HR ANALYTICS CASE STUDY

Group Name:

1. Sanjeev
2. Prakash
3. Neha
4. Bindu

OUR APPROACH

1. Problem Definition – “Higher Attrition Rate”
2. Understanding Objectives and expectations
3. Data collection, import in Platform and Data Understanding
4. Our Assumptions
5. Data Cleaning and creation of derived metrics, if any
6. Conducting Exploratory Data analysis along with relevant Data preparation
7. Collating Data in a Master File for Model Development
8. Identifying Predictor and Response Variables
9. Splitting Data in to Training and Test Data set randomly
10. Developing the model based on training set
11. Fine Tuning the model without overfitting
12. Selection of most appropriate model based on algorithm and Statistical criterion
13. Prediction from test set
14. Assessment of Quality of Prediction of Response Variable
15. Final Recommendations

WHY CONCERN?

- High attrition results in
  - Project delays
  - Missing Deadlines and loss of business
  - Loss of reputation among accounts
  - Increased new recruitment cost and high HR department overheads
  - Further additional Training and Development expenses
  - Negative motivation levels among current employees

WHY HR ANALYTICS INTERVENTION

- To understand
  - ✓ Factors to focus on for curbing attrition rate from growing
  - ✓ What changes are to be made in workplace environment to minimise attrition
  - ✓ Priorities among factors attributing high attrition to attend them immediately.

As a top rated analytics talent in the firm, this assignment has been bestowed on us.

ASSIGNMENT OBJECTIVES

- Developing a model of the ”Probability of Attrition”
- Method for designing model - Logistic Regression

DELIVERABLES

- The Attrition Probability Model based on logistic regression, which will help attaining results to
  - Understand major drivers (factors) for high attrition rate
  - What fine tuning is required in workplace settings, to keep the employees from quitting.

THE DATA GATHERED

1. The Manager Survey Data – Collected from a company wide survey.
2. The Employee Survey Data – Collected from a company wide survey.
3. In-Time Data – Collected from company's attendance Log sheet/ Machine
4. Out – Time Data – Collected from company's attendance Log sheet/ Machine

Rstudio

Data Import and basic understanding

1. All the five tables have been imported in Rstudio.

2. Data has been directly viewed in to grab basic understanding, and

3. Necessary assumptions for data cleaning and handling are made

DATA CLEANING AND DERIVED METRICS

- In this step we have formatted the in time and out time in date and time format for the entire period.
- Calculated the average work hours of each employee using in and out time
- Derived columns
  1. Overtime, 1 indicates yes while 0 = no
  2. Inadequate time means the employee is working mush less than the required hours on average
  3. No of leaves as derived metric

DATA PREPARATION AND RELATED EDA

- All the data files have been merged to form a core data file for analysis
- Distribution of categorical variables have been done to see outlier pattern
- NA from numerical predictors have been filtered and reassigned by median and means adequately
- For Attrition, Gender, Over18 -- 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
  - More than 2 levels -- EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance, JobRole, MaritalStatus, BusinessTravel, Department,Education, EducationField, JobInvolvement, JobLevel, PerformanceRating
- Final dataset has been achieved after this preparation exercises and outliers have been checked and deliberately kept as in our view it is not wise 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 and to avoid power effects, which are Salary, Age.

Final Dataset == hr_analytics_scaled

OUR RESPONSE VARIABLE IS “ATTRITION” (1 == YES, & 0 == NO)

Rest all non constant numeric variables are scaled to aid in regression modelling.

APPROACH FOR LOGISTIC REGRESSION

- 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 (2024.2 after 6 scoring iterations).
- Then based on VIF (variance inflation factor) and P value (with significance) predictors have been filtered and after another 16 iterations we could achieve our final model. with almost all predictors being significant with lowest VIF are present.

