HR ANALTICS CASE STUDY

By:

Savita Upadhyay

Vishnu Poddar

Vividh garg

Mohit Mamgain

Business Understanding

A large company named **XYZ** with 4000 employees have around 15% of employee attrition rate and this level of **attrition**(employees leaving, either on their own or because they got fired) is major concern 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 partners
- 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 acclimatise themselves to the company.

Business Objective

Role of HR Analytics

- To build **logistic regression model** to understand what factors must be focused on, in order to curb attrition. i.e. what changes should be made to workplace, in order to get most of employees to stay.
- Also, the variables which are most important and needs to be addressed right away.

Business Objective

Attributes associated with each employee:

- The Manager Survey Data – Collected from a company survey database.
- The Employee Survey Data – Collected from a company survey database.
- In-Time Data – Collected from company's attendance Records
- Out – Time Data – Collected from company's attendance Records
- General Data – General data includes employee personal data along with education and their satisfaction level for association with XYZ org, etc.

Data Cleaning and Preparation

- We have formatted the in time and out time data in date and time format.
- We also removed following dates from data based on our assumption that these dates are holidays

  X2015.01.01, X2015.01.14, X2015.01.26, X2015.03.05, X2015.05.01, X2015.07.17,

  X2015.09.17, X2015.10.02, X2015.11.09, X2015.11.10, X2015.11.11, X2015.12.25
- All the data files (i.e. Manager Survey Data, Employee Survey Data, In_Time Data , Out _Time Data, General Data have been merged to form a core data file for analysis.
- NA from numerical predictors have been filtered and reassigned by median and means adequately keeping outliers in consideration.

Data Cleaning and Preparation

- For Attrition, Gender, as these are having 2 levels these being realigned as numerical Yes ==1 and No == 0
- Dummy variables for following categorical predictors have been created for: EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance, JobRole, MaritalStatus, BusinessTravel, Department, Education, EducationField, JobInvolvement, JobLevel, PerformanceRating.
- Final dataset has been achieved after creation of dummy variables and derived metrics and outliers have been checked and deliberately kept as in our view it is not good to remove them, since in normal scenario they will be there and represent the company population where such data is bound to exist.
- Relevant predictors have been scaled to aid in regression modelling.

Derived Metrics

We have calculated average worked hours from (out_time- in_time) and then derived new metrics for better analysis.

- Over time : If the average worked hours for an employee is greater than the standard work hours i.e. 8 then the employee works overtime on an average.
- Less work time: If the average work hours for an employee is considerably less standard work hours i.e. less than 7 then the employee on an average has less work hours.
- No. of Leaves : If the employee has NA values for in time for a day which is not an assumed holiday or a weekend. Then employee is assumed to be on leave on that day. Therefore, the Count of all those NA values for each employee is taken as no of leaves.

Among all variable **Attrition is our categorical response variable** and rest are predictor variables.

Our Response variable is “Attrition”
(1 == Yes, & 0 == No)

Approach for Logistic Model Building

- For creating Train and test datasets from final data set:
  1. We fixed seed to 100
  2. Used split ratio of 0.7 for training dataset and remaining data has been assigned to test dataset
- Initial model has been conceived with GLM function, then StepAIC has been applied to arrived at standard model which yielded on iterative predictor selection with out major reduction in AIC Score.
- Then based on VIF (variance inflation factor) and P value (with significance) predictors have been filtered.

