LOGISTIC REGRESSION CASE STUDY

ANUPAM MAJHI
HARI NYSHADAM
LIJO THOMAS
KIRAN VENKAT

OBJECTIVE

Use Logistic Regression to find the **probability of attrition**, for company 'XYZ' which employs about 4000 people at any given time.. The results thus obtained will be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay.

The data provided contains FIVE datasets, which have to be merged in order to understand all the data. The five datasets are,

01
General Data

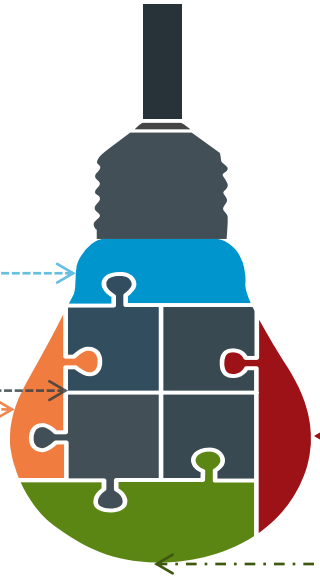
Contains work related metrics of every employee, like Age, Marital Status, Monthly Income, Number of Years Worked, etc

02
Manager Survey Data

Contains evaluation of an employee by their Manager. It contains two fields, Performance Rating and Job Involvement.

03
In Time

Contains the daily log-in time of employees for the duration given.



04
Out Time

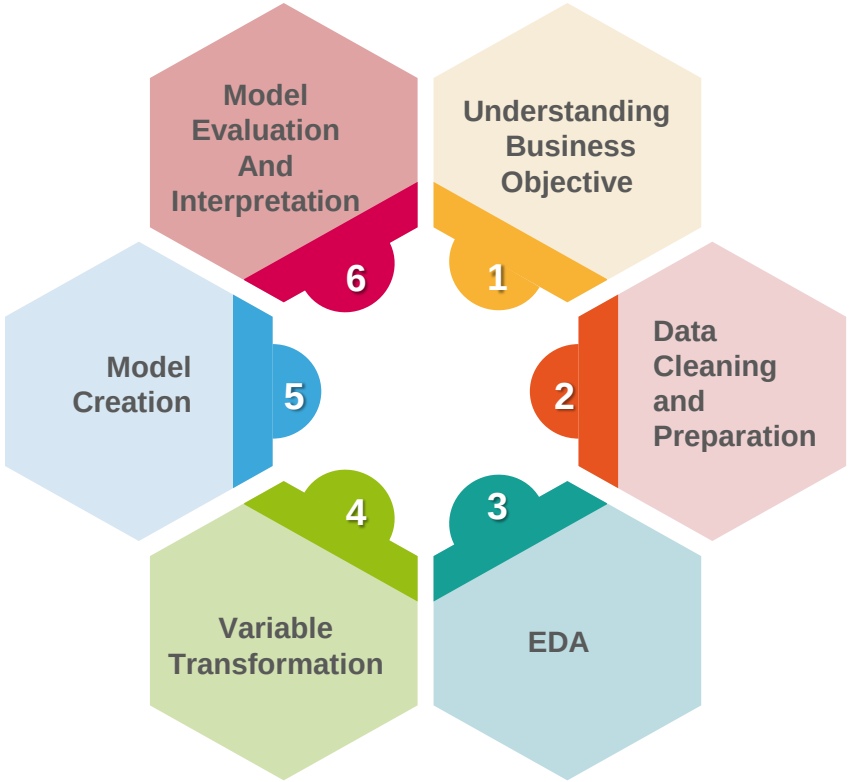
Contains the Log – Out time of employees for the duration given.

05
Employee Survey Data

Contains responses to a survey completed by all employees. It contains fields like Job Satisfaction, Work-Life Balance and Environment Satisfaction..

METHODOLOGY

Our methodology follows the **Cross Industry Standard Process for Data Mining** (CRISP–DM) framework. The process and the steps involved can easily be visualized in the following info graphic

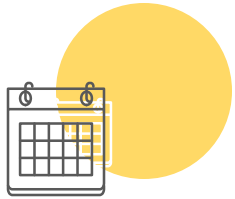


DATA CLEANING

AND

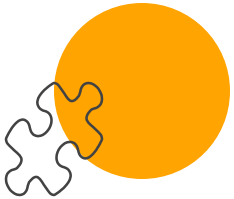
PREPARATION

DATA CLEANING



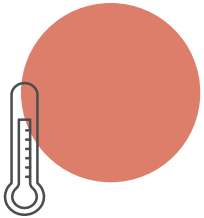
DATE - TIME

The in time and out time datasets were converted to correct date time format. Additional metrics like **Average hours,Leaves Taken**, and **Overtimes** done have been calculated for each employee.



