HR Analytics Case Study

IDENTIFYING ATTRITION FOR XYZ…

Problem Statement:

XYZ facing high attrition among employees…..

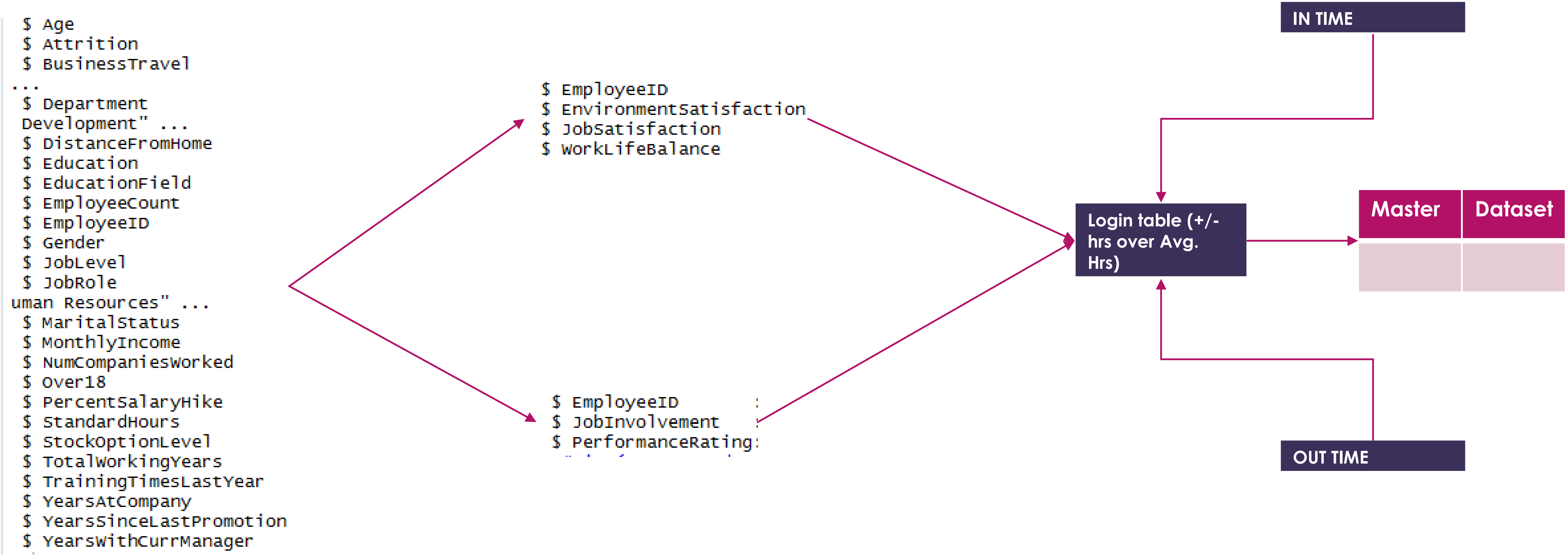
Objective:

MANAGEMENT WANTS TO UNDERSTAND IDENTIFY THE KEY DRIVERS/FACTORS THAT HAVE AN IMPACT ON EMPLOYEES LEAVING THE ORGANIZATION AND USE THIS TO BUILD AN ACTION PLAN TO RETAIN EMPLOYEES AND BRING DOWN ATTRITION

Approach:

- Logistic Regression Algorithm Model has been used on XYZ Employee data for the year 2015 .
- Collective data has general details of employees in company, employee satisfaction and manager survey data.
- Less/more than average hours spent in office is considered from exit and entry data.
- Model generated using above details give striking results.

The Data Summary:



```
 $ Age
 $ Attrition
 $ BusinessTravel
...
 $ Department
 Development" ...
 $ DistanceFromHome
 $ Education
 $ EducationField
 $ EmployeeCount
 $ EmployeeID
 $ Gender
 $ JobLevel
 $ JobRole
uman Resources" ...
 $ MaritalStatus
 $ MonthlyIncome
 $ NumCompaniesWorked
 $ Over18
 $ PercentSalaryHike
 $ StandardHours
 $ StockOptionLevel
 $ TotalWorkingYears
 $ TrainingTimesLastYear
 $ YearsAtCompany
 $ YearsSinceLastPromotion
 $ YearsWithCurrManager
```

```
$ EmployeeID
$ EnvironmentSatisfaction
$ JobSatisfaction
$ WorkLifeBalance
```

```
$ EmployeeID       :
$ JobInvolvement   :
$ PerformanceRating:
```

