HR ANALYTICS

Using R Programming

KHOO Kian Sim, TONG Zi Heng, WU Tong, YEO Yi Xuan, ZHANG Jieyuan

Content

1. Introduction

Employee attrition and performance are some of the core focuses for every company's Human Resources (HR) department. This is because having high turnover leads to poor performance for a company, thus reducing turnover will allow the company to have advantage over its competitors (Spain & Groysberg, 2016). However, according to Deloitte's Global Human Capital Trends 2017 (Deloitte University Press, 2017), "only 9 percent believe they have a good understanding of the talent factors that drive performance".

We aim to enable HR professionals of Company X to dive deeper into the different factors that can affect employee attrition and performance and provide insights on how the HR department can structure Company X's HR policy to achieve lower employee attrition and better employee performance.

The following sections outline the concept, data source and methodology that we will take to provide the foundation on which HR professionals of Company X can leverage and take steps to drive long-term success of the company by implementing informed HR policies.

2. Overall Concept

The traditional HR approach towards employee retention and performance management is through using annual appraisals and getting employees to provide feedback on various metrics such as their individual goals and development. However, that itself is simply not sufficient anymore (Netabai, 2020).

This study aims to utilise data analytics and design R Shiny App to achieve the following:

1. To explore the employee profile across different factors such as age, gender and income
2. To analyse the factors and measure their potential difference of means and association with employee attrition and performance
3. To utilise logistic regression to predict probability of attrition of individual employee based on given factors

With the analysis done in the study, we aim to provide insights for HR to determine the factors that the company's policies should focus on in order to achieve greater success. We will utilize R and its packages to perform the various analysis and studies. For more information on the R packages used, refer to Appendix A.

3. Data Source

The dataset was downloaded from Kaggle.com (https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset). It has data of 1470 employees from Company X, which contains 35 variables, such as gender, age, years at company, department, education background, income, performance rating and attrition (Refer to Appendix B for detailed description of individual variables).

3.1. Data Preparation

Firstly, the categorical data for performance rating is changed from "3" and "4" to "Average" and "Great" respectively for better clarity in its meaning.

Secondly, a new column “AttritionBin” is added to map “Yes” value from the “Attrition” column to “1” integer value and “No” value from the “Attrition” column to “0” integer value as binary integer is required in fitting the logistic regression model.

Thirdly, the class of “Education”, “EnvironmentSatisfaction”, “JobInvolvement”, “JobLevel”, “JobSatisfaction”, “RelationshipSatisfaction”, “StockOptionLevel” and “WorkLifeBalance” are changed to factor data type to reflect the categorical type of data and set to be ordered=TRUE to reflect that the variables exist in naturally occurring ordered categories.

Lastly, a new column “NewOldEmployee” is added to classify employees with less than a year tenure in the company as “New” and employees with one or more years tenure in the company as “Old”.

3.2. Descriptive Statistics

From Figure 1, 60% of the employees are male. In addition, majority of the employees are aged between 33 and 37 and most employees have been with the company for 3 to 7 years.

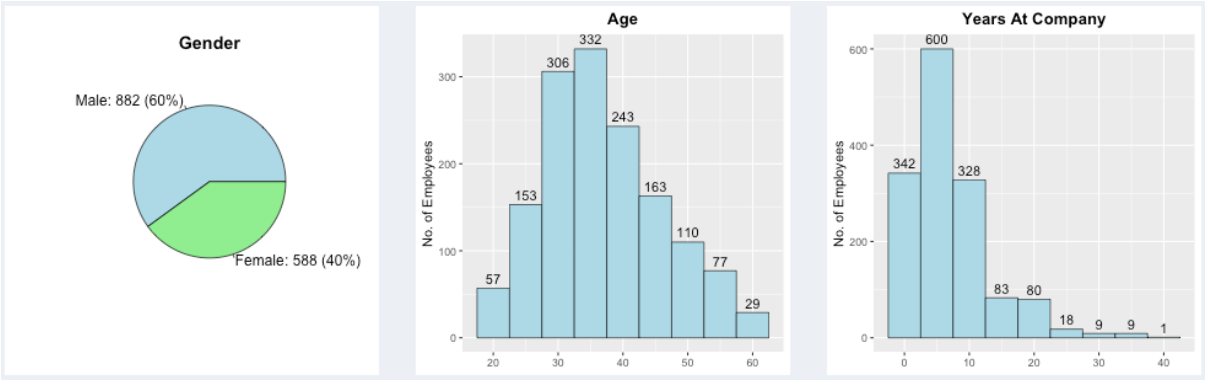


Figure 1: Distribution of employees by gender, age and tenure

From the distribution of employees across the departments, 65.3% of the employees are in the Research & Development department and 72.8% of the employees have education background in either life sciences or medical fields, as calculated from Figure 2.

