HR ANALYTICS CASE STUDY

SUBMISSION

Group Name:

1. Soura Dutta
2. Jasaswini Mohanty
3. Gaurav Makkar
4. Abhik Mitra

Business Understanding

- Company XYZ suffers from employees attrition with the rate of 15% employees leaving the company despite the company hiring around 4000 employees at any given point of time.
- This kind of attrition is bad for the company due to the following reasons :
    1. The former employees' projects get delayed, which makes it difficult to meet **timelines**, resulting in a reputation loss among consumers and partners.
    2. A sizeable department has to be maintained, for the purposes of **recruiting** new talent.
    3. The new employees have to be **trained** for the job
- Hence, the main target is to analyse the data and predict the employees who are likely to leave the company so that the management can come up with strategies to retain their employees.

Data Understanding

- The data is divided into different files namely, employee survey data, manager survey data, general data, in-time of the employees and the out-time of the employees.
- The employee survey data consists of data about how each employee feels about his/her job and the working environment.
- Manager survey data consists of each employees' performance rating.
- The general data consists of the basic information about each employee. It also has information about the target variable attrition.
- Both in-time and out-time are used for the purpose of attendance of each employee. The difference of both the datasets has been used for computing the average working hours of each employee on a monthly basis.
- The data has quality issues such as missing values which have been treated with medians of the respective fields.
- Moreover, outliers have also been detected which have been treated using the appropriate statistical measures.

EDA

- Employee survey data, general data and manager survey data were merged and the merged variable has 4410 records and 29 variables.
- The merged variable has 111 missing values which were addressed column-wise. The variables environment satisfaction, job satisfaction, work life balance were found to have missing values.
- The missing values have been replaced with the highest frequency values.
- The data had 5 numeric variables having outliers which have been treated.
- The date-time variables have been standardised using appropriate methodologies.
- The missing date-time values have been replaced with 0 to ease the calculation of average working hours.

Data Preparation

- ❖ Initially, all the numeric variables have been scaled to prevent a biased model and prediction.
- ❖ The creation of dummy variables has been carried out in two steps:
  1. Creation of binary level dummy variables.
  2. Creation of multiple level dummy variables.
- ❖ Creation of difference and average working hours data frame to compute store the result of average hours spent by each employee on the job per month.
- ❖ The average working hours data is then merged with the master dataset and further EDA is carried out on the merged data frame.

Model Building

- ❖ The master data has been split into train and test data with 70:30 ratio.
- ❖ For all the models created, the number of significant variables, p values and the VIF values have been considered and the insignificant ones removed.
- ❖ The most optimal model has 11 significant variables with VIF values for each variable less than 2 (as per business standards).

List of Significant Variables

- These following variables have been found to be significant in case of predicting attrition –
  - Age
  - Number of Companies Worked
  - Training Times Last Year
  - Years Since Last Promotion
  - Years with Current Manager
  - Marital Status – Single
  - Environmental Satisfaction – Levels 2, 3 and 4
  - Job Satisfaction – Level 4
  - Average Working Hours – October

Model Evaluation

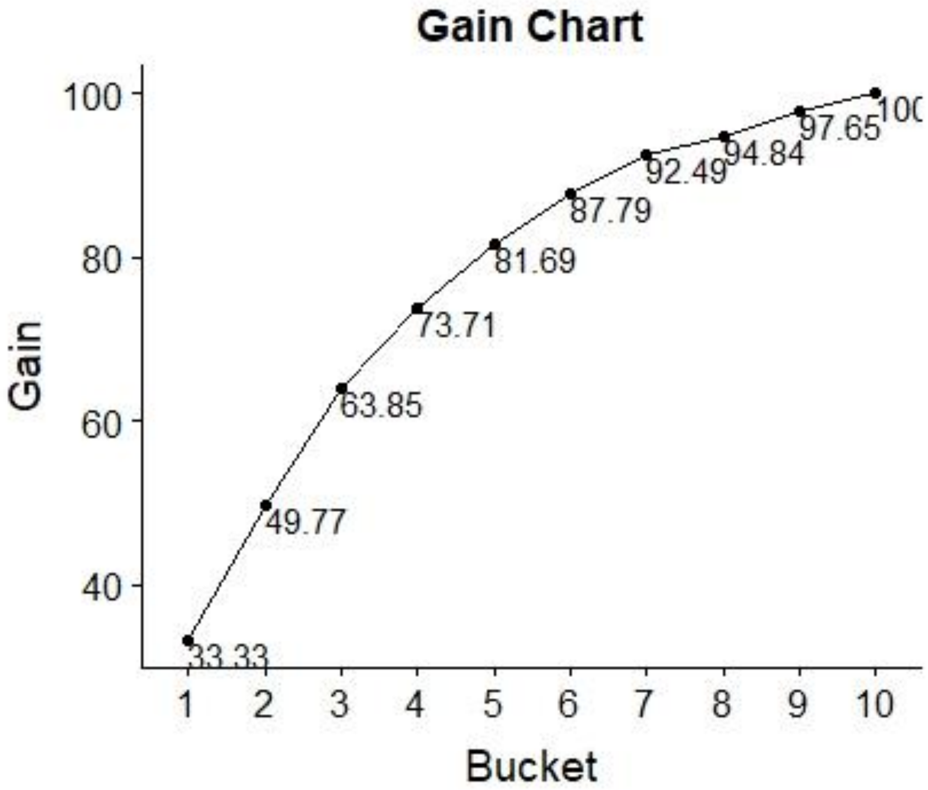
- The threshold value should have been chosen to create a balance between sensitivity and specificity to get the optimal results and optimum accuracy of the model. A higher value of sensitivity gives higher true positive value and lower false negative values. Therefore, according to business demands we needed a fair amount of actual attrition to be predicted. Hence, the **threshold for predicted attrition has been taken as 0.158**.
- The confusion matrix -