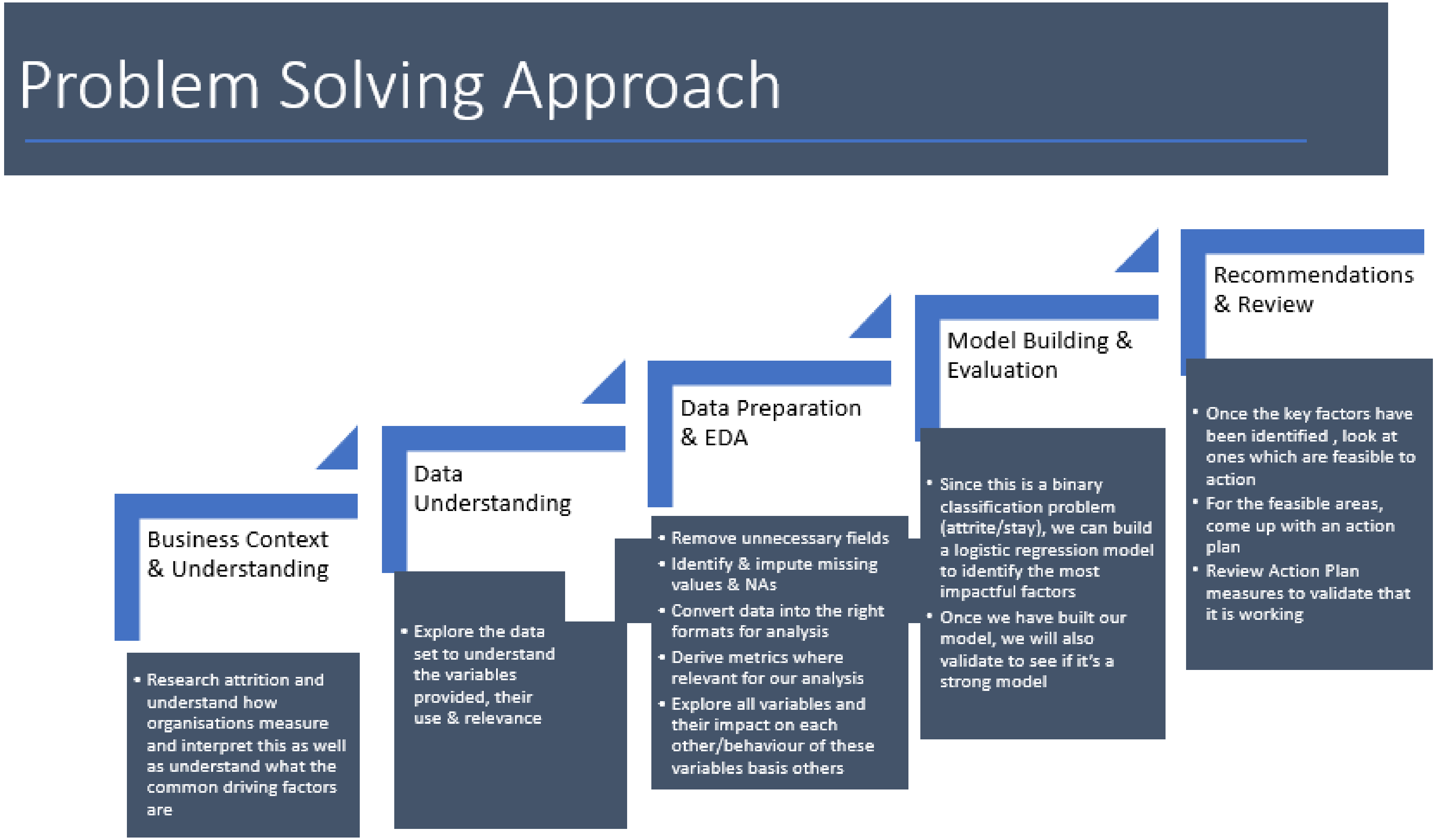
IN TIME

Login table (+/- hrs over Avg. Hrs)

OUT TIME

Master	Dataset

The Data Journey:



Problem Solving Approach

Business Context & Understanding

- Research attrition and understand how organisations measure and interpret this as well as understand what the common driving factors are

Data Understanding

- Explore the data set to understand the variables provided, their use & relevance

Data Preparation & EDA

- Remove unnecessary fields
- Identify & impute missing values & NAs
- Convert data into the right formats for analysis
- Derive metrics where relevant for our analysis
- Explore all variables and their impact on each other/behaviour of these variables basis others

Model Building & Evaluation

- Since this is a binary classification problem (attrite/stay), we can build a logistic regression model to identify the most impactful factors
- Once we have built our model, we will also validate to see if it's a strong model

Recommendations & Review

- Once the key factors have been identified , look at ones which are feasible to action
- For the feasible areas, come up with an action plan
- Review Action Plan measures to validate that it is working

Understanding and preparation

Business Understanding

- Attrition is a common challenge across many organisations and it is a core measure for HR as well as managers and department heads on how they are doing on the employee front
- Usual factors linked with attrition to name a few are years of experience, growth/promotions, job satisfaction, work life balance, travel time, managers worked with, performance ratings, utilisation and productivity, etc

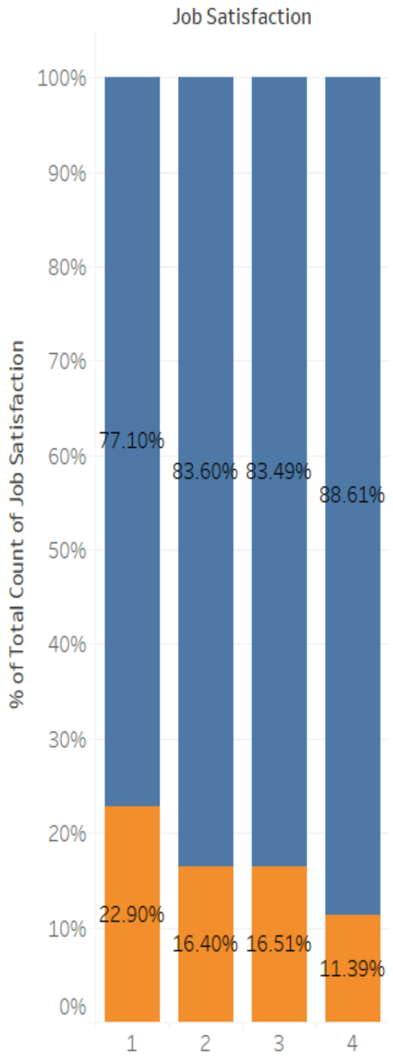
Data Preparation & Understanding

- The data provided has a good set of variables that we can work with, covering most of the factors above, plus employee and manager survey scores
- We also have access to daily timesheets for a year, which will help us identify the hours logged by employee and their days off
- Using this and standard hours provided, we can also calculated the average logged hours and the variance versus the expected hours to see if they are over/under-utilised
- We will also need to treat the data for outliers to get actionable and relevant insights
- Converting data formats to the required ones for analysis will also help us for our exploratory data analysis., as well as building our model

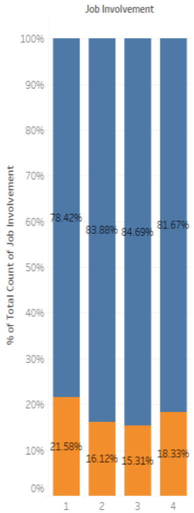
Lets take a closer look:

Longer login hours seem to drive attrition

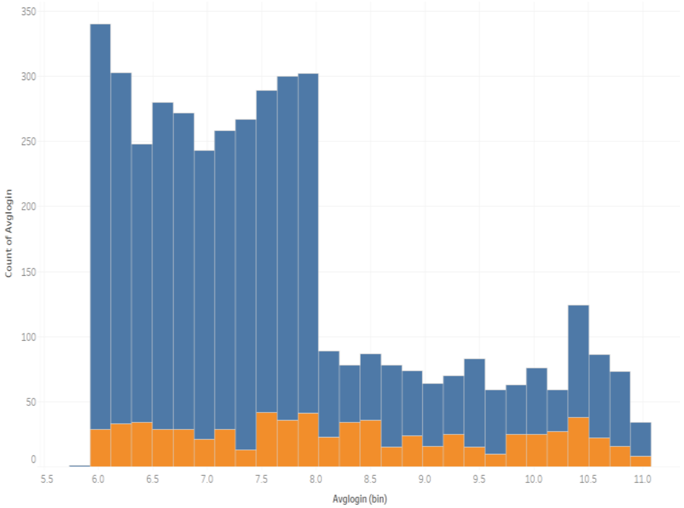
Job Satisfaction Rating Based Attrition



Job Involvement based Attrition



Average Login hours



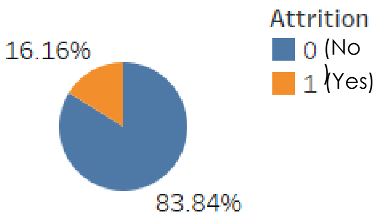
….continued..