Final Model

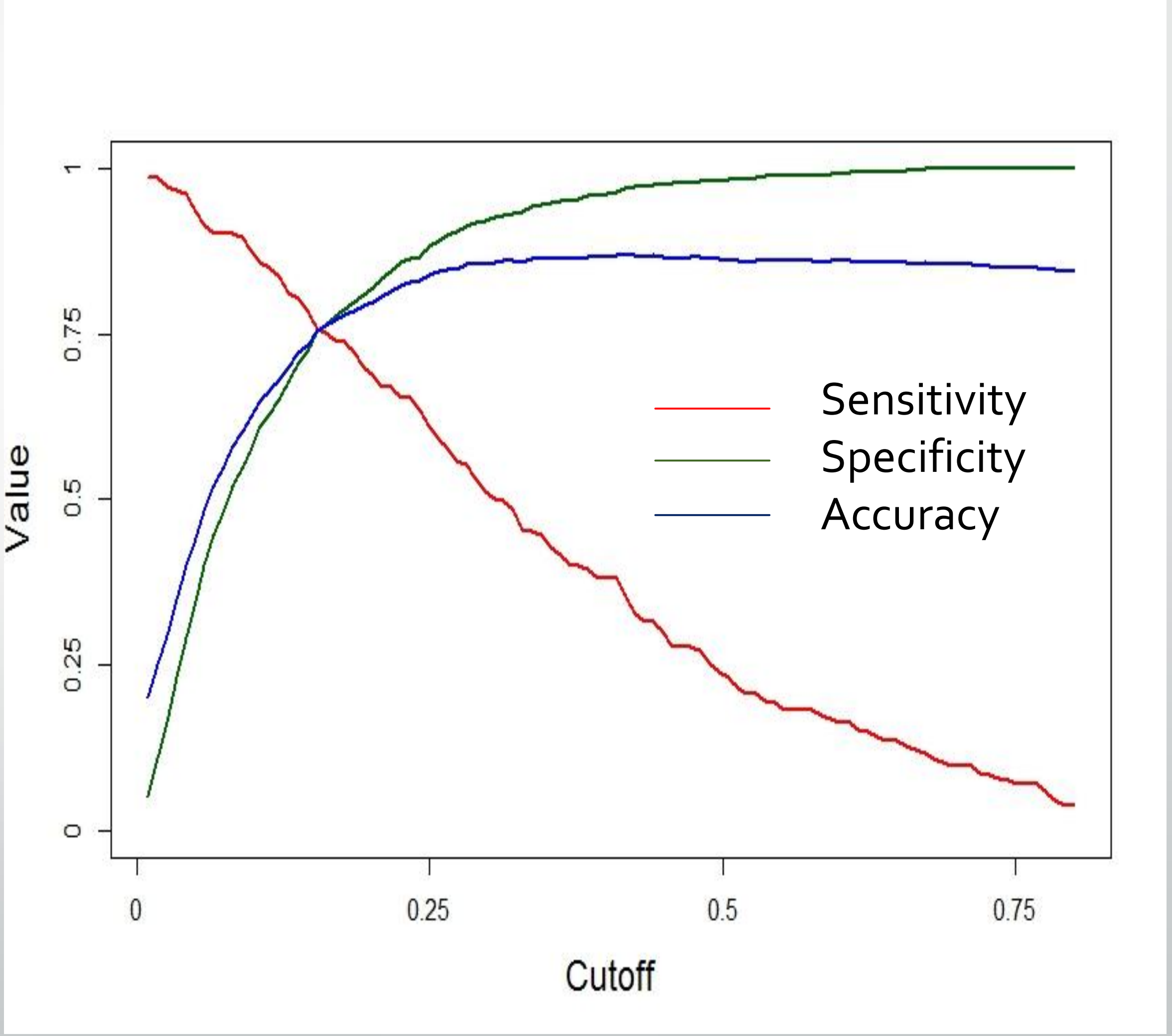
```
• model_26 <- glm(formula = Attrition ~ Age + NumCompaniesWorked+
                  TotalWorkingYears + TrainingTimesLastYear +
                  YearsSinceLastPromotion + YearsWithCurrManager+
                  Over_time + EnvironmentSatisfaction.xlow +
                  JobSatisfaction.xlow + JobSatisfaction.xvery.high +
                  WorkLifeBalance.xbetter +
                  BusinessTravel.xTravel_Frequently +
                  MaritalStatus.xSingle, family = "binomial", data = train)
```

