HR Analytics CASE STUDY

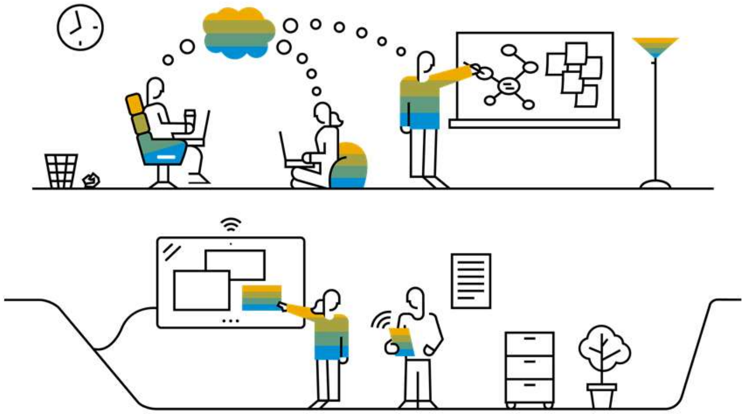
Maxim Rohit

Abhishek Ranjan

Yogesh Kulkarni

Srivatsan Santhanam

Problem Statement



A large company named XYZ, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition(employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons -

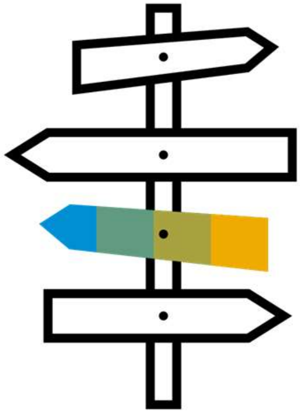
The former employees' projects get delayed, which makes it difficult to meet timelines, resulting in a reputation loss among consumers and partner. A sizeable department has to be maintained, for the purposes of recruiting new talent

More often than not, the new employees have to be trained for the job and/or given time to acclimatize themselves to the company

Hence, the management has contracted an HR analytics firm

1. **To understand what factors they should focus on, in order to curb attrition.**
2. **what changes they should make to their workplace, in order to get most of their employees to stay**
3. **which of the variables is most important and needs to be addressed right away**.

Approach



Logistic Regression technique approach was followed

- Data Understanding
-

- Data Preparation & EDA

- Model Building
  - Separate Data into Train and Test
    - Ensuring Test data has similar distribution to the data set rpovided
  - Use STEP AIC and model

- Model Evaluation
  - Accuracy, Sensitivity, and Specificity
  - KS Static
  - Lift and Gain chart

- Boot strapping to test the stability of the model

Data Understanding

employee_survey_data has details of the Work Environment Satisfaction Level, Job Satisfaction Level and Work life balance level for each employee

manager_survey_data has data of the Job Involvement Level & Perf rating for last year as given by the manager.

general_data has data of a number employee attributes. There are 10 Continuous Variables & 14 Categorical Variables in the general_data dataset.

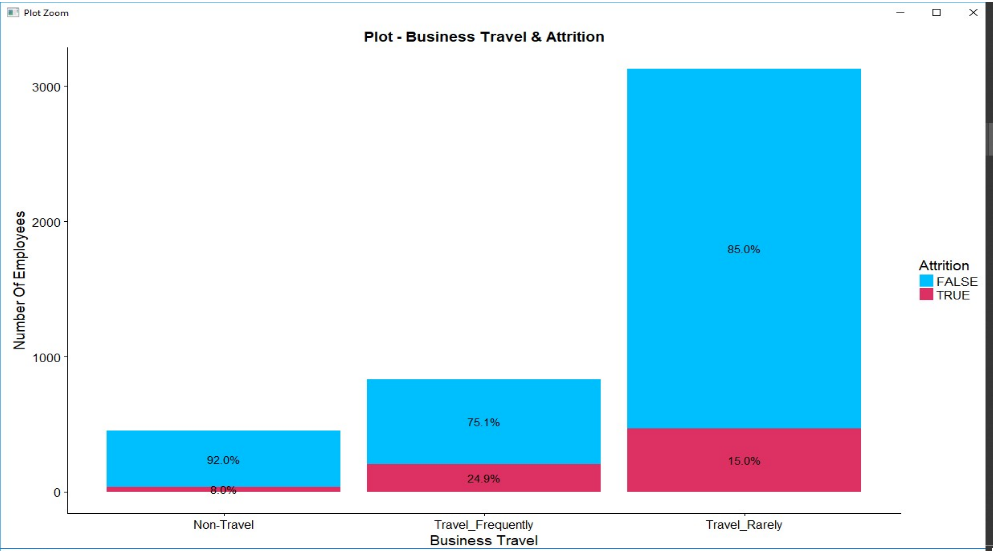
in_time and out_time : in_time gives the date & time when employee enters the office. out_time dataset gives the date & time when employee leaves the office.