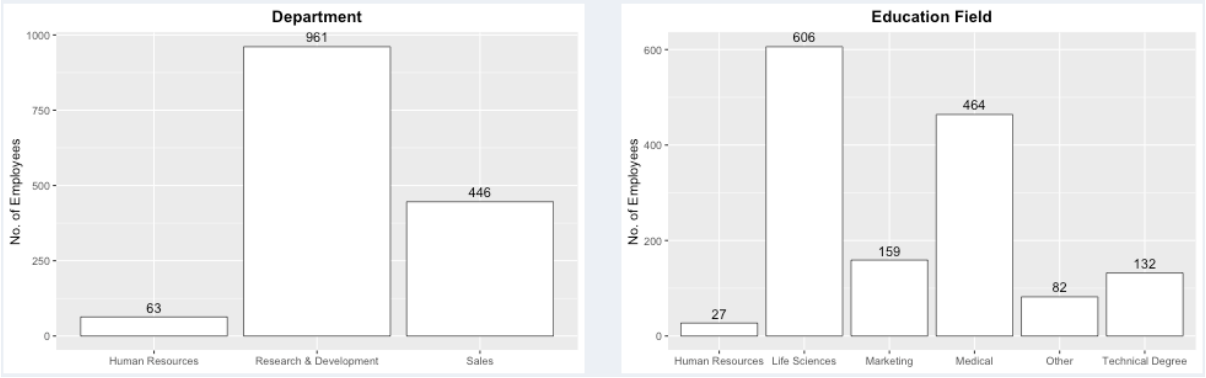


Figure 2: No. of employees by department and education field

With regards to monthly income, Figure 3 shows that the Sales department has the highest median monthly income, which might indicate that the company prioritises paying a higher salary to its sales staff. In line with the norm, the median income increases according to the employees’ job level. However, it is unexpected that the median monthly income of employees with a “College” education level is slightly higher than that of employees with “Bachelor” education level, which possibly indicate that the company value both equally. This is supported by the fact that

the average working experience of employees with “College” and “Bachelor” education level is around the same at 10.71 years and 11.26 years respectively.

Figure 3: Monthly income in USD by department, job level and education level

The monthly income of new and old employees is shown in Figure 4. A new employee is defined as an employee with less than a year tenure in the company while an old employee is defined as an employee with one or more years tenure in the company. From Figure 4, new employees with Job Level 2 and 3 in the Research & Development department and new employees with Job Level 3 in the Sales department have higher median monthly income as compared to the old employees. If this is not managed well by the Human Resource department, it could lead to dissatisfaction amongst old employees. In addition, there is only one new employee in Research & Development department as seen under the columns of Job Level 4 and 5, which likely indicate the company’s preference to promote employees from within. This is supported by the median tenure of employees with Job Level 4 and 5 in the company being 12.5 years and 18 years respectively.

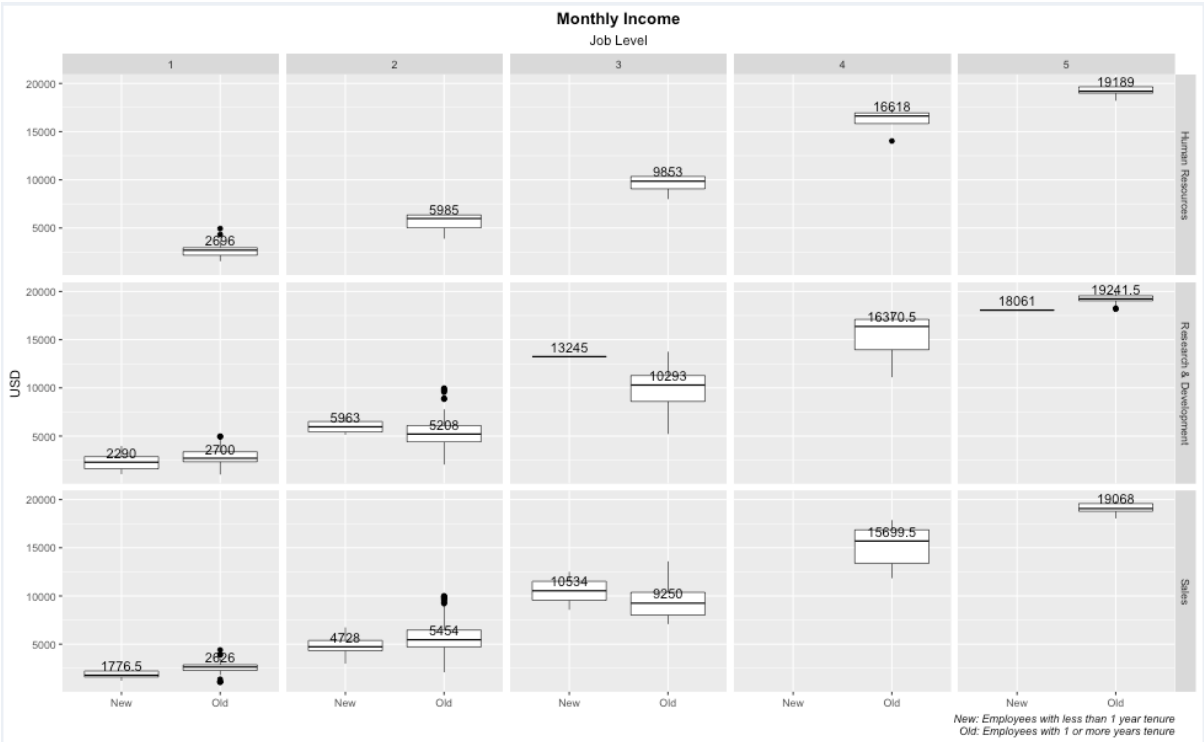


Figure 4: Monthly income in USD by employee tenure, department and job level

The monthly income by gender is shown in Figure 5. When broken down into department and job level, the median income of female and male are largely comparable, which might indicate that the company places importance in ensuring gender pay equality. However, there is exception for

employees with Job Level 4 in the Human Resource department and employees with Job level 4 in the Sales department.