	Reference	
Prediction	**No**	**Yes**
No	781	61
Yes	329	152

- For the balanced model, the **Accuracy has been found to be 70.5%** ,
  Sensitivity has been found to be 71.4%,
  Specificity has been found to be 70.4%.
- KS statistic is a measure of the quality of a model. A KS stat value greater than 40 is indicative of a good stable model which is achieved by the model built. Our final model has **KS value 41.7**.

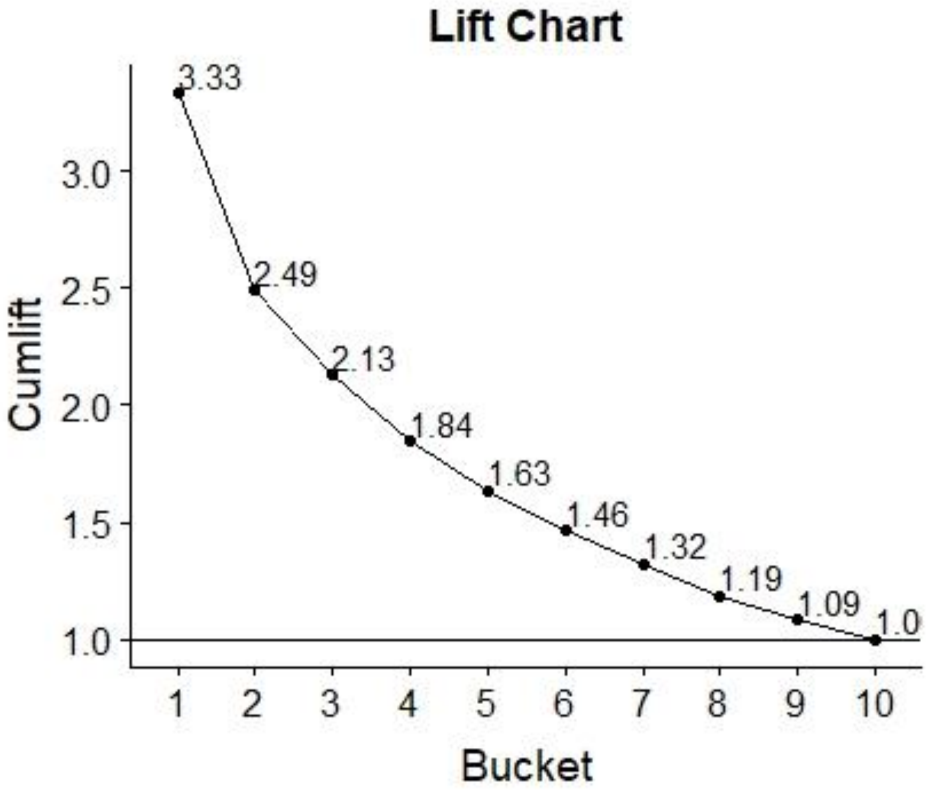
Gain Chart

- Gain chart gives a measure of the effectiveness of the model.
- With each decile, we can derive the percent of employees that can be accurately predicted for a positive attrition.
- From the gain chart, we can deduce that for the 4th decile, the percentage of accuracy of predicting the number of employees likely to undergo attrition is 73.71, till 5th decile we get 81.69% attritions covered and till 6th we can cover nearly 88% of attritions correctly.



Lift Chart

- ❖ Lift chart, tells us the ratio of gain chart for the R predicted model to the random model.
- ❖ It tells us how well R predicted model would predict as compared to a random generated model for a given decile.
- ❖ For the 1st decile, the gain of the R generated model predicts 3.33 times better, for the 2nd 2.49 times, for 3rd 2.13 times and for 4th decile,1.84 times better.



Conclusions

- Our model is 71.4% successful in predicting the actual churners correctly, which was the target of building this model. The predicted churners can be targeted by administration with different marketing strategies to help them retain.
- Both the gain and lift charts signify that the predicted amount of employees leaving the company is fairly correct and which is again verified by a good KS statistic value.
- The overall achieved model accuracy of 70.5% , high sensitivity and specificity say that the significant variables selected, are statistically and intuitively correct in case of predicting attritions.