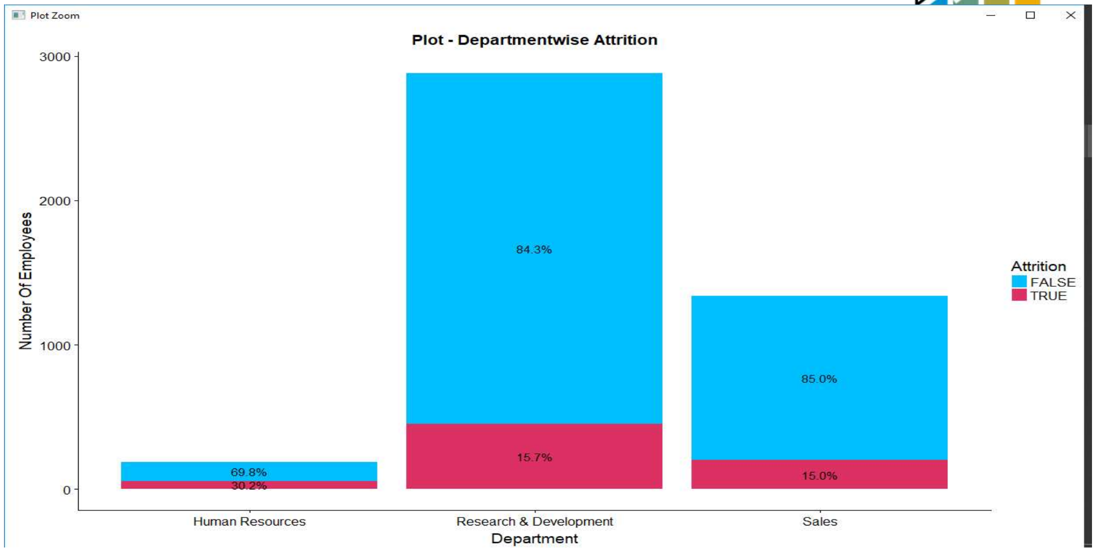
Data Preparation

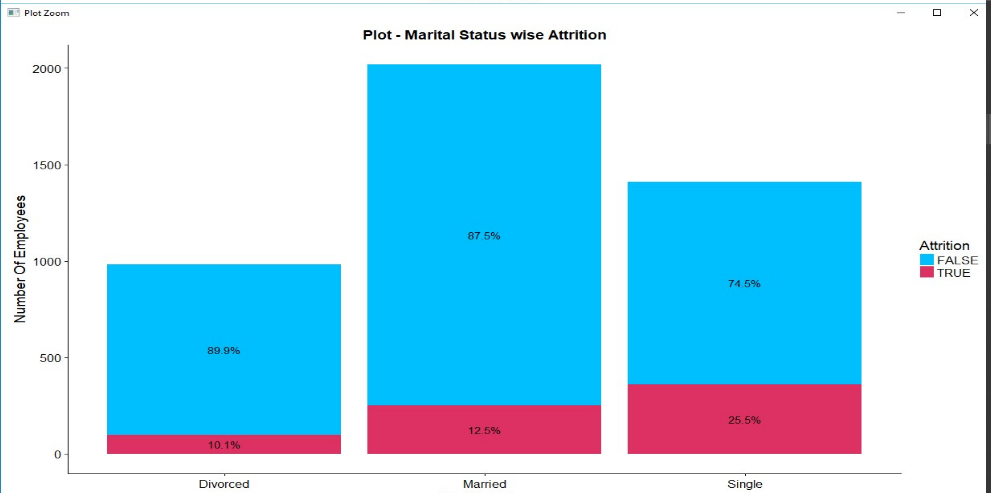
Following were done as part of Data preparation step on the DFs