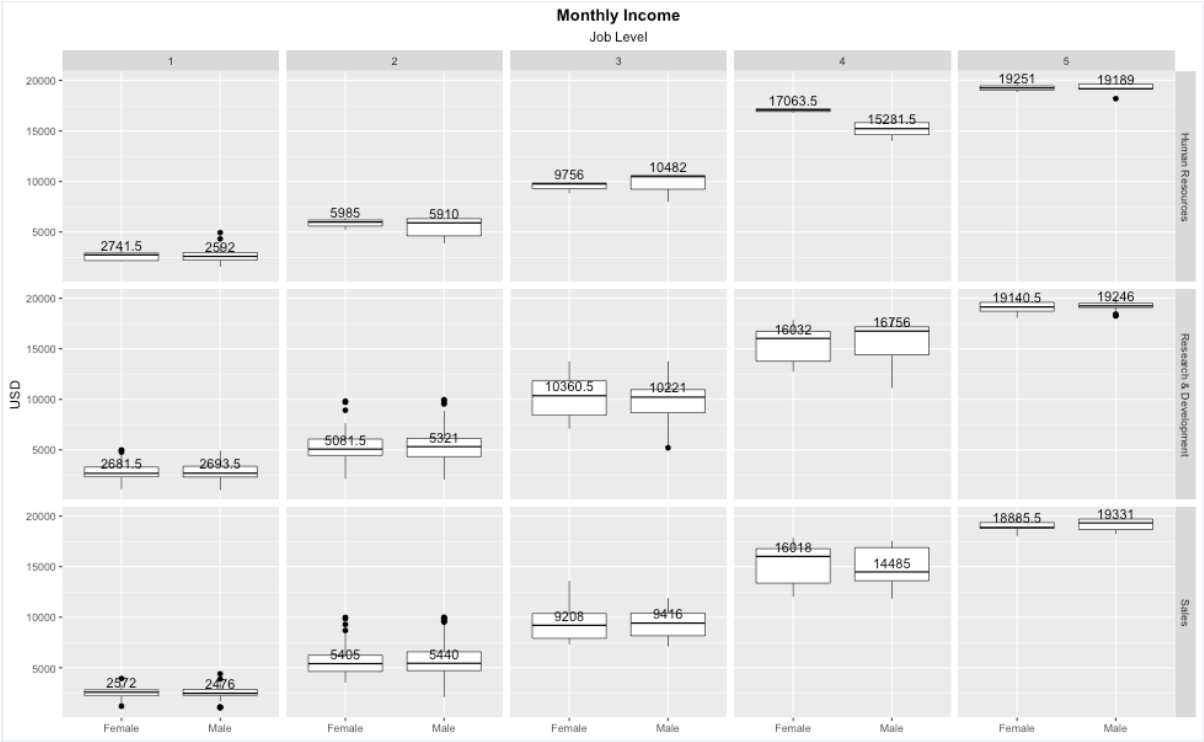


Figure 5: Monthly income in USD by gender, department and job level

From Figure 6, 15.37% of the employees have a "Great" performance rating and 16.12% of the employees have attrited.

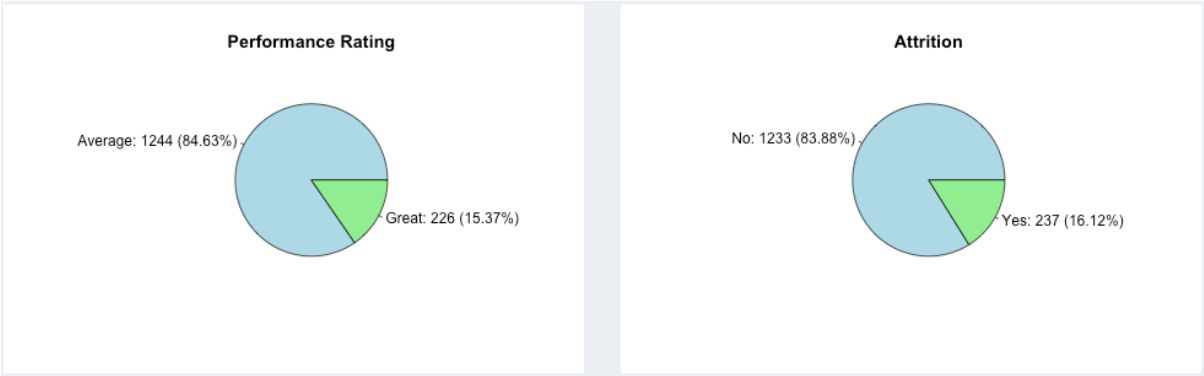


Figure 6: Breakdown of employees by performance rating and attrition

4. Methodology

4.1. Comparison of Mean

4.1.1. Methodology Overview

We are interested in finding if there are significant differences in mean across different factors between employees who were rated as average (n = 1,244) and great (n=226) as well as between employees who attrited (n=237) and those who remain with the company (n=1,233).