Summary

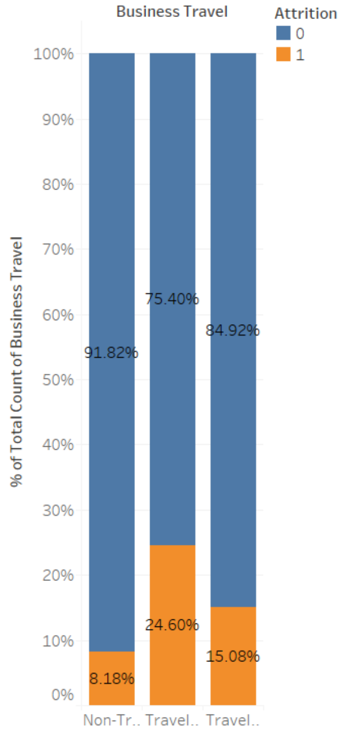
- As login hours go more than the standard 8 hours, we tend to see a higher rate of attrition
- Those who log lesser hours seem to have lesser attrition
- Data shows, in line with what you could expect, those who have low job satisfaction or job involvement have higher attrition rates
- However, we also see a higher than average attrition with those who are highly involved in their job

Around 16% of employees have attrited

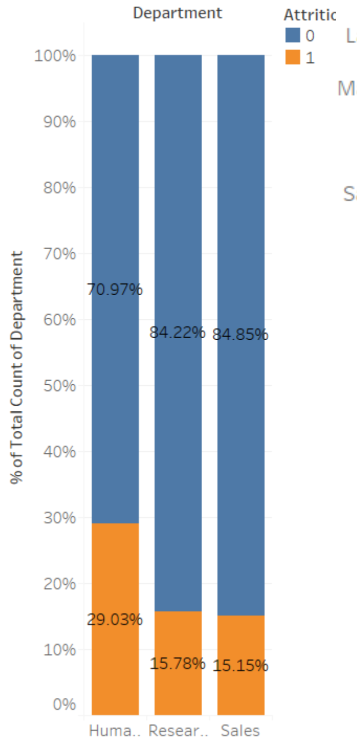
Attrition



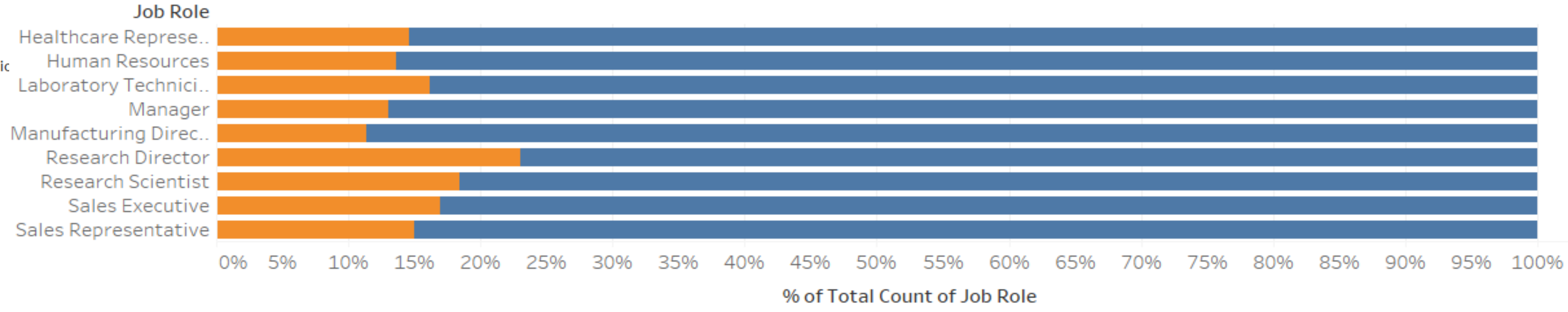
Attrition by Business Travel



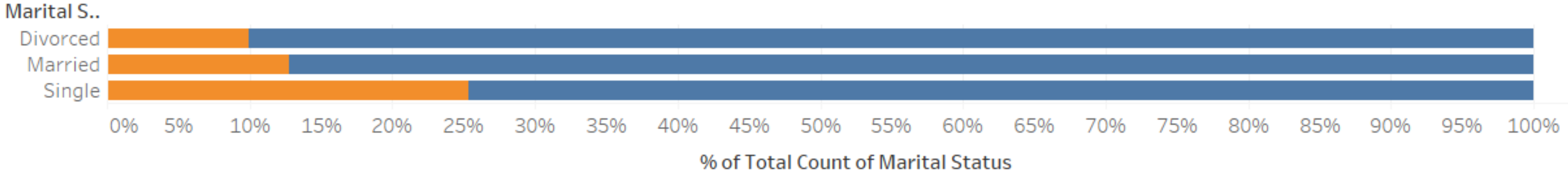
Attrition by Department



Attrition by Role



Attrition by Marital Status

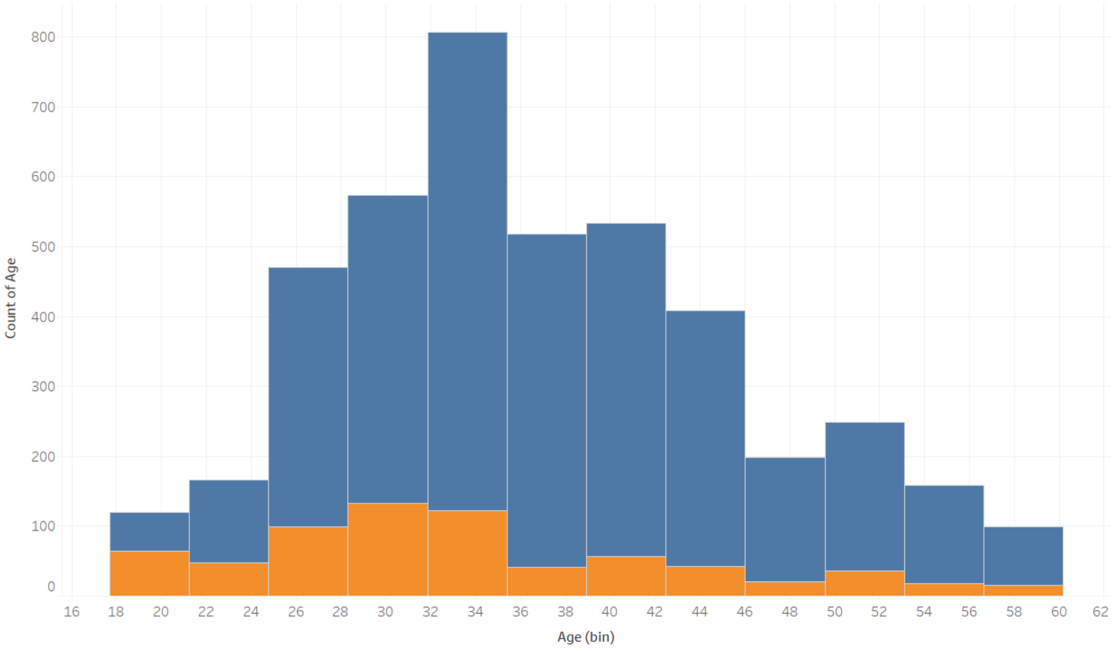


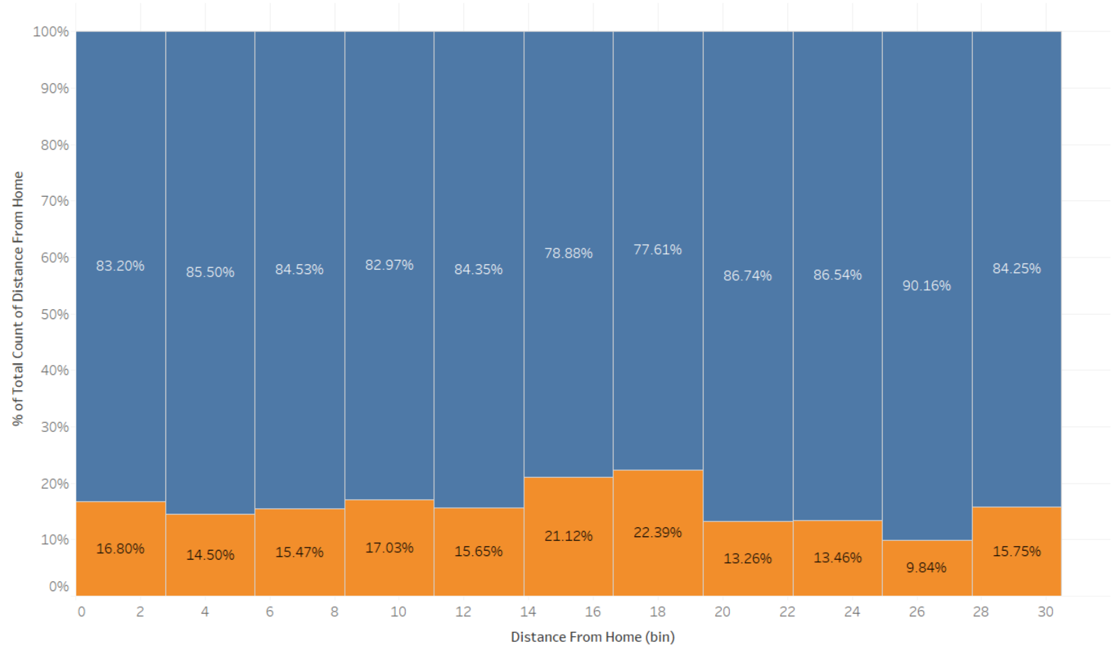
….continued..

Summary

- Attrition is around 16% in this data set
- We see a higher attrition by those who travel on business frequently
