HR ANALYTICS CASE STUDY

SUBMISSION

Group Name:

1. Nirmal Maheshwari
2. Vijay Mane
3. Shwetank Pandey
4. Vijaya Raman

Business Objective

XYZ company is a large company with 4000 employees at any point of time, 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

Hence, the management has contracted an **HR analytics** firm to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away.

Goal of this case study is to find the factors affecting the attrition rate of the employee using Logistic Regression and present them in business form so that the company can take appropriate steps to retain the employees

Problem Solving Methodology- CRISP DM Framework

Business Objective : Explained Above Slide

Data : Present in Excel Format, Information about all companies and details about Investment in them and mapping of primary sectors to their main sectors

Data Preparation : Cleaning of the data available by replacing the NA values with meaningful values wherever necessary.

Modelling: Applying a Model following the constraints above and sorting the results step by step coming to a conclusion for completing the objective.

Evaluation: Evaluating the results in different tools, reviewing the process and summarizing the results keeping the business success constraints in mind(Above Slide)

Tools Used : R, Excel, Tableau

Data Understanding And Preparation

Strategy And Steps Followed

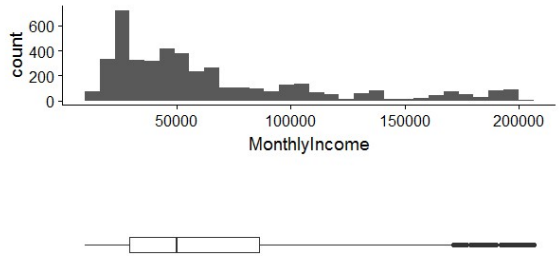
1. Reading Excel Data available in R.
2. Merging of the Data in a single master frame for further analysis.
3. Converting the time data into R compatible date format and finding the time difference for each day for every employee in a separate column
4. Removing the rows having NA values for column Total working years and Number of companies worked as the NA values have very less percentage
5. Removing those columns which are having only one kind of values in all rows.
6. Imputation of the remaining NA values with the mode of the corresponding columns, Columns are EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance
7. Assumption taken for Number of Companies worked column that Current Company is not counted in this column and the data is cleaned accordingly for this column

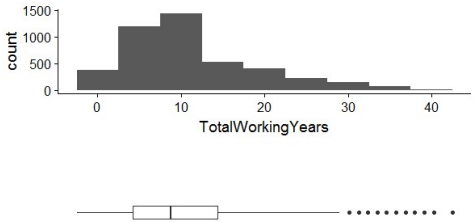
Outlier Removal and Converting Discrete Variables into Categorical

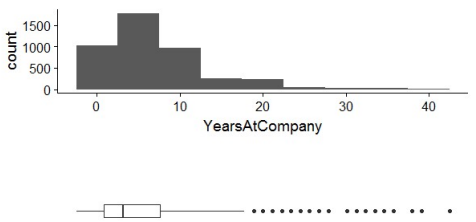
Strategy And Steps Followed

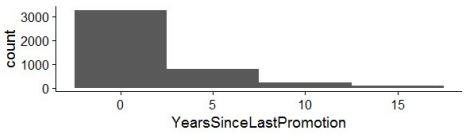
1. Making the box plot for every continuous variable and checking for the outliers.
2. Imputing the outliers using the Interquartile function.
3. Outliers found in the columns Monthly Income, Total Working Years, Years At Company, Years since Last Promotion ,Years with Current Manager.
4. Converting discrete variables into categorical variables using as.factor command in R.

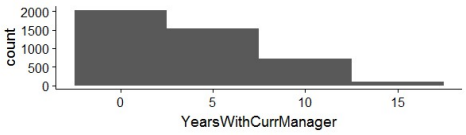
Box Plots Showing Outliers











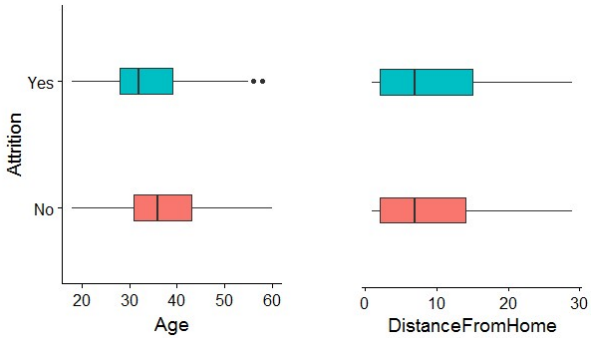
EDA for Continuous Variables

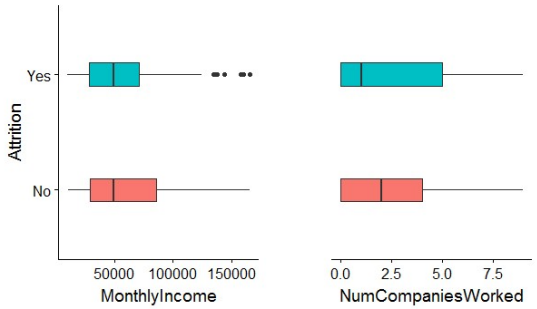
Deriving two new columns Average time spent on each day and number of leaves taken by each employee in an year.