Since we are comparing the differences in mean, this will be a comparison of continuous variables such as age, daily rate, distance from home, hourly rate, monthly income, monthly rate, number of companies work at, total working years, number of trainings received last

year, number of years at company, numbers of years at current role, number of years since last promotion, number of years with current manager and salary hike percentage.

Z-test will be used to measure as n>30 for the respective groups in both the performance ratings and attrition comparison. Below is a code snippet for the z-test.

```
Average_performance <-HRdata[HRdata$PerformanceRating == "Average",]$YearsWithCurrManager
Great_performance <-HRdata[HRdata$PerformanceRating == "Great",]$YearsWithCurrManager
result1 <- z.test(Average_performance,Great_performance, sigma.x = sd(Average_performance), sigma.y = sd(Great_performance))
```

Figure 7: Code snippet for comparison of means using z-test

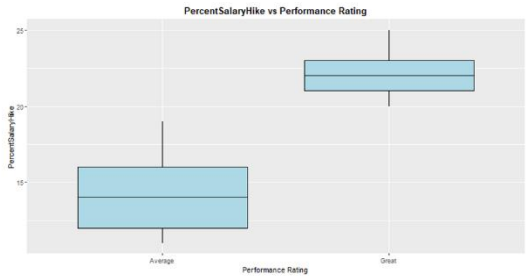
4.1.2. Performance Rating

Based on the z-test, we only have enough statistical evidence to reject the null hypothesis at 95% significant level and support the statement that there is a difference in mean percentage salary hike between employees who were rated to have greater performance and those who were rated as average performer.

Variable	Great performance mean	Average performance mean	Diff	p-value
Age	37.0	36.9	0.0	0.94
Daily Rate	802.9	802.4	0.5	0.99
Distance From Home	9.7	9.1	0.6	0.33
Hourly Rate	65.8	65.9	-0.1	0.93
Monthly Income	6313.9	6537.3	-223.4	0.51
Monthly Rate	14149.3	14342.9	-193.6	0.72
Number of company work at	2.6	2.7	-0.1	0.58
Total Working Years	11.4	11.3	0.1	0.80
Number of trainings received last year	2.8	2.8	-0.1	0.55
Number of years at company	7.1	7.0	0.1	0.89
Numbers of years at current role	4.5	4.2	0.4	0.20
Number of years since last promotion	2.3	2.2	0.2	0.52
Number of years with current manager	4.3	4.1	0.2	0.38
Salary hike percentage	**21.8**	**14.0**	**7.8**	**< 2.2e-16**

Figure 8: Summary of means for employee with average performance and great performance and p-value for comparison of means

As seen from the boxplot and z-test summary below, employees who were rated to have great performance (mean = 21.8%) have a higher mean percentage salary hike compared to employees that were rated to have average performance (mean = 14.0%). This is likely due to performance ratings being a factor in determining the salary hike percentage.




```
        Two-sample z-Test

data:  Average_performance and Great_performance
z = -63.389, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.088949 -7.603735
sample estimates:
mean of x mean of y
 14.00322  21.84956
```

Figure 9: Mean of salary hike percentage for employee with average performance and great performance and z-test results summary

4.1.3. Attrition

At 95% significant level, there are enough evidence to reject the null hypothesis and support the statement that there are significant differences in means between employees who attrited and the employees who remained at the company for the following factors: age, daily rate, distance from home, monthly income, total working years, number of trainings received

last year, number of years at the company, number of years at current role and number of years at current manager.

Variable	Attrited mean	Non-attrited mean	Diff	p-value
Age	**33.6**	**37.6**	**-4.0**	**6.5E-10**
Daily Rate	**750**	**813**	**-62**	**0.03**
Distance From Home	**10.6**	**8.9**	**1.7**	**2.7E-03**
Hourly Rate	65.6	66.0	-0.4	0.79
Monthly Income	**4,787**	**6,833**	**-2,046**	**5.5E-10**
Monthly Rate	14,559	14,266	294	0.56
Number of company work at	2.9	2.6	0.3	0.10
Total Working Years	**8.2**	**11.9**	**-3.6**	**2.9E-11**
Number of trainings received last year	**2.6**	**2.8**	**-0.2**	**0.02**
Number of years at company	**5.1**	**7.4**	**-2.2**	**2.0E-07**
Numbers of years at current role	**2.9**	**4.5**	**-1.6**	**4.6E-10**
Number of years since last promotion	1.9	2.2	-0.3	0.21
Number of years with current manager	**2.9**	**4.4**	**-1.5**	**1.4E-09**
Salary hike percentage	15.1	15.2	-0.1	0.61

Figure 10: Summary of means for employee who attrited and non-attrited and p-value for comparison of means

Employees who attrited are on average younger, stay further from the workplace, have lower monthly income, fewer working years, received less training last year, have been company for a shorter period and have been in the current role and manager for a shorter number of years.

4.2. Test of Association

4.2.1. Methodology Overview

We will proceed to test if there are associations between the categorical variables with employee performance and attrition. Factors that will be examined for association are frequency of business travel, department, education level, education field, environment satisfaction, gender, job involvement, job level, job role, job satisfaction, marital status, overtime, performance rating, relationship satisfaction, stock option level, work life balance.