Model Evaluation

CONFUSION MATRIX

PREDICTION		REFERENCE	
	NO	YES	TOTAL
NO	848	53	901
YES	262	160	422
TOTAL	1110	213	1323

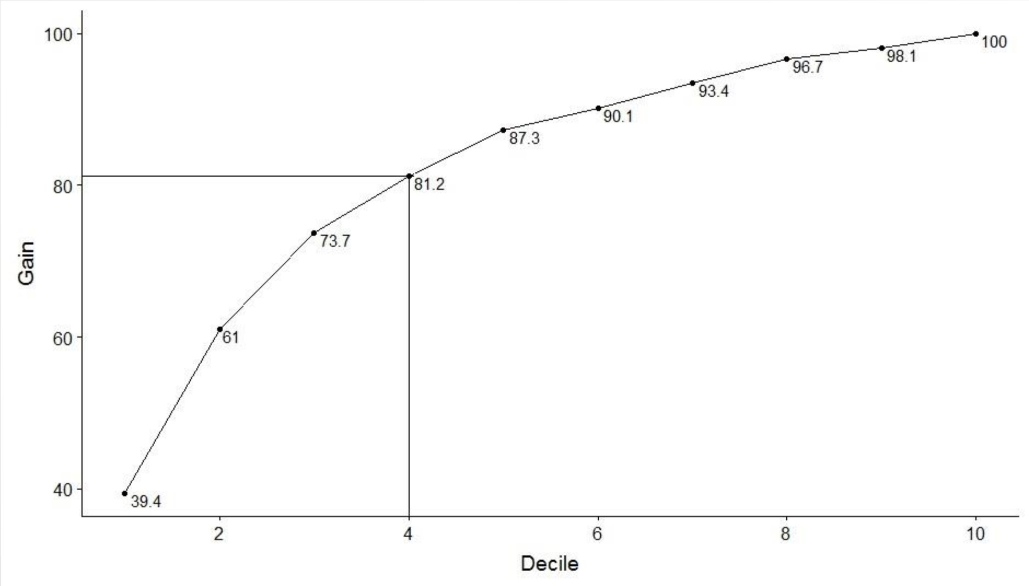
- Accuracy: 0.7619
- Sensitivity: 0.7512
- Specificity: 0.7640

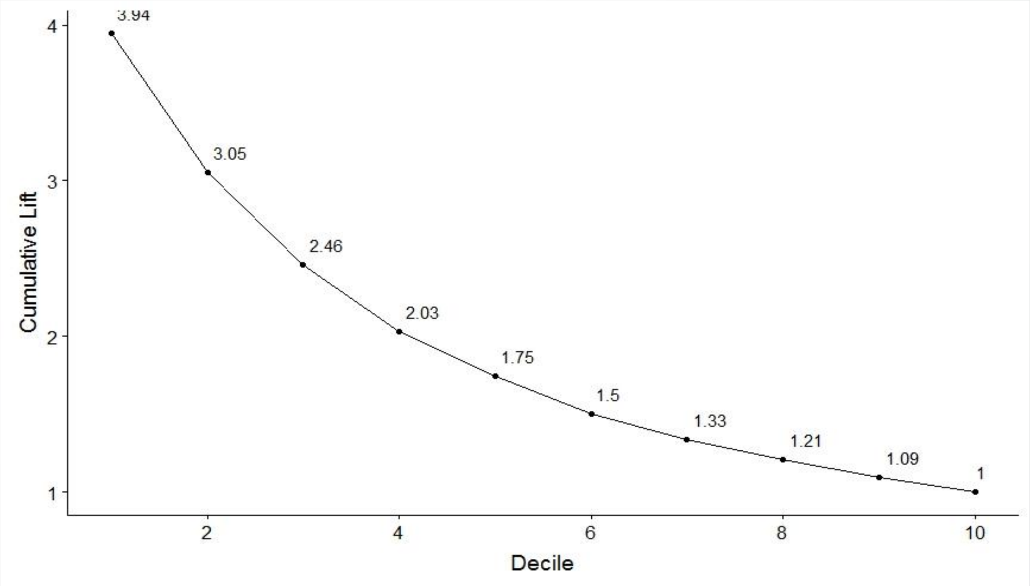


Model Assessment (GAIN & LIFT charts)

