Backward elimination	86.41%	0.64	0.34
	Fs via CFS subset evaluation	85.32%	0.52	0.05
Decision Tree	All Data	84.23%	0.61	0.25
	Smote 1	85.57%	0.86	0.71
	Smote 2	92.22%	0.92	0.84
	Fs via Backward elimination	83.96%	0.64	0.31
	Fs via CFS subset evaluation	84.23%	0.61	0.25
Random Forest	All data	87.72%	0.59	0.26
ANN	All data	85.86%	0.64	0.33
SVM	All data	83.9%	0.58	0.19
XGBOOST	All data	88.04%	0.66	0.4
	Smote 1	91.41%	0.92	0.82
	Smote 2	87.35%	0.87	0.74

The interpretation is simple SMOTE data is the winner and decision tree stole the show!

But what are we missing at times it may be necessary to explain the HR our results, why our model believes our person is planning to leave, in that case decision tree or logistic regression are the best options.

However, if why decision made was not required we can simply use more powerful models like XGBoost, because as data increases the accuracy of XGBOOST will be higher than a decision tree because it’s a powerful ensemble algorithm.

Who made this – Mohammed

Who reviewed this – Patricia , Valerio

Clustering

For clustering our goal was to check if our dataset falls into 2 perfect clusters or not. If not what additional insights does it provide.

A K-means clustering algorithm was used on the original dataset.

1. The silhouette index tells us about the effectiveness of the classification and using the K-means (with two clusters) on the last set of variables, we obtained a 0.7 (Silhouettes average width).
2. We all pointed out that it was a very good result, plus upon reviewal of the work everyone correctly pointed that we are setting 2 as the number of cluster for testing matters, but the elbow plot at the beginning of the scripts shows that 3 or 4 is the best number to choose.
3. In order to explain such a high silhouette score, a reasoning was done to check why people in the clusters are likely to leave or not leave. Do they have something in common? That would have required further analysis on the clusters. And a future scope of research could be done on this data point

We wanted to work using clustering on this project because we wanted to improve our knowledge on the project and we still feel there is much to learn.

Who made this – Valerio

Who reviewed this – Sanchita , Mohammed

Challenges faced –

- The main difficulty was plotting. When it comes to clustering, plots become hard to code and interpret.
- Luckily enough all group members supported the clustering and a great repository of examples from this course.
- Ultimate struggle faced was also with time. A lot of time was spent learning and correct methods of coding. A lot of simple random bugs took up more time than it should have.

Conclusion of Author-

I wish to take this project out again in the future and try to answer the unanswered questions.

Association

Conclusion