For the test of association, chi-squared test is conducted using R. Below is an example of the code used for the test.

```
result2 <- chisq.test(HRdata$WorkLifeBalance,HRdata$Attrition,correct=FALSE)
```

Figure 11: Code snippet for chi squared test for association

4.2.2. Performance

For employee performance, we do not have sufficient statistical evidence to reject the null hypothesis at 95% significant level and is unable to support the claim that there are association between employee performance and any of the other factor of interest.

Variable	p-value
Frequency of Business Travel	0.59
Department	0.45
Education level	0.65
Education field	0.84
Environment Satisfaction	0.56
Gender	0.60
Job Involvement	0.58
Job level	0.73
Job role	0.53
Job Satisfaction	0.26
Martial Status	0.91
Overtime	0.87
Relationship Satisfaction	0.58
Stock option level	0.59
Work life balance	0.81

Figure 12: P-values of chi squaretest of association with employee performance

4.2.3. Attrition

At 95% significant level, we have enough statistical evidence to reject the null hypothesis and support the claim that there is an association between attrition and the following factors: frequency of business travel, department, education field, environment satisfaction, job involvement, job level, job role, job satisfaction, marital status, overtime, stock option level and work-life balance.

Environment and Job Satisfaction: Based on the difference in observed value and expected value, a larger number of employees that attritted scored low in environment and job satisfaction. Running further study and analysis to try to understand what factors affect the dissatisfaction of employee in these areas could help improve employee retention.

Department: It was observed that the sales department had a higher than expected number of employees that attritted. Future analysis can be carried out to try to understand the factors that leads to sales department having the higher than expected attrition given that the department have a higher median income as noted earlier in the report.

Overtime and work life balance: It is noted that among that attritted, a larger than expected proportion of them need to overtime and rated their work life balance at the lowest level. This could be an area where the HR and senior management team could take proactive step to ensure employee retention.

Variable	p-value
Frequency of Business Travel	**5.6E-06**
Department	**4.5E-03**
Education level	0.55
Education field	**6.8E-03**
Environment Satisfaction	**5.1E-05**
Gender	0.26
Job Involvement	**2.9E-06**
Job level	**6.6E-15**
Job role	**2.8E-15**
Job Satisfaction	**5.6E-04**
Martial Status	**9.5E-11**
Overtime	**2.2E-16**
Performance Rating	0.91
Relationship Satisfaction	0.16
Stock option level	**4.4E-13**
Work life balance	**9.7E-04**

Figure 13: P-value of chi square test of association with employee attrition

		Observed - Expected	
No.	**Variables**	**Non-Attrited**	**Attrited**
1	Business Travel	+12 Non-Travel, +12 Travel Rarely	+24 Travel Frequently
2	Department	+22 Research & Department	+20 Sales
3	Education Field	+9 Life Sciences, +11 Medical	+9 Marketing, +10 Technical Degree
4	Environment Satisfaction	+11 High, +12 Very High	+26 Low
5	Job Involvement	+15 High, +10 Very High	+15 Low, +11 Medium
6	Job Level	+34 Level 2, +12 Level 4	+55 Level 1
7	Job Role	+12 Healthcare Representative, +11 Manager, +13 Manufacturing Director, +11 Research Director	+20 Laboratory Technician, +20 Sales Representative
8	Job Satisfaction	+22 Very High	+19 Low
9	Marital Status	+20 Divorced, +25 Married	+44 Single
10	Overtime	+60 No overtime	+60 Have overtime
11	Stock option level	+40 Level 1, +13 Level 2	+52 Level 0
12	Work life balance	+17 Better	+12 Bad

Figure 14: Selected residuals (observed count – expected count) for factors with significant association with employee attrition

4.3. Logistic Regression

As part of the application of the tool, we aim to predict the probability of attrition of individual employees based on various factors. Logistic regression is selected as the dependent variable (Attrition) is categorical with two possible outcomes (Yes or No).

4.3.1. Methodology Overview

Dataset is then randomized to produce training and testing set. Test set is used to access the accuracy of the model in making prediction.

```
glm_fit=glm(AttritionBin~ Age+BusinessTravel+DailyRate+Department+DistanceFromHome
            +Education+EducationField+EnvironmentSatisfaction+Gender+HourlyRate+JobInvolvement+JobLevel
            +JobRole+JobSatisfaction+MaritalStatus+MonthlyIncome+MonthlyRate+NumCompaniesWorked+OverTime
            +PercentSalaryHike+PerformanceRating+RelationshipSatisfaction+StockOptionLevel
            +TotalWorkingYears+TrainingTimesLastYear+WorkLifeBalance+YearsAtCompany+YearsInCurrentRole
            +YearsSinceLastPromotion+YearsWithCurrManager, family = binomial, data = train
            , contrasts = list(Education = 'contr.treatment', EnvironmentSatisfaction = 'contr.treatment',
                               JobInvolvement ='contr.treatment',JobLevel = 'contr.treatment',
                               JobSatisfaction = 'contr.treatment', RelationshipSatisfaction = 'contr.treatment',
                               StockOptionLevel = 'contr.treatment',WorkLifeBalance = 'contr.treatment'))
# Contrast level for ordered variable change to contr.treament from the default polynomial orthogonal contrasts
# as we are unable to assume the variable are equally spaced which is required for polynomial contrasts

summary(glm_fit)
```

```
#Accuracy for testing set
glm_prob<-predict(glm_fit,newdata = test,type = "response")
glm_pred <- ifelse(glm_prob > 0.5, "Yes", "No")
table(glm_pred, test$Attrition)
mean(glm_pred==test$Attrition)
```

Figure 15: Code snippet for logistic regression and calculation of accuracy

Using all the relevant variables, the model has AIC of 751 and an accuracy of 88% when tested with the test data set. To further improve the model fit and accuracy, stepwise selection is done to find the subset of variables that results in the best performing model.