MISSING VALUES

There were 111 missing values in the dataset. Survey data missing values were imputed with mean value of that category . Continuous variable missing values were imputed with the median value.



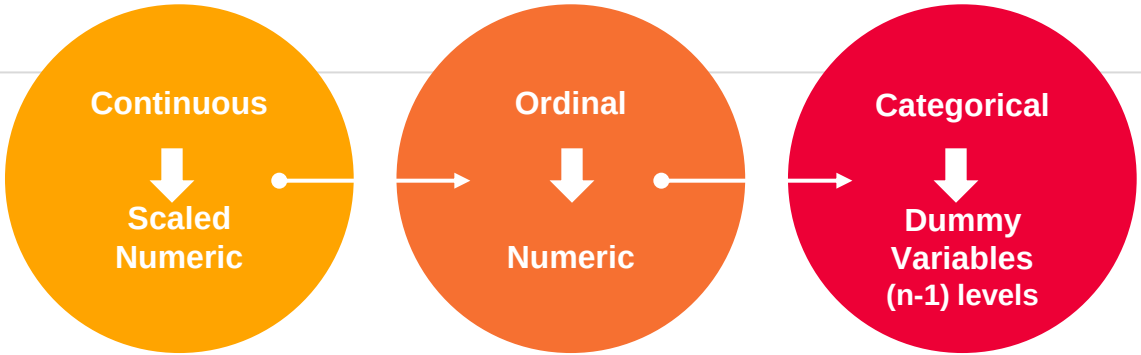
OUTLIERS

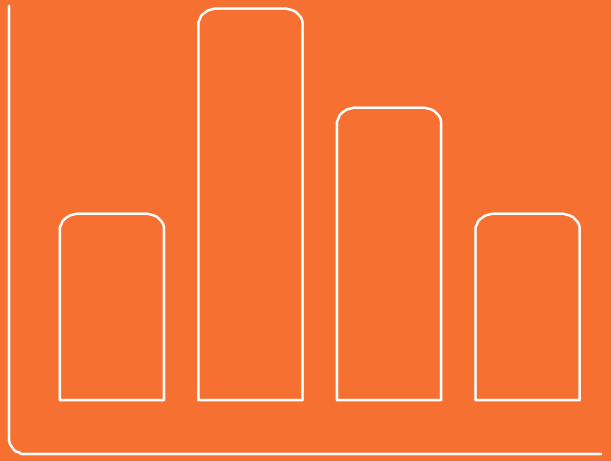
Outliers were detected in the dataset. The outliers were capped with the appropriate values.

VARIABLE TRANSFORMATION

Before modelling the variables have to be transformed into the right format, so as to keep the model stable and ensure efficient prediction. This is the process we have followed,

- Continuous variables are kept as numeric. They are scaled using the scale function so as to ensure the same scale across all variables
- Ordinal variables are kept as numeric
- Categorical variables are converted into dummy variables using model.matrix command.