- Similarly those who work in Human Resources seem to be more likely to attrite though the overall numbers are lower as HR is a support function and has lesser employees
- We also notice that those working as Sales Executives, Research Directors and Research Scientist tend to have a higher rate of attrition, while those who are single also display a similar trend

Over 53% attrition for those between 18 and 21


Attrition by age


Attrition by distance

<u>Summary</u>

- We see that younger employees have a higher attrition rate

- 18-21 seems to be high risk with over 53% of employees leaving and then dropping to around 29% up to 25

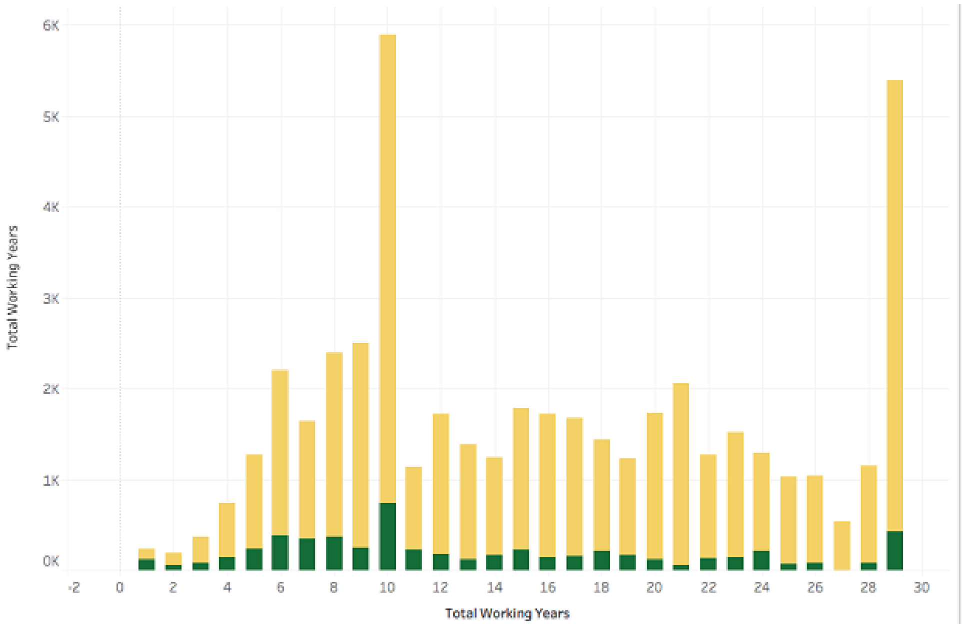
- We also see that those living 14 to 19kms from office have a higher than average attrition rate

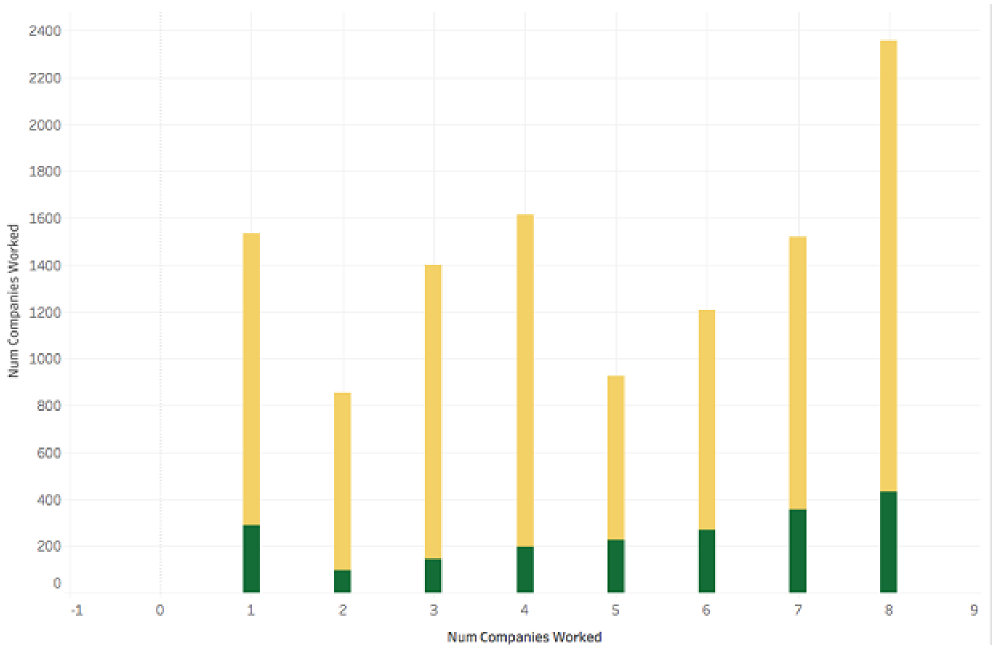
Less effective Factors:

Max 14% attrition for those people ~ 10years exp.

Iteration is found to be more among people with one company switch post
That it becomes stagnant.