Below is the outcome after the step AIC. The revised model has the following significant variables with a reduced AIC (730) and improved accuracy (89%).

```
library(MASS)

step<-stepAIC(glm_fit,direction="both")
summary(step)
```

Figure 16: Code snippet for stepwise selection

4.3.2. Significant Variables

```
Call:
glm(formula = AttritionBin ~ BusinessTravel + DailyRate + DistanceFromHome +
    EducationField + EnvironmentSatisfaction + Gender + JobInvolvement +
    JobLevel + JobRole + JobSatisfaction + NumCompaniesWorked +
    OverTime + RelationshipSatisfaction + StockOptionLevel +
    TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
    YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
    YearsWithCurrManager, family = binomial, data = train, contrasts = list(Education = "contr.treatment",
    EnvironmentSatisfaction = "contr.treatment", JobInvolvement = "contr.treatment",
    JobLevel = "contr.treatment", JobSatisfaction = "contr.treatment",
    RelationshipSatisfaction = "contr.treatment", StockOptionLevel = "contr.treatment",
    WorkLifeBalance = "contr.treatment"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6630  -0.4423  -0.2067  -0.0603   3.3491

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    3.2923063  1.5064567   2.185 0.028855 *
BusinessTravelTravel_Frequently 2.2767641  0.5055168   4.504 6.67e-06 ***
BusinessTravelTravel_Rarely     1.3340437  0.4655890   2.865 0.004166 **
DailyRate                      -0.0005344  0.0002598  -2.057 0.039715 *
DistanceFromHome                0.0507604  0.0128267   3.957 7.58e-05 ***
EducationFieldLife Sciences    -0.8521770  1.0670383  -0.799 0.424501
EducationFieldMarketing        -0.4211937  1.1054355  -0.381 0.703188
EducationFieldMedical          -1.0027311  1.0717801  -0.936 0.349492
EducationFieldOther            -0.7451448  1.1317920  -0.658 0.510296
EducationFieldTechnical Degree  0.3302587  1.0902389   0.303 0.761948
EnvironmentSatisfaction2       -1.1806823  0.3316389  -3.560 0.000371 ***
EnvironmentSatisfaction3       -1.5061196  0.3067886  -4.909 9.14e-07 ***
EnvironmentSatisfaction4       -1.4021386  0.2980276  -4.705 2.54e-06 ***
GenderMale                      0.4610887  0.2165852   2.129 0.033262 *
JobInvolvement2                -1.0423919  0.4305839  -2.421 0.015483 *
JobInvolvement3                -1.3529497  0.4045487  -3.344 0.000825 ***
JobInvolvement4                -1.7754140  0.5601423  -3.170 0.001527 **
JobLevel2                      -1.2729803  0.4744165  -2.683 0.007291 **
JobLevel3                       0.0364823  0.6081028   0.060 0.952161
JobLevel4                      -1.2531901  1.0612624  -1.181 0.237663
JobLevel5                       1.4407977  1.3744672   1.048 0.294519
JobRoleHuman Resources          0.0491885  0.8693790   0.057 0.954881
JobRoleLaboratory Technician    0.6543695  0.6312814   1.037 0.299935
JobRoleManager                 -1.1542067  1.0886046  -1.060 0.289025
JobRoleManufacturing Director   0.4186957  0.6044032   0.693 0.488471
JobRoleResearch Director       -1.9840976  1.0909554  -1.819 0.068960 .
JobRoleResearch Scientist      -0.3199454  0.6505341  -0.492 0.622847
JobRoleSales Executive          1.0552905  0.5094599   2.071 0.038322 *
JobRoleSales Representative     1.2770860  0.7024562   1.818 0.069060 .
JobSatisfaction2               -0.7425631  0.3209725  -2.313 0.020696 *
JobSatisfaction3               -0.8732638  0.2880627  -3.032 0.002433 **
JobSatisfaction4               -1.3644225  0.3024638  -4.511 6.45e-06 ***
NumCompaniesWorked              0.2095026  0.0462317   4.532 5.85e-06 ***
OverTimeYes                     2.0978475  0.2306034   9.097  < 2e-16 ***
RelationshipSatisfaction2      -0.9253283  0.3218382  -2.875 0.004039 **
RelationshipSatisfaction3      -1.1274101  0.2980196  -3.783 0.000155 ***
RelationshipSatisfaction4      -0.9876899  0.2919903  -3.383 0.000718 ***
StockOptionLevel1              -1.4831782  0.2420171  -6.128 8.88e-10 ***
StockOptionLevel2              -1.6951376  0.4410422  -3.843 0.000121 ***
StockOptionLevel3              -0.8914842  0.4870929  -1.830 0.067218 .
TotalWorkingYears              -0.1188105  0.0316345  -3.756 0.000173 ***
TrainingTimesLastYear          -0.1858238  0.0864149  -2.150 0.031526 *
WorkLifeBalance2               -1.0554723  0.4272418  -2.470 0.013495 *
WorkLifeBalance3               -1.5542196  0.4006482  -3.879 0.000105 ***
WorkLifeBalance4               -1.3120403  0.4963794  -2.643 0.008212 **
YearsAtCompany                  0.1352538  0.0463901   2.916 0.003550 **
YearsInCurrentRole             -0.1674056  0.0569413  -2.940 0.003282 **
YearsSinceLastPromotion         0.1826927  0.0507045   3.603 0.000314 ***
YearsWithCurrManager           -0.1594729  0.0545449  -2.924 0.003459 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1053.26  on 1175  degrees of freedom
Residual deviance:  632.19  on 1127  degrees of freedom
AIC: 730.19
```