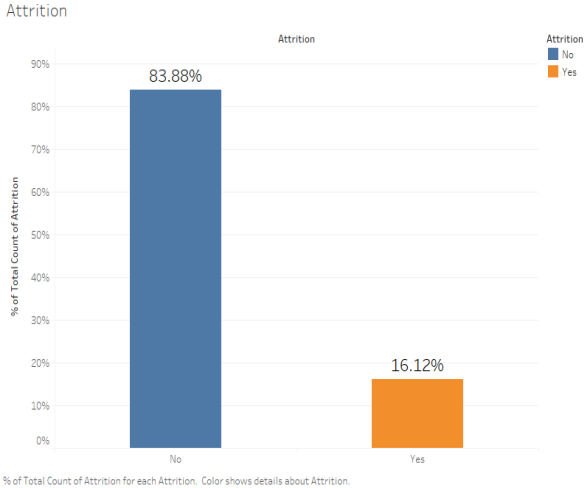
EXPLORATORY DATA ANALYSIS

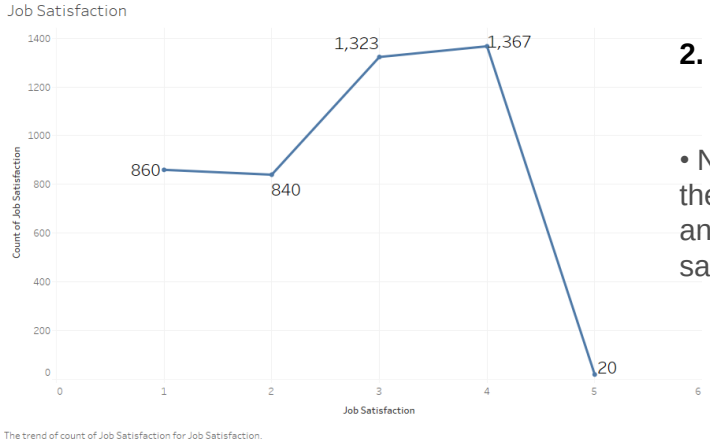
Univariate Analysis

1. Attrition

•Overall Attrition is 16%

• Approximately 700 people leave the company every year





2. Job Satisfaction

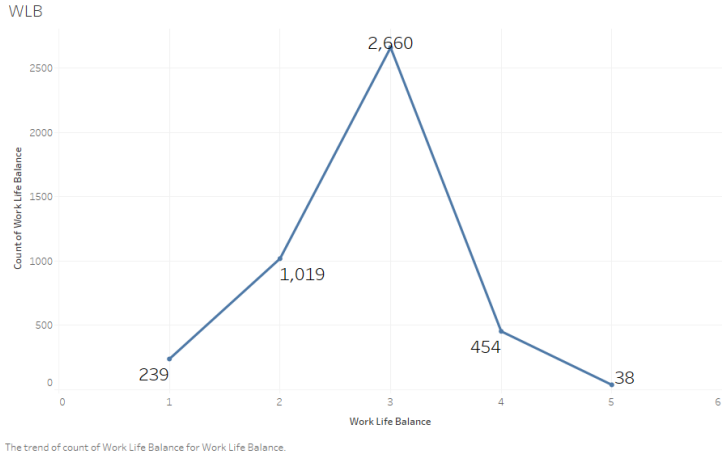
• Nearly 60% of all employees scored their satisfaction with their job between 3 and 4 indicating that were relatively satisfied with their jobs.

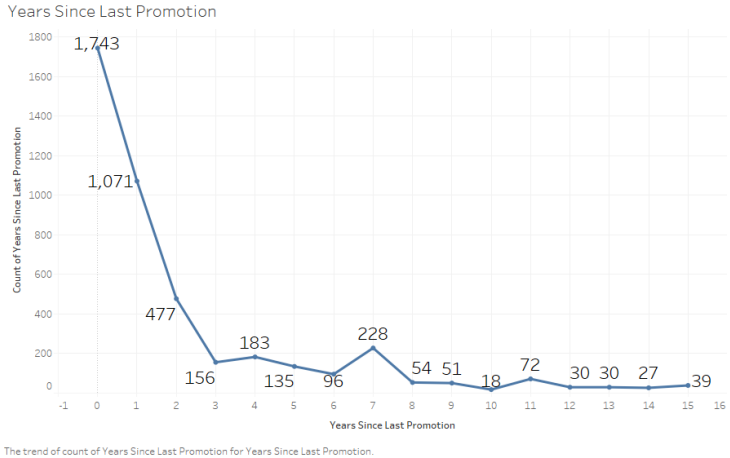
Univariate Analysis

3. Work Life Balance

•Nearly 88% of employees scored their Work – Life Balance below 3

• The most scored rating was 3





4. Years Since Last Promotion

•Surprisingly we have several employees who have not been promoted in more than 12 years.

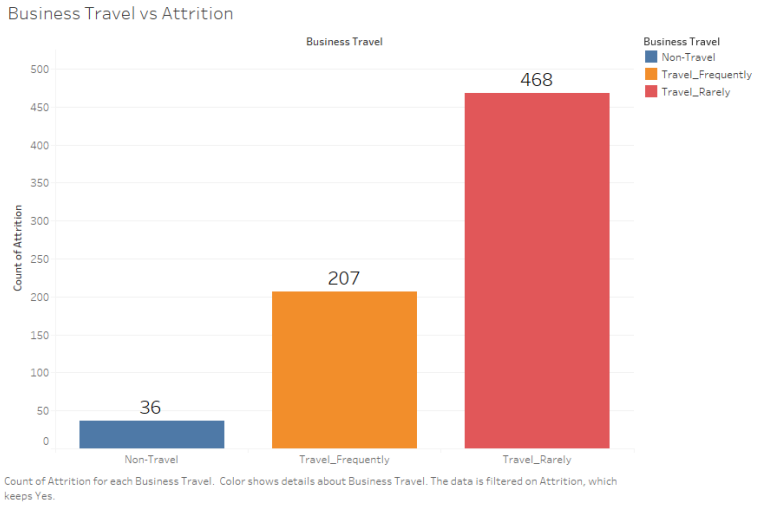
• The maximum number of years for which an employee has not been promoted is 15 years.