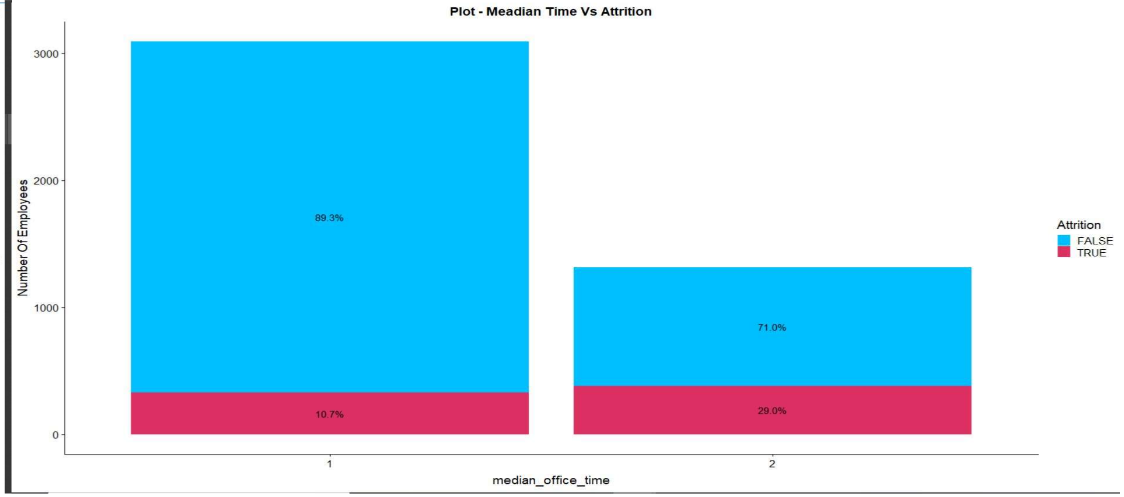
- Check for missing values
- Data format handling
- Check for NA: replace with WoE (TotalWorkingYears, NumCompaniesWorked, EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance)
- ColName Enrichment where ever missing
- Derived Columns: Daily work hours, Leave Count, Income to expr ratio, Marital status combined with Gender
- Removal of insignificant/empty Cols
- Binning of Continuous Variables
- Ordinal Categorical Variable handling via dummy variable