Decile	Observations	Churn	Cum- Churn	Gain(%Cum-Churn)	Gain (Random Model)	Lift
1	133	84	84	39.4	10	3.94
2	132	46	130	61.0	20	3.05
3	132	27	157	73.7	30	2.46
4	**133**	**16**	**173**	**81.2**	**40**	**2.03**
5	132	13	186	87.3	50	1.75
6	132	6	192	90.1	60	1.50
7	133	7	199	93.4	70	1.33
8	132	7	206	96.7	80	1.21
9	132	3	209	98.1	90	1.09
10	132	4	213	100	100	1.00
Total	1323	213				

GAIN AND LIFT CHARTS





- Gain are in PERCENTAGE.
- Gain for a given model, is **81% by the 4th decile**, what it basically means if all customers are sorted according to probability, then among the top 40% customers of this sorted list, you would find 81% of all employees likely to have attrition.

- The lift tells you the factor by which model is outperforming a random model.
- Our model's lift is equal to 2.0 by the 4th decile, it means that our model gain by the end of the 4th decile is 2.0 times that of a random model's gain at the end of 4th deciles.
- In other words, the model catches 2.0 times more attrition than a random model would have caught.

KS Statistics

Decile	Observations	Churn	Cum-Churn	% Cum-Churn	Non-Churn	Cum-Non-Churn	%Cum-Non-Churn	(%Cum-Churn) - (%Cum-Non-Churn)
1	133	84	84	39.4	49	49	4.4	35.0
2	132	46	130	61.0	86	135	12.2	48.8
3	**132**	**27**	**157**	**73.7**	**105**	**240**	**21.6**	**52.1**
4	133	16	173	81.2	117	357	32.2	49.0
5	132	13	186	87.3	119	476	42.9	44.4
6	132	6	192	90.1	126	602	54.2	35.9
7	133	7	199	93.4	126	728	65.6	27.8
8	132	7	206	96.7	125	853	76.8	19.9
9	132	3	209	98.1	129	982	88.6	9.5
10	132	4	213	100	128	1100	100	0
Total	1323	213			1110			

- The highest value of the term (%cum-churn - %cum-non-churn) is called the KS statistic.
- A high KS statistic means that our model have all attrition at the top, and all non-attrition at the bottom.
- For a good model, KS statistic would be more than 40% and would lie in the top few deciles (1st to 4th).
- **In our model KS Statistic is 52% and lie in 3rd decile**

Model Assessment

- The model has an increasing Gain and a decreasing Lift.
- The Model predicts more than 80% of the attritions within the 4th Decile with 76% accuracy.
- The KS statistic shows that the model is very good in distinguishing between employees who will leave the company and employees who won't.

Final Predictors Correlations

Column	
Age	-0.32328
NumCompaniesWorked	0.36851
TotalWorkingYears	-0.53756
TrainingTimesLastYear	-0.20081
YearsSinceLastPromotion	0.51299
YearsWithCurrManager	-0.50304
Over_time	1.35032
EnvironmentSatisfaction.xlow	1.06734
JobSatisfaction.xlow	0.58213
JobSatisfaction.xvery.high	-0.62814
WorkLifeBalance.xbetter	-0.37565
BusinessTravel.xTravel_Frequently	0.76335
MaritalStatus.xSingle	1.01651

Recommendations

- Environment Satisfaction, Job Satisfaction and Work life balance, the better these are for employees the less are their chances of leaving the company.
- The more an employee works overtime on an average the more are the chances that he/she will leave the company.
- If an employee works with the same manager for a longer period of time the lesser are the chances that employee will leave the company.
- Hire people with more experience as they are less likely to leave the company. But if the person has worked in many companies then the chances that he/she will leave the company increases.
- Employees who are unmarried are prone to leaving the company.

THANKS