Figure 17: Output for best fit model after stepwise selection

Business Travel: Travelling frequently and travelling rarely are significantly associated with a 2.27 and 1.33 increase in the log odds of attrition respectively, relative to those who do not travel for business. Hence, business travel seems to be one of the key factors affecting attrition. The need for business travel might mean less time spent with family which might be one of the reasons for attrition.

Overtime: Overtime is another factor that is statistically significant with the log odds of attrition for employees that have to work overtime increasing by 2.1. The need to overtime might result in the lack of work life balance and fatigue among employee which in turn led to a motivation for employee to seek new opportunities.

DistanceFromHome: This is a continuous variable. Every 1 unit of distance away from home will increase in the log odds of Attrition by 0.05. Hence, longer daily traveling distance to work is associated with higher attrition likely due to greater inconvenience and less personal time.

Job Involvement/Job Satisfaction: These ordered factors are significantly associated with the log odds of Attrition with negative coefficient. The absolute value of the coefficient increases with higher satisfaction level. The higher job involvement and job satisfaction will tend to decrease the chance of attrition.

Environment Satisfaction/Relationship Satisfaction/Worklife Balance: Although these factors are negatively associated with Attrition, 3 has highest absolute value of coefficient, followed by 4 and then 2. Hence, assuming other variables remain the same, having level 3 for these factors will reduce the log odds of Attrition the most.

Stock Option Level: Similarly, while this factor is negatively associated with Attrition, level 2 has the highest value for the coefficient, followed by level 1. The coefficient for level 3 is not statistically significant (95% significance level).

Job Roles: Looking at the various job roles, only the coefficient of Sales Executives is significantly associated with Attrition at 95% significantly level. This could be due to Sales roles having more transferrable skillsets that enable the employees to seek new opportunities more easily in another company/industry vs the other more technical roles that are research related. Additionally, Sales Executives generally have higher job level vs Sales representatives which might mean that they are more experienced in the domain of sales, which in turn increases their chance of getting new job opportunities.

Other variables: The more companies worked at previously, longer years at company and longer years since last promotion are factors that are positively associated with Attrition. While total working years, number of training times last year, years in current role, years with current manager and daily rate are negatively associated with Attrition.

4.3.3. Limitations

Imbalanced dataset

The dependent variable (Attrition) comprised of 1233 “No” and 237 “Yes” with a roughly 5:1 ratio. This imbalanced dataset has resulted in forming an imbalanced prediction model. From the confusion matrix below, it is observed that the accuracy predicting “No” for attrition is around 244/(244+7) ≈ 97%, while accuracy for predicting “Yes” is around 18/(25+18) ≈ 42%. Therefore, the model is more accurate at predicting 'No' than predicting 'Yes'.

```
glm_pred  No Yes
      No  244  25
      Yes   7  18
> mean(glm_pred==test$Attrition)
[1] 0.8911565
```

Figure 18: Confusion matrix of predicted vs observed outcome

4.3.4. Application

The predictive logistic model is built as a feature of the Shiny App, allowing user to choose or key in parameter values. After the user has filled up all the parameters and press on the "Predict" button, the probability of attrition will be shown on the bottom of the screen.