Attrition Vs Independent Variables

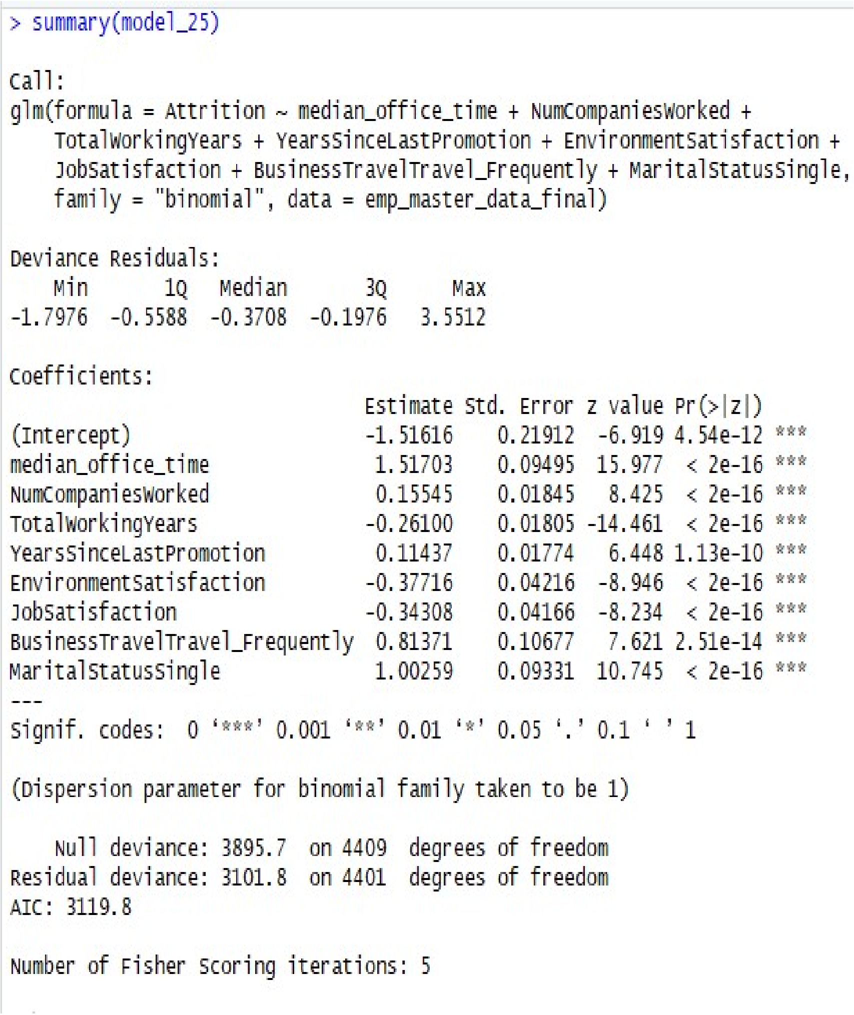








Model Building



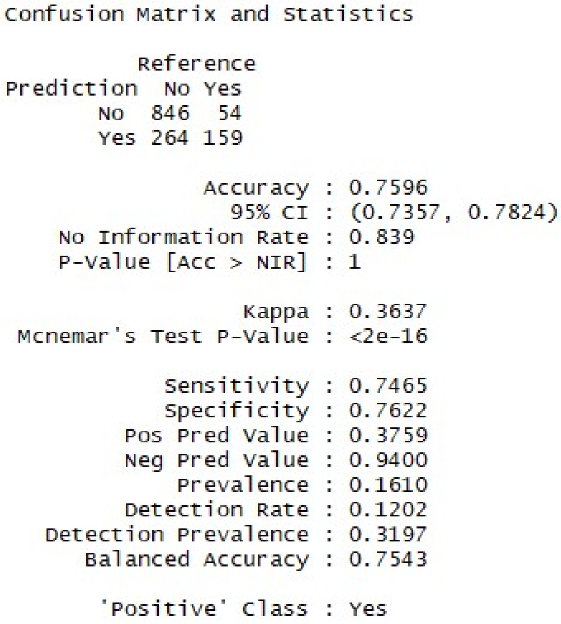
```
> summary(model_25)

Call:
glm(formula = Attrition ~ median_office_time + NumCompaniesWorked +
    TotalWorkingYears + YearsSinceLastPromotion + EnvironmentSatisfaction +
    JobSatisfaction + BusinessTravelTravel_Frequently + MaritalStatusSingle,
    family = "binomial", data = emp_master_data_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7976  -0.5588  -0.3708  -0.1976   3.5512

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.51616    0.21912  -6.919 4.54e-12 ***
median_office_time               1.51703    0.09495  15.977  < 2e-16 ***
NumCompaniesWorked               0.15545    0.01845   8.425  < 2e-16 ***
TotalWorkingYears               -0.26100    0.01805 -14.461  < 2e-16 ***
YearsSinceLastPromotion          0.11437    0.01774   6.448 1.13e-10 ***
EnvironmentSatisfaction         -0.37716    0.04216  -8.946  < 2e-16 ***
JobSatisfaction                 -0.34308    0.04166  -8.234  < 2e-16 ***
BusinessTravelTravel_Frequently  0.81371    0.10677   7.621 2.51e-14 ***
MaritalStatusSingle              1.00259    0.09331  10.745  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3895.7  on 4409  degrees of freedom
Residual deviance: 3101.8  on 4401  degrees of freedom
AIC: 3119.8

Number of Fisher Scoring iterations: 5
```

Model 25 have comparable sensitivity,accuracy and specifity with minimum set of variable

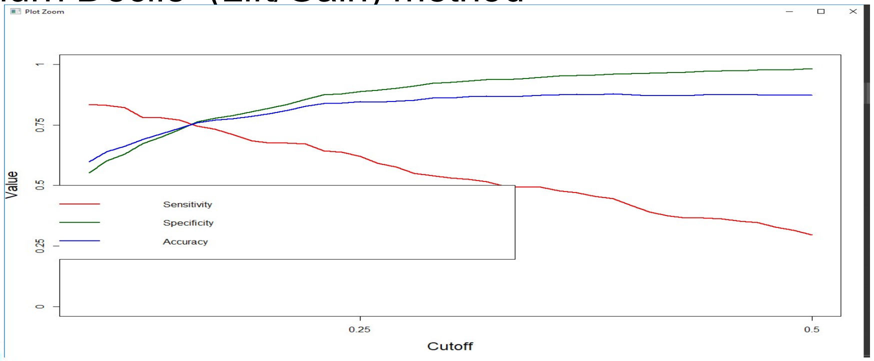
	Accuracy	Sensitivity	Specificity	Variable Count
Model 22	0.8715042	0.2863850	0.9837838	11
Model 23	0.8722600	0.2816901	0.9855856	10
Model 24	0.8745276	0.2957746	0.9855856	9
Model 25	0.8722600	0.2957746	0.9828829	8
Model 26	0.8662132	0.2676056	0.9810811	7