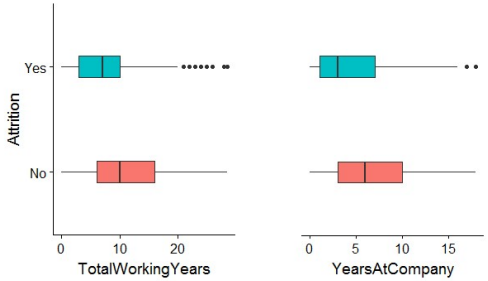
Observation by Boxplots of numeric variables relative to attrition status

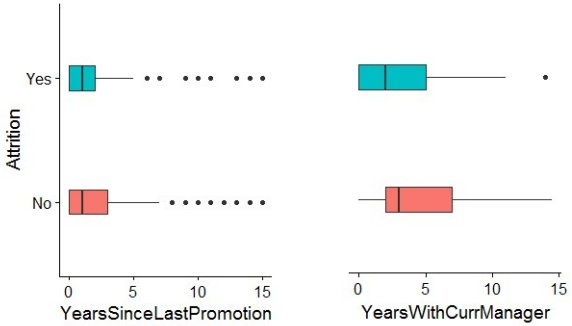
1. People who resigns from a company comes from the lower Age bracket.
2. The 50-75 percentile of the person who resigns have to travel more distance from home.
3. Monthly income for the persons who resigns is in less bracket than who stays with the company and the number of company worked bracket is more for the person who resigned.
4. Less experienced people resign more and person who have spent less years with the company resigned more.
5. Promotion bracket for the person who leaves the company is less than person who stays and as expected the years with current manager is also for the person who leaves the company.
6. Interestingly that the average time spent by the persons who leaves the company is more than the person who stays with the company