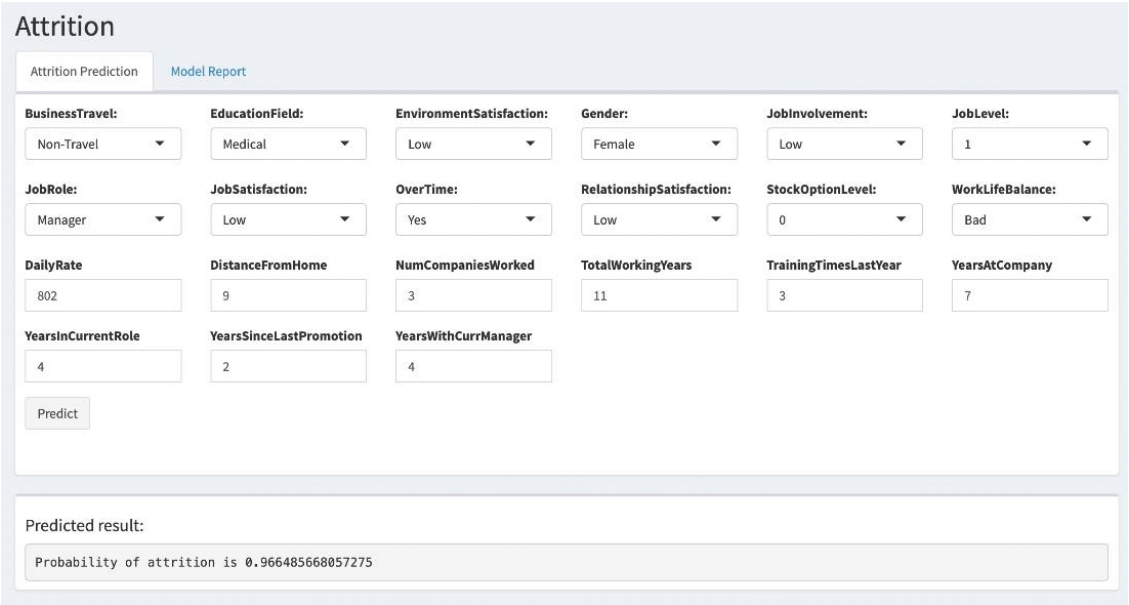


Figure 19: Feature designed in Shiny App to predict the probability of attrition for individual

5. Conclusion & Recommendations

This project aimed to enable HR professionals to derive insights and determine the factors that the company's policies should focus on. Besides designing various descriptive charts to further understand the profile of employees, comparison of mean (using z-test) and test of association (using Chi-Sq) were done to assess the association between individual factors vs performance/attrition. Lastly, logistic regression was carried out to predict the likelihood of attrition for individual employees using statistically significant key factors.

Further studies to understand the potential reasons of why a rating of 3 (vs a rating of 4) for Environment Satisfaction/Relationship Satisfaction/Worklife Balance factors is associated with the highest decrease in the log odds of Attrition in the predictive analysis, might also help in retention of employee who are high satisfied with the culture and environment in Company X.

Factors associated with performance is another aspect that we have explored on beyond attrition. Unfortunately, we are unable to find a difference in mean for all factors when we compare average vs great performer. Additionally, salary hike percentage was the only statistically significant factor in the Chi-Square test, and it is more likely that performance is the factor affecting salary hike. This could be due to the limitations of data such that employees are either graded as 3 (average) or 4 (great). We would recommend Company X to design performance appraisal system that

allows greater differentiation among employees. With that, further analysis could be to understand factors associated with outstanding performance. In addition, other factors such as GPA, cognitive test scores, personality/psychometric test results could also be collected to see if they are associated with and are predictors for employee performance.

By taking proactive actions, Company X stands to benefit from increasing the retention rate of outstanding employees who can continue to contribute at a high-performance level over a longer period.

6. References

Spain, E., & Groysberg, B. (2016, April). Making Exit Interviews Count. Retrieved from Harvard Business Review: https://hbr.org/2016/04/making-exit-interviews-count

Deloitte University Press. (2017). Rewriting the rules for the digital age. Retrieved from 2017 Deloitte Global Human Capital Trends: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/About-Deloitte/central-europe/ce-global-human-capital-trends.pdf

Netabai, L. (2020, May 19). Talent retention: why traditional methods are no longer enough. Retrieved from BusinessChief: https://businesschief.eu/leadership-and-strategy/talent-retention-why-traditional-methods-are-no-longer-enough-1

Appendix A – R Packages

Below is the list of R packages that were used for our project:

- shiny & shinydashboard: For building the Shiny App
- dplyr & tidyverse: For data preprocessing and manipulation
- ggplot2, ggpubr, ggmosaic, cowplot: To provide visualization tools such as boxplot, bar charts, scatterplots, mosaic plots etc.
- MASS: To run StepAIC for model selection
- BSDA & PASWR: To use for hypothesis testing

Appendix B – Detailed description of each variable

Column name (Variables)	Description	Data type
Age	Age of employee	Quantitative
Attrition	Whether the employee has left the company	Qualitative
BusinessTravel	Frequency of business trip	Qualitative
Hourly/Daily/Monthly Rate	Cost of employee*	Quantitative
Department	Department of the employee	Qualitative
DistanceFromHome	Distance from company to the employee's home	Quantitative
Education	Education level of the employee	Qualitative
EnvironmentSatisfaction	Satisfaction of the working environment	Qualitative
Gender	Gender of employee	Qualitative
JobInvolvement	Level of involvement of employee	Quantitative
JobLevel	Job level of employee	Qualitative
JobRole	Title/Position of employee	Qualitative
JobSatisfaction	Level of job satisfaction	Qualitative
MaritalStatus	Marital status of employee	Qualitative
Monthly Income	Monthly salary of employee	Quantitative
NumCompaniesWorked	Number of companies the employee has worked at prior to current company	Quantitative
OverTime	Whether the employee has worked overtime	Qualitative
PercentSalaryHike	Percentage increase in salary	Quantitative
PerformanceRating	Performance rating of employee	Qualitative
RelationshipSatisfaction	Level of satisfaction of interpersonal relationship	Qualitative
StockOptionLevel	Tier of compensation package	Qualitative
TotalWorkingYears	Number of years joining workforce	Quantitative
TrainingTimesLastYear	Number of hours spent on training last year	Qualitative
WorkLifeBalance	Whether the employee have work life balance	Qualitative
YearsAtCompany	Number of years working at the company	Quantitative
YearsinCurrentRole	Number of years in current working position	Quantitative
YearsSinceLastPromotion	Number of years since last promotion	Quantitative
YearsWithCurrManager	Number of years working under the current manager	Quantitative

*Cost of employee includes cover salary, social insurance, administration, logistics etc.