FINAL LOGISTIC REGRESSION MODEL

- Assessment_Model = glm(formula = Attrition ~ NumCompaniesWorked + TotalWorkingYears +

  ```
  YearsSinceLastPromotion + YearsWithCurrManager + Over_time +
  EnvironmentSatisfaction.x2 + EnvironmentSatisfaction.x3 +
  EnvironmentSatisfaction.x4 + JobSatisfaction.x2 + JobSatisfaction.x3 +
  JobSatisfaction.x4 + WorkLifeBalance.x2 + WorkLifeBalance.x3 +
  WorkLifeBalance.x4 + BusinessTravel.xTravel_Frequently +
  JobRole.xManufacturing.Director + MaritalStatus.xSingle,
  family = "binomial", data = train)
  ```

MODEL EVALUATION

- P_Cutoff>=0.5
- Confusion Matrix and Statistics

Prediction	Reference		
	No	Yes	Total
No	1068	170	873
Yes	26	59	450
Total	1094	229	1323

- Accuracy : 0.8519
- Specificity : 0.97623
- Pos Pred Value : 0.69412
- Neg Pred Value : 0.86268
- Prevalence : 0.17309
- Detection Rate : 0.04460
- Detection Prevalence : 0.06425
- Balanced Accuracy : 0.61694
- 'Positive' Class : Yes

Sensitivity : 0.25764

95% CI : (0.8316, 0.8706)

No Information Rate : 0.8269

P-Value [Acc > NIR] : 0.008179

Kappa : 0.3113

Mcnemar's Test P-Value : < 2.2e-16

MODEL EVALUATION

- P_Cutoff>=0.4
- Confusion Matrix and Statistics

Prediction	Reference		
	No	Yes	Total
No	1047	151	873
Yes	47	78	450
Total	1094	229	1323

- Accuracy : 0.8503
- 95% CI : (0.83, 0.8691)
- No Information Rate : 0.8269
- P-Value [Acc > NIR] : 0.01223
- Kappa : 0.3628
- Mcnemar's Test P-Value : 2.482e-13
- Detection Prevalence : 0.09448
- Balanced Accuracy : 0.64882
- 'Positive' Class : Yes

Sensitivity : 0.34061

Specificity : 0.95704

Pos Pred Value : 0.62400

Neg Pred Value : 0.87396

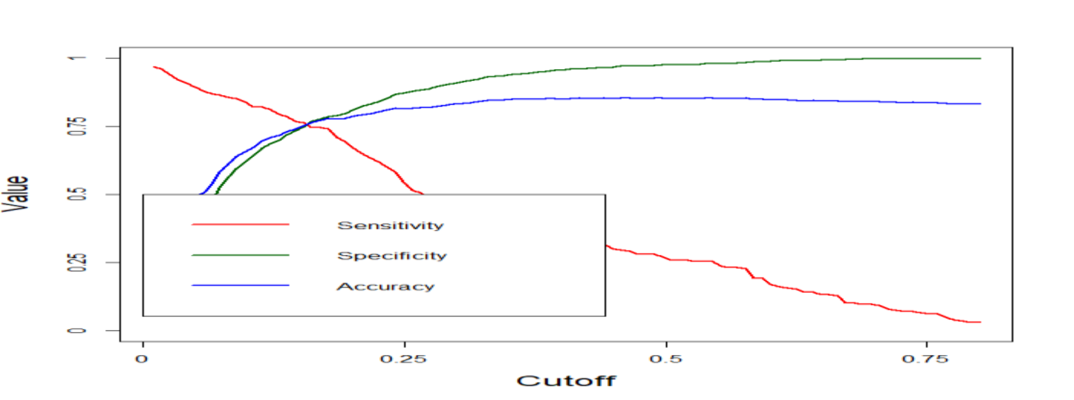
Prevalence : 0.17309

Detection Rate : 0.05896

PREDICTION BASED ON FINAL MODEL

Confusion Matrix and Statistics

	Reference		
Prediction	No	Yes	Total
No	819	54	873
Yes	275	175	450
Total	1094	229	1323



Accuracy : 0.7513

95% CI : (0.7271, 0.7744)