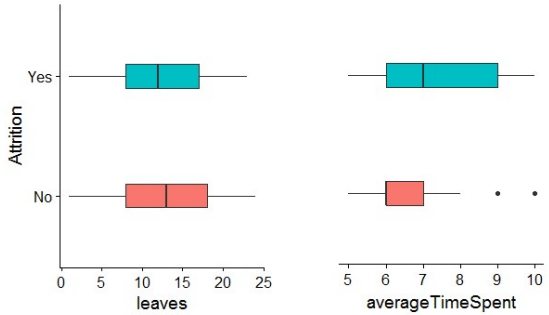
Graphs Proving Above Observation





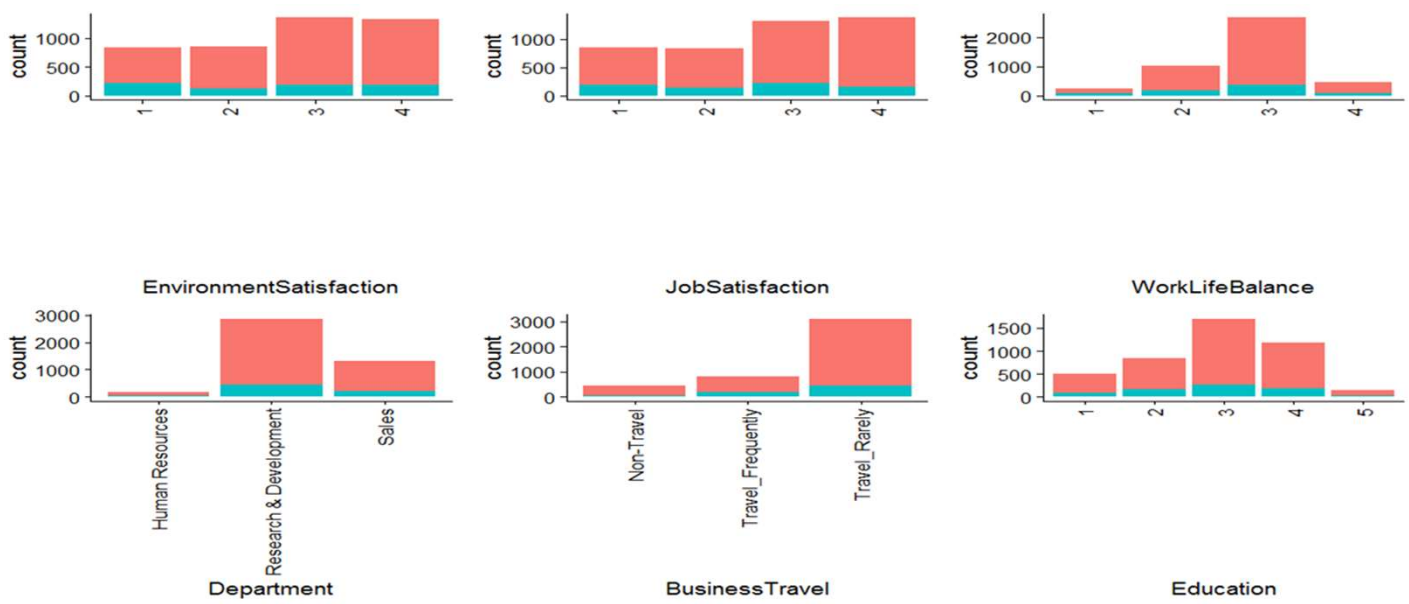






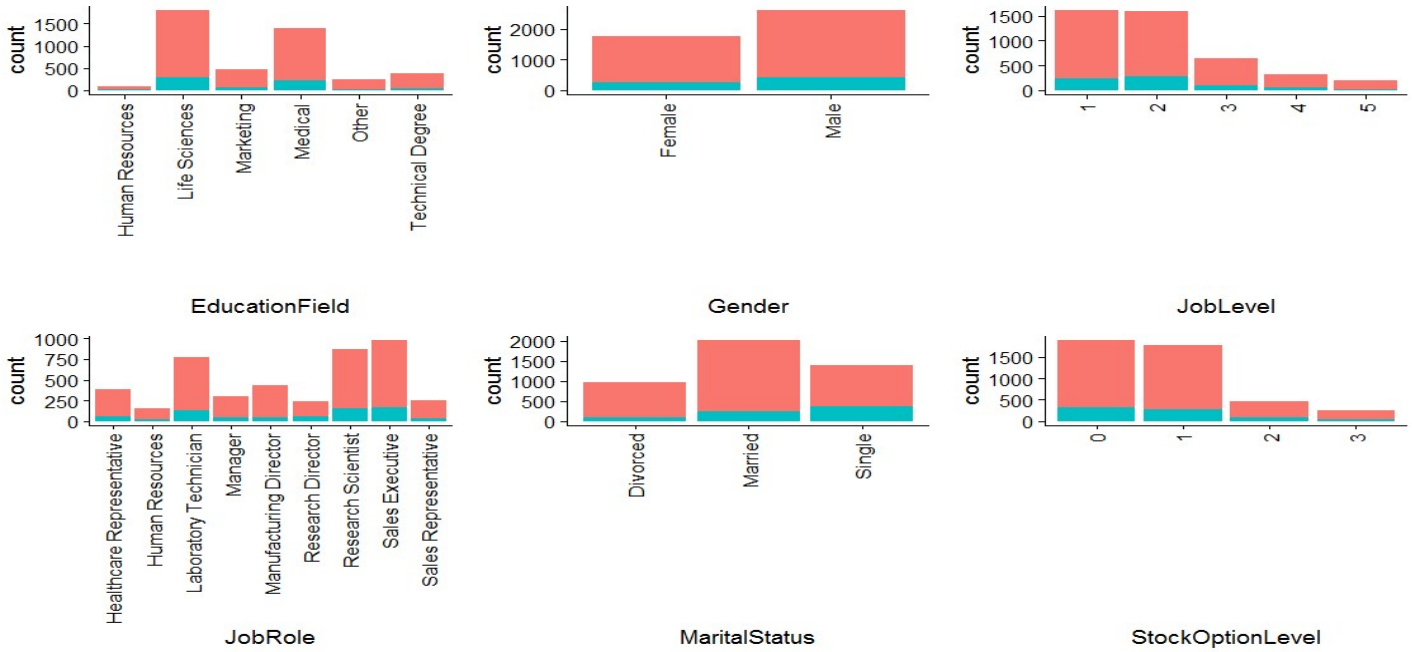
EDA for Categorical Variables

- Below plot will be revealing moderate contrast of attrition rate with Environment satisfaction and the Business Travel and Work Life Balance.