Model Evaluation

Based on the final model on test data and choosing a cut-off value of 0.16 for final model, we have the following: (KS Statistic and Churn Decile (Lift/Gain) method



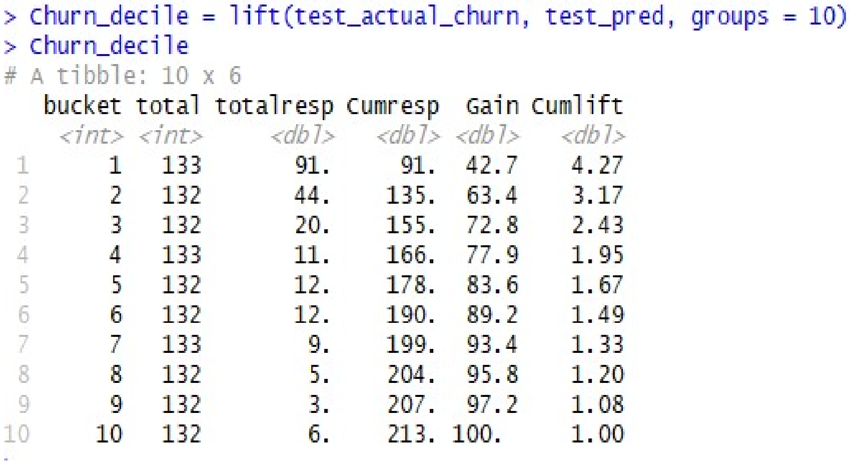
```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  846  54
       Yes 264 159

               Accuracy : 0.7596
                 95% CI : (0.7357, 0.7824)
    No Information Rate : 0.839
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3637
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7465
            Specificity : 0.7622
         Pos Pred Value : 0.3759
         Neg Pred Value : 0.9400
             Prevalence : 0.1610
         Detection Rate : 0.1202
   Detection Prevalence : 0.3197
      Balanced Accuracy : 0.7543

       'Positive' Class : Yes
```



Red- Sensitivity

Green – Specificity

Blue - Accuracy



```
> Churn_decile = lift(test_actual_churn, test_pred, groups = 10)
> Churn_decile
# A tibble: 10 x 6
   bucket total totalresp Cumresp  Gain Cumlift
    <int> <int>     <dbl>   <dbl> <dbl>   <dbl>
 1      1   133       91.     91.  42.7    4.27
 2      2   132       44.    135.  63.4    3.17
 3      3   132       20.    155.  72.8    2.43
 4      4   133       11.    166.  77.9    1.95
 5      5   132       12.    178.  83.6    1.67
 6      6   132       12.    190.  89.2    1.49
 7      7   133        9.    199.  93.4    1.33
 8      8   132        5.    204.  95.8    1.20
 9      9   132        3.    207.  97.2    1.08
10     10   132        6.    213. 100.     1.00
```

KS Statistic: 0.50864(50.86%)

Inferences & Recommendations

- **Work environment related issues**
- Higher working hours point to extra effort due to work pressure is the biggest contributes to attrition
    - Recommend to review work planning for workload balancing so that only some employees do not get overstretched
    - In cases when it can't be avoided may be rewarding such employees might help reduce the impact
- Environment and Job satisfaction are big motivator for retaining an employee
    - Working on improving the general work culture might go a long way in retaining the employees
- **Employee human behavior**
- More experienced employee are stable and less likely churn
    - The firm should focus its attention on retaining employees with less experience as there’s higher churn in this category.
- Single employees form the majority of less experience employees, so these two factor combined need more attention.
- Promotions in recognition of performance tend to keep the employee happy and reduce attrition.
    - Available performance data shows only 2 ratings, 5 job levels, 9 job titles across 3 departments
    - Review and improvements of performance management along with familywise job architecture could help improve attrition.
- People being asked to travel frequently are leaving,
    - Even if the their role required still the travel frequency can be looked at in more details and only willing employees asked to travel
- **Hiring related issues**
- Prior job switches being greater in number represents the employee habit to churn.
    - Firm should look at an prospect’s past job tenures closely while hiring
- **Additionally**
- A high churn percentage on the HR department needs to be closely reviewed as it can keep check in overall percentage as well as stabilize HR practices.
- Factor like percentage hike are not in the final model , representing their neutrality towards churn whereas its is expected to have influence in attrition.
    - Firm should look into this space and offer competitive percentage hikes based on performance to retain talent.