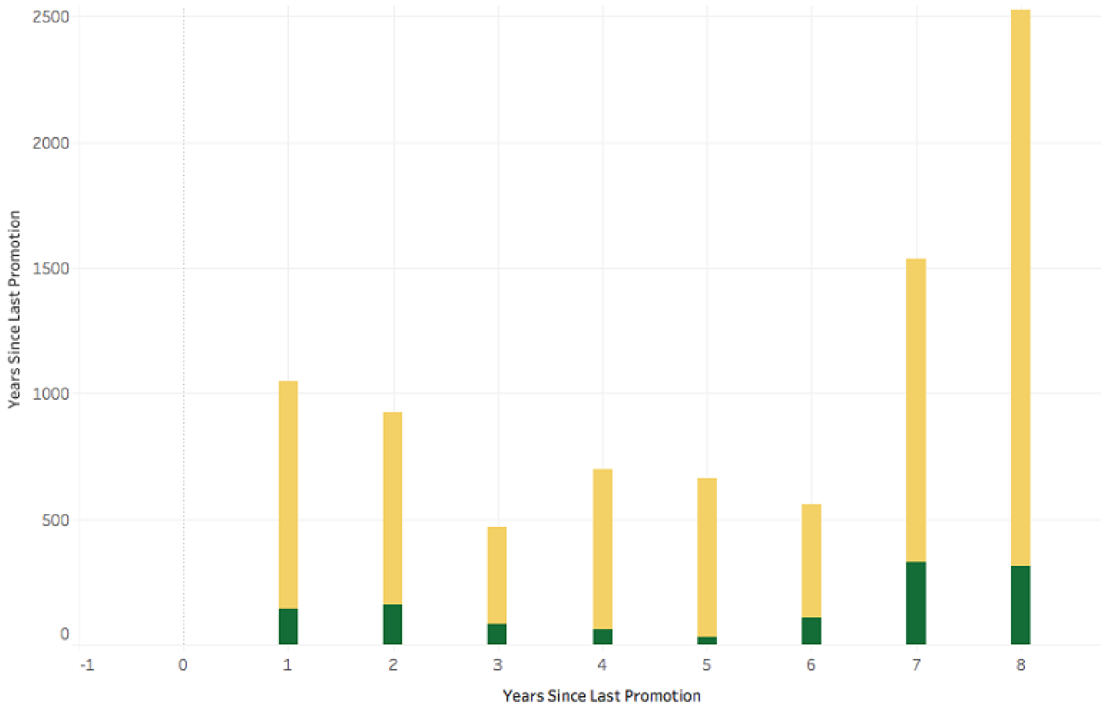
People in early stage career like 1 year tend to switch more. After that people
With (5-7) years of experience and around 10 years are more likely to leave the company.

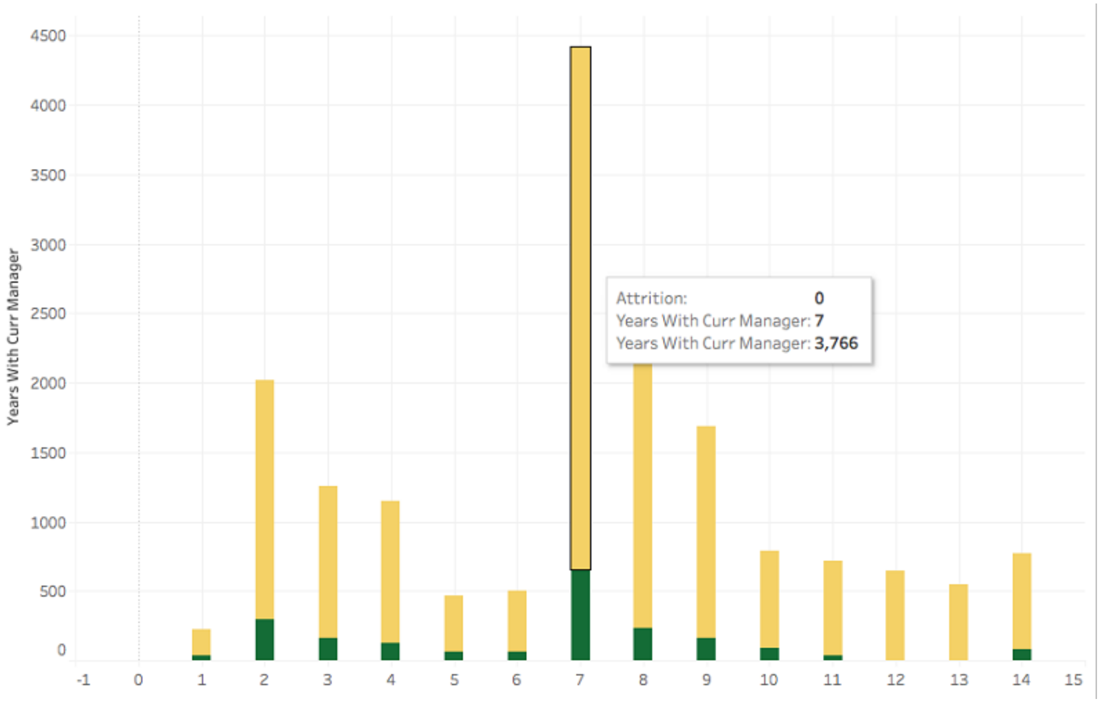




Over 14% attrition for those not promoted from last 8 years

High attrition found when less promotion opportunities and attrition decreases when Same manager is continued for employees.



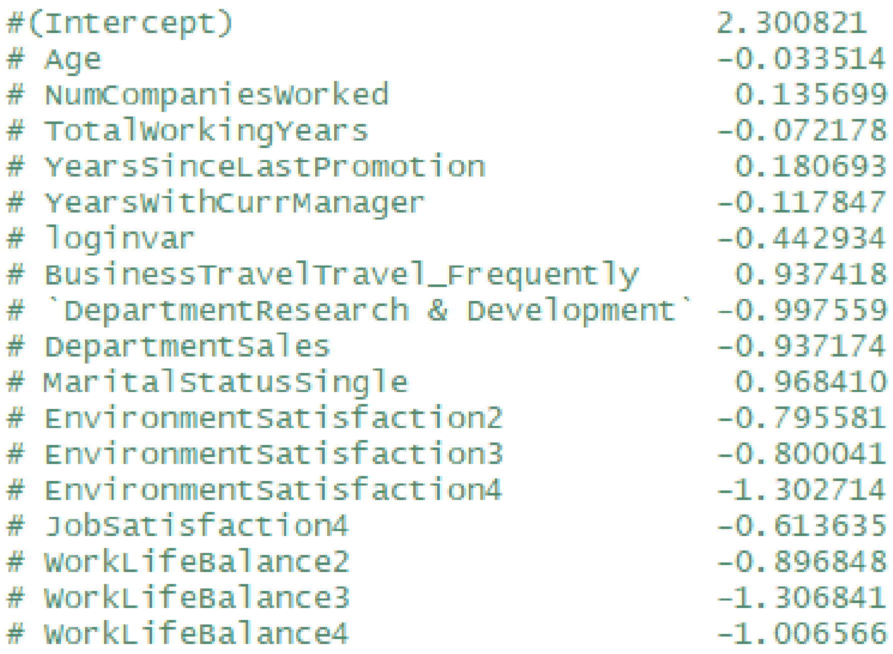


- Now that we know some of the indicators, we need to validate our observations and use a modelling technique to get statistically significant factors and quantify their impact

- We have also calculated some derived metrics basis some of the key metrics that seemed to impact attrition (e.g. – average login hours, variance to standard login hours, leaves, etc)

- Since this is a binary classification problem of either leaving or staying, we will use logistic regression

17 significant variables identified (12 categories)