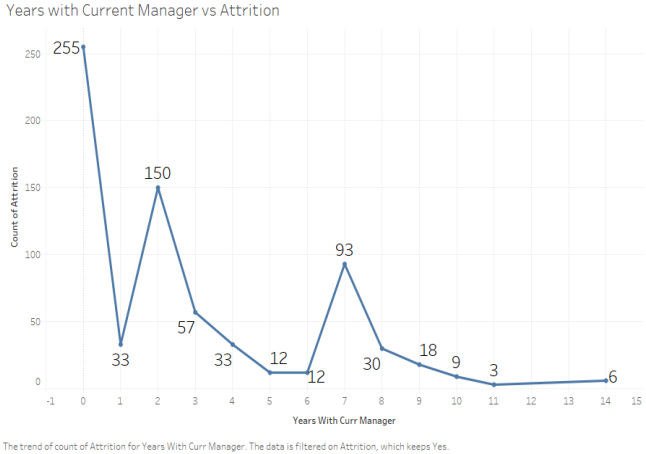
• About 25% of employees do not get promoted for three years

Bivariate Analysis

1. Business Travel vs Attrition

• The Highest Attrition occurs for those employees that travel very rarely





2. Years With Current Manager vs Attrition

•The highest attrition occurs for those employees who have been with their managers for less than a year.

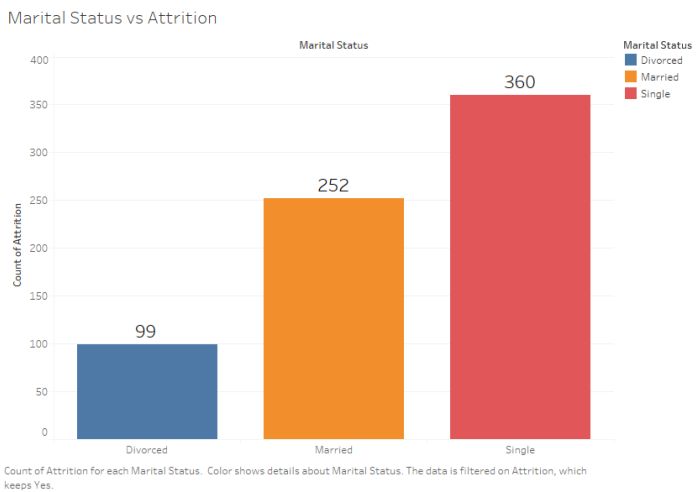
• Attrition rates generally decrease as the number of years with current manger increases.

Analysis

3. Marital Status vs Attrition

•Employees whose Marital status is *Single* , tend to contribute to higher attrition

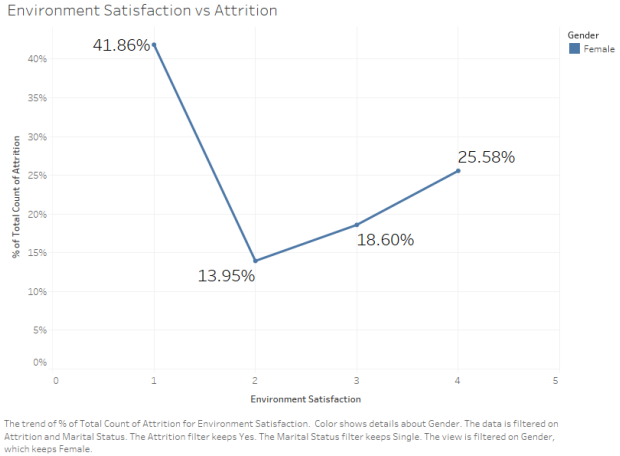
•Employees that are Married or Divorced tend to stick with the company longer



4. **Marital Status –** Single
Gender - Female

•An interesting thing to note is that , out of all the Single, Female employees that attrite, nearly 50% of them score their environment satisfaction as 1.

• This could be an easy indicator of attrition. More importantly, 50% of Single Female Employees feeling their environment is uncomfortable is a **RED** Flag, and a major cause for concern



MODELING THE DATA

Model Creation

•The final data set after variable preparation has 42 independent variables .