No Information Rate : 0.8269

P-Value [Acc > NIR] : 1

Kappa : 0.3712

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

Balanced Accuracy : 0.7564

'Positive' Class : Yes

Sensitivity : 0.7642

Specificity : 0.7486

Pos Pred Value : 0.3889

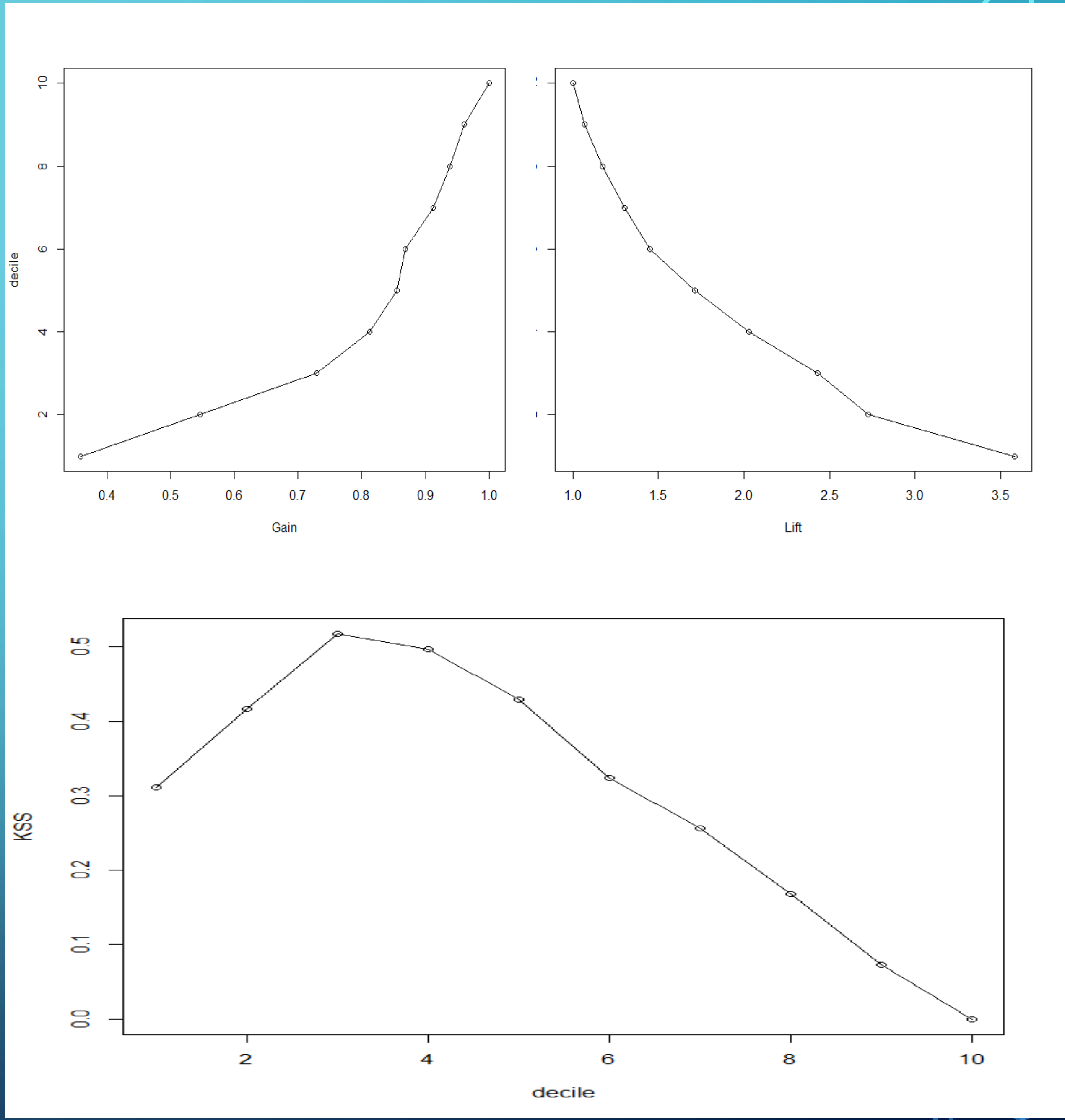
Neg Pred Value : 0.9381

Prevalence : 0.1731

Detection Prevalence : 0.3401

MODEL ASSESSMENT (GAIN & LIFT AND KSS)

Decile	Observations	Churn	Cum- Churn	Gain(%Cum-Churn)	Gain (Random Model)	Lift	KSS
1	133	82	82	35.80%	10%	3.58	0.3114607
2	133	43	125	54.60%	20%	2.73	0.4169667
3	133	42	167	72.90%	30%	2.43	0.5171918
4	132	19	186	81.20%	40%	2.03	0.4968706
5	132	10	196	85.60%	50%	1.71	0.4290213
6	132	3	199	86.90%	60%	1.45	0.3242059
7	132	10	209	91.30%	70%	1.3	0.2563566
8	132	6	215	93.90%	80%	1.17	0.1673838
9	132	5	220	96.10%	90%	1.07	0.0731301
10	132	9	229	100.00%	100%	1	0
Total	1323	229					



MODEL ASSESSMENT SUMMARY

- The model has an increasing Gain and a decreasing Lift.
- The Model predicts more than 80% of the attritions within the 4th Decile with 75% 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.

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.

THANK YOU