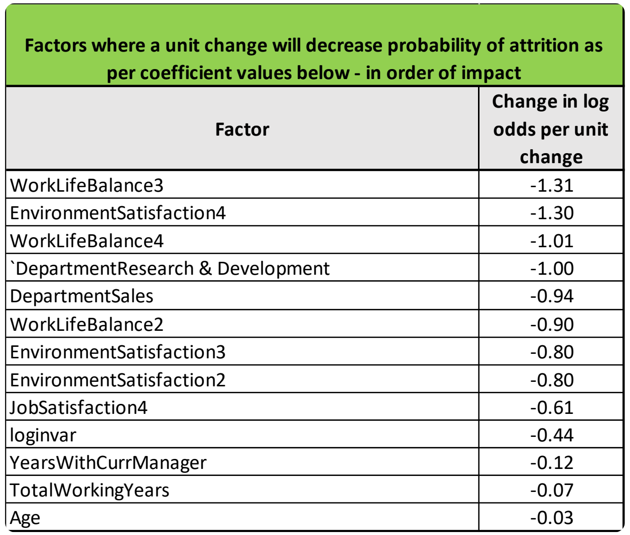


```
#(Intercept)                         2.300821
# Age                               -0.033514
# NumCompaniesWorked                 0.135699
# TotalWorkingYears                 -0.072178
# YearsSinceLastPromotion            0.180693
# YearsWithCurrManager              -0.117847
# loginvar                          -0.442934
# BusinessTravelTravel_Frequently    0.937418
# `DepartmentResearch & Development` -0.997559
# DepartmentSales                   -0.937174
# MaritalStatusSingle                0.968410
# EnvironmentSatisfaction2          -0.795581
# EnvironmentSatisfaction3          -0.800041
# EnvironmentSatisfaction4          -1.302714
# JobSatisfaction4                  -0.613635
# WorkLifeBalance2                  -0.896848
# WorkLifeBalance3                  -1.306841
# WorkLifeBalance4                  -1.006566
```

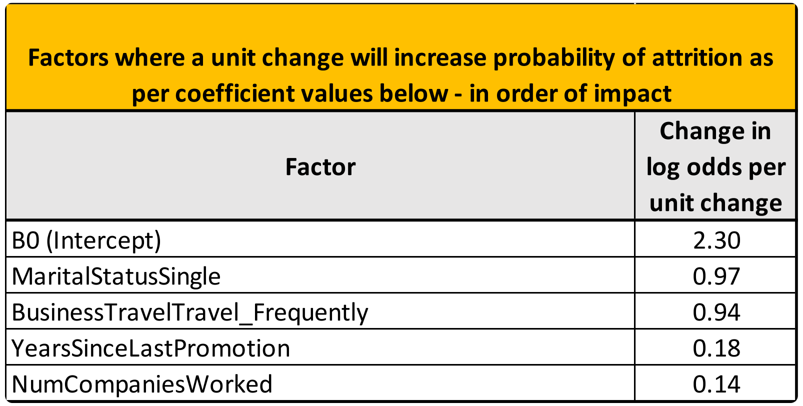
….continued..

- The logistic regression model gave us 17 variables which are statistically significant predictors of attrition which are actually 12 broad categories
- The 12 broad categories are Age, the number of companies an employee has worked with before, total work experience, time since last promotion, years with the current manager, Variance in login hours to standard hours, Business Travel, Department, Marital Status, Work Environment Satisfaction, Job Satisfaction and Work-life balance
- The chart below shows the specific 17 variables and the number to their right is the Beta coefficient
- This model is also validated and adjusted for measures of accuracy, specificity, sensitivity and has a strong KS statistic

Interpretation of the model



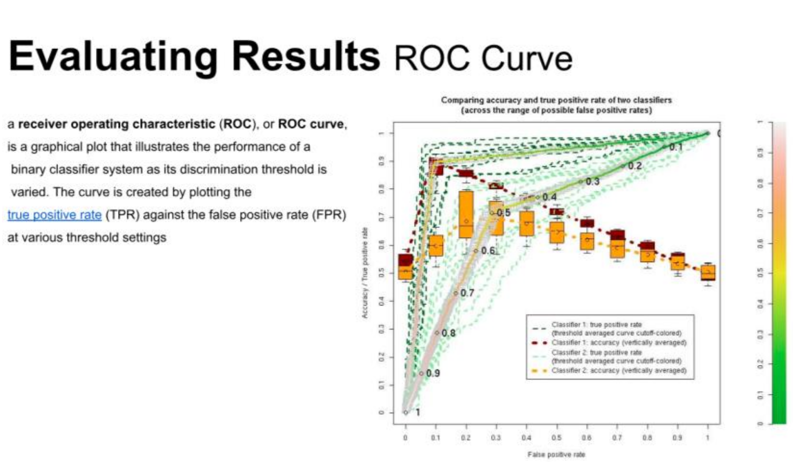
Factors where a unit change will decrease probability of attrition as per coefficient values below - in order of impact	
Factor	**Change in log odds per unit change**
WorkLifeBalance3	-1.31
EnvironmentSatisfaction4	-1.30
WorkLifeBalance4	-1.01
`DepartmentResearch & Development	-1.00
DepartmentSales	-0.94
WorkLifeBalance2	-0.90
EnvironmentSatisfaction3	-0.80
EnvironmentSatisfaction2	-0.80
JobSatisfaction4	-0.61
loginvar	-0.44
YearsWithCurrManager	-0.12
TotalWorkingYears	-0.07
Age	-0.03



Factors where a unit change will increase probability of attrition as per coefficient values below - in order of impact	
Factor	**Change in log odds per unit change**
B0 (Intercept)	2.30
MaritalStatusSingle	0.97
BusinessTravelTravel_Frequently	0.94
YearsSinceLastPromotion	0.18
NumCompaniesWorked	0.14

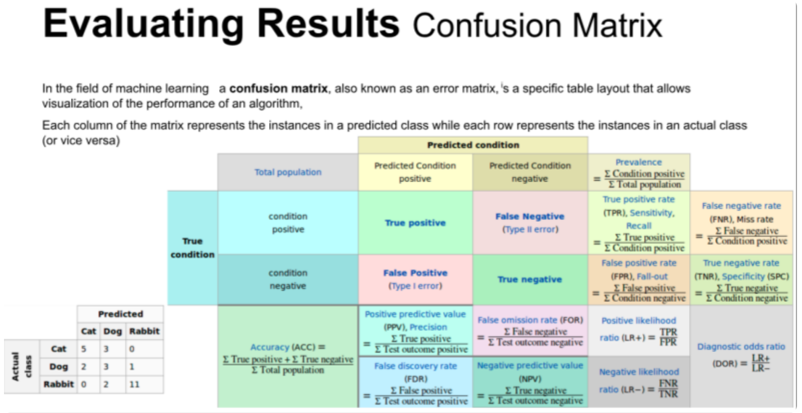
….continued..