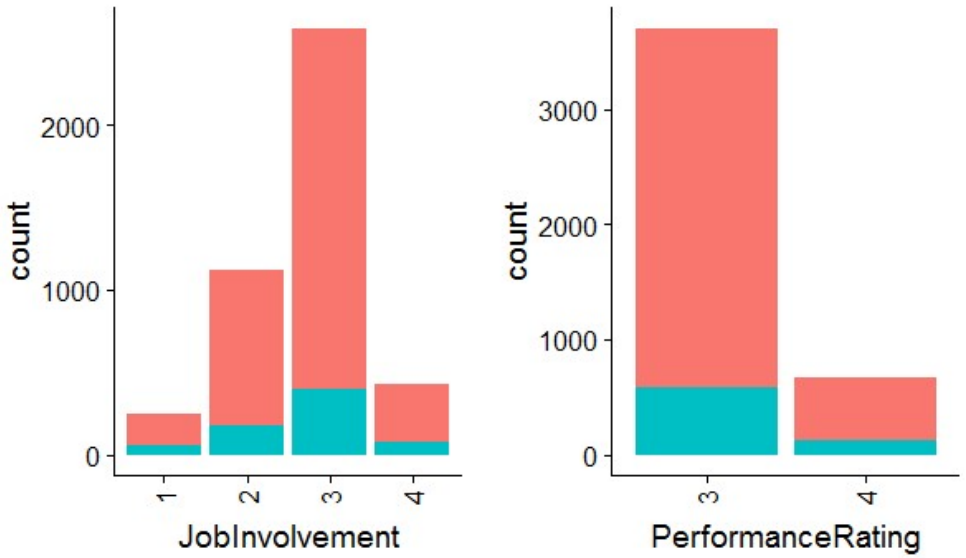
EDA for Categorical Variables

Below plot reveals strong contrast of attrition rate with Education Field Job Level Job Role and Marital Status



EDA for Categorical Variables

- Below plot reveals strong contrast of attrition rate with Job Involvement



Modelling

Steps Followed

1. Scaling of all the continuous variables.
2. Converting categorical variables into dummies so that they can be converted into numeric variables.
3. Total Attrition Rate comes out to be 16.08 percent of total.
4. Splitting data into training and test data set.
5. Prepare Model using Logistic Regression and removing the variables having high VIF and low significance.

Significant Variables and there relation

Overall there are 15 significant variables with positive and negative coefficients, positive coefficient means they will be positively more the value of the variable more is the chance of attrition and negative coefficient means they will be affecting attrition rate negatively. In case of categorical variables like EnvironmentSatisfaction the coefficient is negative as as the satisfaction level is high (with highest rating =4), the attrition is lower, hence they are inversely proportional to each other and hence the negative co-efficient.

Significant Variables List Along with Coefficient Values.

1. NumCompaniesWorked = 0.2546
2. TotalWorkingYears = -0.7418
3. YearsSinceLastPromotion = 0.4778
4. YearsWithCurrManager = -0.4582
5. averageTimeSpent = 0.6382
6. EnvironmentSatisfaction.x2(Level 2) = -0.9893
7. EnvironmentSatisfaction.x3(Level 3) = -0.8954
8. EnvironmentSatisfaction.x4(Level 4) = -1.2967
9. JobSatisfaction.x4(Level 4)= -0.8055
10. WorkLifeBalance.x3(Level 3) = -0.3750
11. BusinessTravel.xTravel_Frequently = 0.8974
12. Department.xResearch...Development = -0.9304
13. Department.xSales = -0.9817
14. JobRole.xManufacturing.Director = -0.7477
15. MaritalStatus.xSingle = 1.0304

Model Evaluation

- Generating the confusion matrix with different cutoff levels and then comparing the sensitivity specificity and accuracy of the model.
- Finding the optimal cutoff value and creating the graph showing the optimal cutoff

Cutoff Probability	Sensitivity	Specificity	Accuracy
50 %	0.22170	0.97008	0.8517
40%	0.32547	0.97280	0.8494
25%	0.51415	0.84769	0.7939
Optimal Cutoff(17%)	0.73113	0.73708	0.7361

Conclusion

- Model and EDA lay out the following factors for the company.

1. Company has to take the factor seriously about the number of companies candidate has worked before and then should give him the offer to join.
2. Company has to improve the Working environment and then should manage the work properly as some employees might be working more because of work pressure then the other and they tend to leave because of these working condition.
3. Company has to look into the number of years the employee has to wait for promotion so that he would be satisfied with his job and has to take care of the Sales and R and D department separately as the model shows significant attrition in these 2 departments.
4. Then there are some factors such as Job role, Work Life balance which are interlinked with the working conditions and also the employees who are single tend to leave the company more.

Accuracy Specificity and Sensitivity