•The GLM has been built on this data set. The first model has AIC of 2099

•The many variables which have high VIF and low significance have been removed iteratively, and with the remaining variables, a new model is created .

•This is repeated 14 times until all the variables are significant and their VIF is low .

•Our final model has AIC of 2113

•The Model can be viewed with the intercept values on the right hand side of the slide.

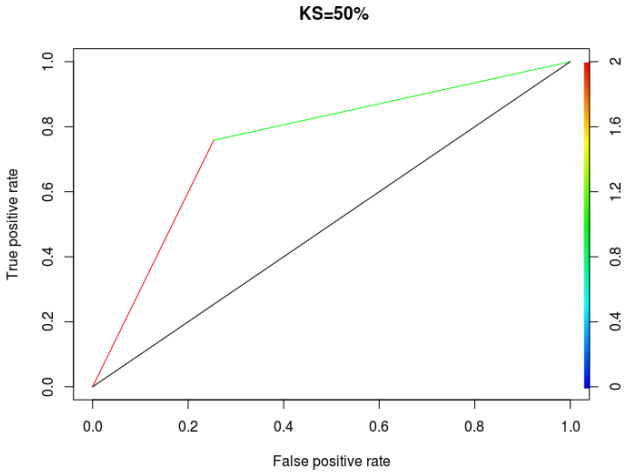
Coefficients:

	Estimate
(Intercept)	-0.36057
BusinessTravel	0.84086
EnvironmentSatisfaction	-0.40736
JobSatisfaction	-0.37267
WorkLifeBalance	-0.33036
JobRole.xManufacturing.Director	-0.92259
MaritalStatus.xSingle	0.95396
Age	-0.37371
NumCompaniesWorked	0.41384
TotalWorkingYears	-0.50362
YearsSinceLastPromotion	0.60549
YearsWithCurrManager	-0.54495
overtime_count	0.79147

Model Evaluation

To predict using our model, we need to set the cut off . Cutoff depends on the business scenario. We will settle at **0.2** since that gives us better *Sensitivity* and *Specificity* , and has a good Balanced *Accuracy*. Depending on the business requirement changes, we can vary the cutoff value to suit the needs

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  875  55
       Yes 241 152

               Accuracy : 0.7763
                 95% CI : (0.7528, 0.7985)
    No Information Rate : 0.8435
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3795
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7343
            Specificity : 0.7841
         Pos Pred Value : 0.3868
         Neg Pred Value : 0.9409
             Prevalence : 0.1565
         Detection Rate : 0.1149
   Detection Prevalence : 0.2971
      Balanced Accuracy : 0.7592

       'Positive' Class : Yes
```



•KS Statistic = **0.505**

•Area under the curve(AUC) is **0.7524349** , where 1 is perfect model and 0.5 is random model

•K-Fold Cross-validation estimate of accuracy is **0.862** . We see that the accuracy is quite high after 10 folds, hence we conclude that the model is quite stable

Suggestions

• **Business Travel** : We find that employees that travel frequently and those that travel very rarely tend to leave the company. the fastest. Therefore opportunities must be given to employees that travel rarely , while simultaneously reducing the travel load of frequent travelling employees.

• **Marital Status-Single** : Those employees that are single generally have no dependencies. As a result they are always on the move, looking for the next big opportunity. Therefore , steps should be taken to entice these employees to stay longer by means of incentives, stock options, etc

• **Number of Companies Worked** : This is one of the clearest indicators of Attrition. Employees that have a huge list of companies worked for, generally tend to stay for a short time. Hiring such employees should be considered a risky proposition.

• **Years since Last Promotion** : We find that as the number of years since last promotion increases, rate of Attrition increases. Therefore, care must be taken to ensure that years between promotions is fewer.

• **Overtimes** : Employees that are consistently overworked tend to leave the fastest. While looking at the data, there are numerous instances of employees consistently working overtimes. Additional resources must be provided for project completion and measures taken to ensure that employees are never unnecessarily overworked.

The following Factors negatively affect Attrition,

•**Environment Satisfaction, Job Satisfaction, Work Life Balance** : Improving these three factors, reduces Attrition.

•**Age , Total Working Years** : As Employees get older, they tend to settle down and resist the need to change. This reduces their Attrition rate dramatically.

•**Years with Current manager** : Employee that generally have spent long periods of time with their current manager, tend to have lower Attrition rates.. Therefore, care should be taken to ensure that managers are not frequently changed or replaced.