- The negative coefficients indicate that a unit change in these will reduce the log odds of attrition while positive coefficients indicate an increase in log odds of attrition
- Number of companies an employee has worked for, the years since they were last promoted, Frequent Business Travel and Being Single (Marital Status) all have positive coefficients - indicates that these have a positive effect on attrition
- Age, total working years, years working under the current manager, being part of the R & D or sales dept, medium and above satisfaction with the work environment, Very high job satisfaction and a Good or above work-life balance rating have negative coefficients and reduce log odds of attrition
- Variance of login hours versus the standard hours also has a negative coefficient which means that if the login hours are less than standard hours, then it'll reduce log odds of attrition but if it more than standard hours, then there is a higher chance of attrition



Evaluating Results ROC Curve

a **receiver operating characteristic** (**ROC**), or **ROC curve**, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings



Evaluating Results Confusion Matrix

In the field of machine learning a **confusion matrix**, also known as an error matrix, 's a specific table layout that allows visualization of the performance of an algorithm,

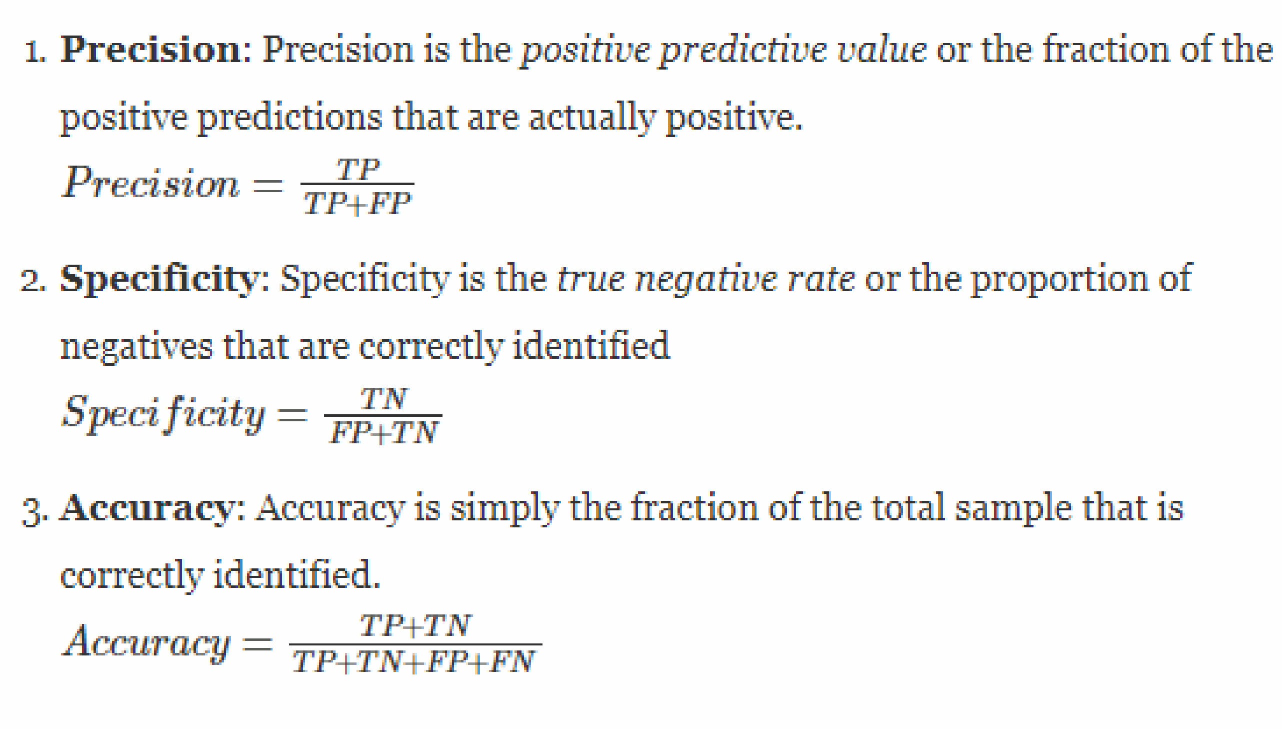
Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice versa)

			Predicted	
		Cat	Dog	Rabbit
Actual class	Cat	5	3	0
	Dog	2	3	1
	Rabbit	0	2	11

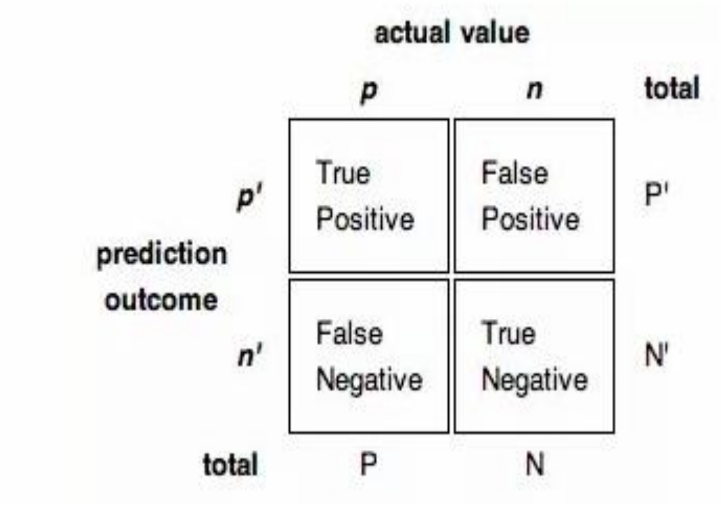
		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	**True positive**	**False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	**False Positive** (Type I error)	**True negative**	False positive rate (FPR), Fall-out $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
		False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$	

Evaluating Results:

Some Statistical terms…..



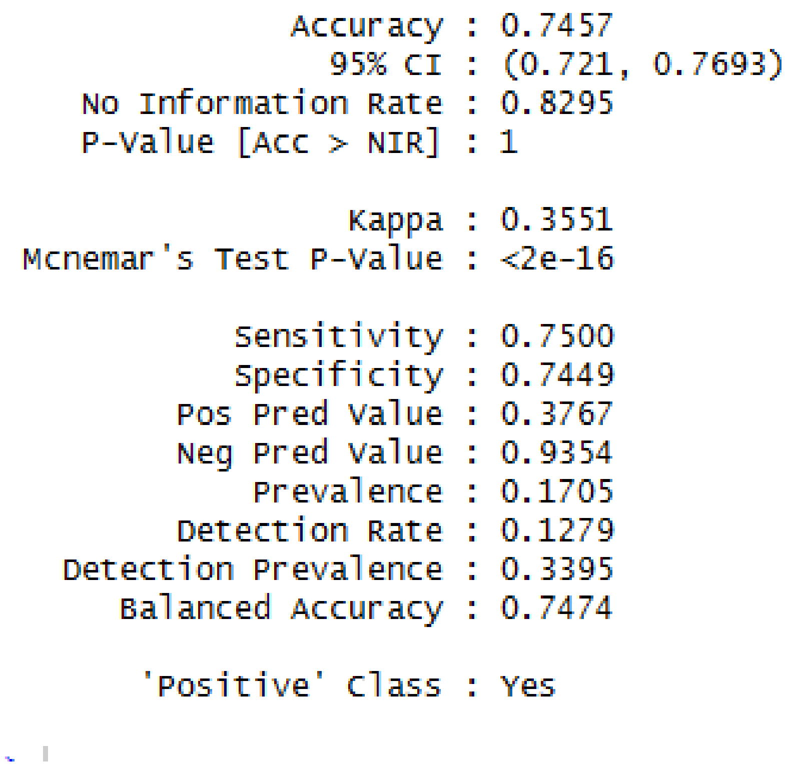
1. **Precision**: Precision is the *positive predictive value* or the fraction of the positive predictions that are actually positive.
   $$Precision = \frac{TP}{TP+FP}$$
2. **Specificity**: Specificity is the *true negative rate* or the proportion of negatives that are correctly identified
   $$Specificity = \frac{TN}{FP+TN}$$
3. **Accuracy**: Accuracy is simply the fraction of the total sample that is correctly identified.
   $$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$



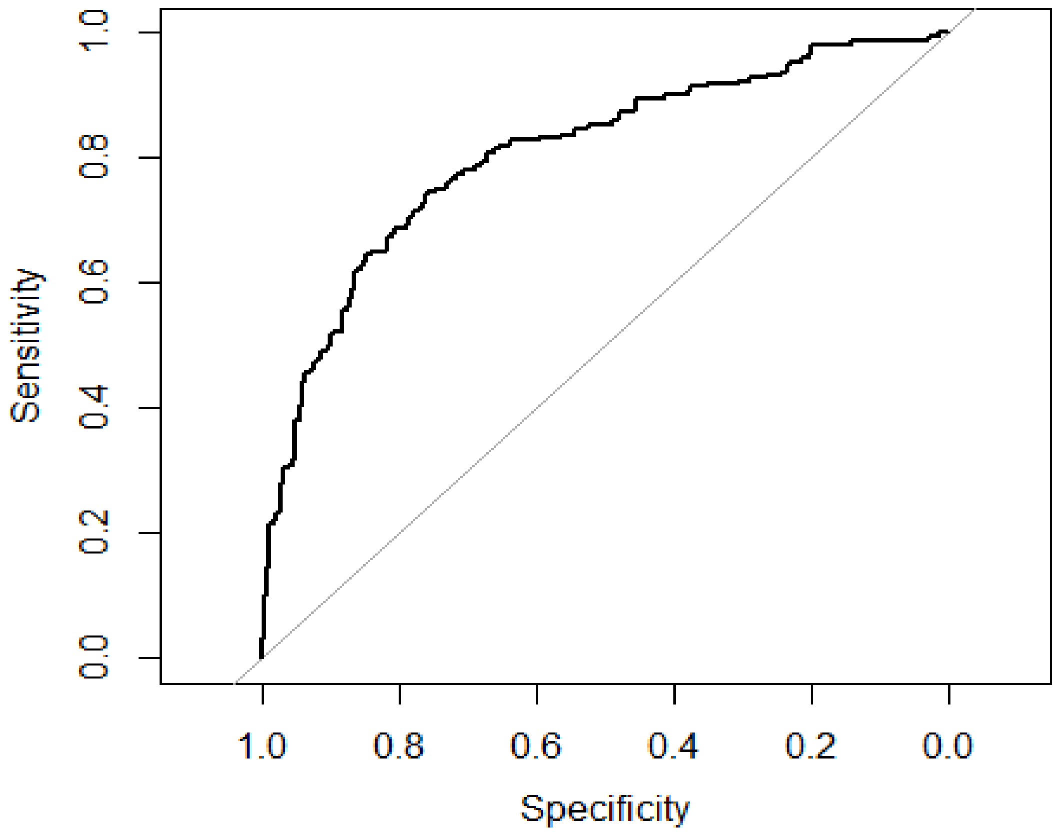
		actual value		
		p	*n*	total
prediction outcome	*p'*	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	total	P	N	

Precise Evaluation Measures for The Model:

The accuracy , sensitivity and specificity of the model looks good and the area under curve is 81%

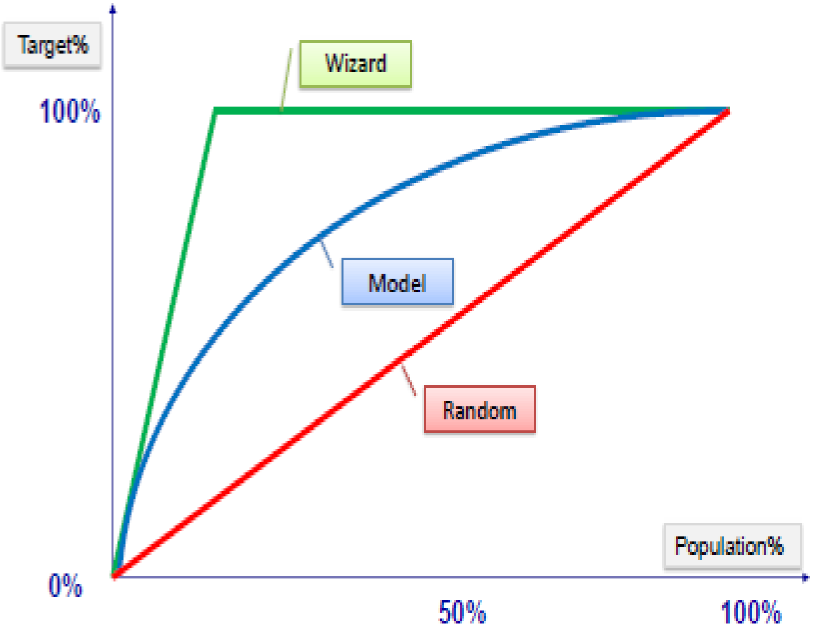


```
               Accuracy : 0.7457
                 95% CI : (0.721, 0.7693)
    No Information Rate : 0.8295
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3551
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7500
            Specificity : 0.7449
         Pos Pred Value : 0.3767
         Neg Pred Value : 0.9354
             Prevalence : 0.1705
         Detection Rate : 0.1279
   Detection Prevalence : 0.3395
      Balanced Accuracy : 0.7474

       'Positive' Class : Yes
```



Area under curve = 81%

….continued..



K-statistic - An indicator of how well your model discriminates between the random model and the perfect model.

- shows 49.4%

Gain chart – Factor by which your model is outperforming random model

$$\text{Lift} = \frac{\text{GainforCurrentModel}}{\text{GainforRandomMode}}$$

shows upliftment from 1 to 3.5

Recommendations for mitigation

- Hire candidates with more work experience than we currently do (automatically leading to increase in age of hires as well) as our model indicates that this will reduce the probability of attrition

- Build a stronger career development plan and potentially re-look at org structures to see if more relevant levels can be created so employees get promoted more often longer time between promotion increases the attrition risk

- Avoid too many manager changes and re-look at these reshuffles with the reasons to validate if they are really required. When employees spend more time with their current managers, they have a lower chance of leaving

- Put in place a real-time or daily tracking of employee utilisation and task managers with ensuring that employees are not stretched beyond standard login hours. Those who have to log in longer have a higher chance of leaving.

....continued..

- Those who have lower login hours also have a higher chance of staying so once the above is implemented and tracked for a while, if there is an opportunity to slightly cut down on required hours, this would be good to implement

- Frequent business travel has a significant impact on probability of attrition so it might be worth reviewing the need for business travel as well as making sure that the opportunity is split across employees so that some employees don't have to travel too much

- Conduct a survey to identify some specific improvement initiatives that employees want to see and then come up with actions on that feedback to drive improvement in the work environment. Employees with 'Medium' or above satisfaction have a lower chance of leaving and it increase as the rating increases

….continued..

- Some of the above actions on login hours tracking for better and more uniform utilisation along with some work from home policies, flexible working hours, holidays, etc would help drive work-life balance and this will help reduce chances of attrition

- Conduct detailed career discussions to ensure that the right people are mapped to the right jobs and also ensure uniform utilisation, etc as those with a high job satisfaction are more likely to stay back. A survey to understand what are the key factors for employees that have rated job satisfaction high will help in identifying and potentially replicating this for other employees

- Increase the mix of candidates who are married or separated in the overall mix as they have a lower probability of attrition

Thank -you