
EMPLOYEE ATTRITION ANALYSIS & PREDICTION

Branata Kurniawan (34534388)
Ng Zhi Hui (34557319)



TABLE OF CONTENT

- BACKGROUND
- ABOUT THE DATA
- DATA PREPARATION
- ALGORITHMS
- DATA EXPLORATORY
- FINDINGS
- CONCLUSION



BACKGROUND

01

CHANGES IN EMPLOYEES' BEHAVIOURS

COVID-19 pandemic has significantly impacted employee behaviour, prioritising the work-life balance, better compensation, and many others.

02

MAINTAINING ATTRITION RATE

Due to the above changes there is a need for organisations to maintain a low attrition rate.



ABOUT THE DATA

- **Link:** <https://www.kaggle.com/datasets/thedevastator/employee-attrition-and-factors>
 - Containing employees' personal information that focus on areas such as employee attrition, personal and job-related factors, and also financials.
 - **Target variable:**
 - Attrition
 - **Feature variables:**
 - Consist of 34 both categorical and continuous variables
 - Such as: Age, Gender, Marital Status, Business Travel Frequency, Daily Rate of Pay, Departmental Information such as Distance From Home Office or Education Level Obtained by the employee in question, and others.
-

DATA PREPARATION

- **Data cleaning**
 - By omitting the NA values out of the data frame
- **Data pre-processing**
 - **Label encoding:**
 - Transforming non-numeric variables into numeric variables before they can be used.
 - **Checking correlation:**
 - Checking the correlation between the features against target variables, and then remove the features with empty (NA) correlation.
- **Incorporating n-fold cross validation**
 - Using 5-fold cross validation to obtain more robust and reliable performance results



DATA CLEANING

```
# Read the data
#data <- read_csv("/Users/zhakaeric-macairm1/Downloads/ICT515 Source Code/HR_Analytics.csv")
data <- read_csv("/Users/branata.kurniawan/Documents/personal/ICT515/HR_Analytics.csv")
glimpse(data)

# Cleaning the data
data <- na.omit(data)
```

- Right here after reading the data, we omit any NA value out of the data table
- This would ensure that incomplete data are removed
 - more effective than manual verification
 - time-consuming and error-prone.



DATA PRE-PROCESSING



DAT
A

LABEL ENCODING

```
# BusinessTravel  
levels(data$BusinessTravel) <- seq(length(levels(data$BusinessTravel)))  
data$BusinessTravel <- as.numeric(data$BusinessTravel)
```

We will be doing the same process for: Department, Education Field, Gender, Job Role, Marital Status, and Over Time

CHECKING CORRELATION

```
# Calculate the variable correlation value with Attrition  
correlation_table <- cor(data %>% select(Attrition, everything())) %>% as.data.frame()  
correlation_table <- correlation_table[, 1, drop=FALSE] %>% arrange(-Attrition)
```

DATA PRE-PROCESSING (CONT)

VARIABLES	ATTRITION
OVERTIME	0.25
MARITALSTATUS	0.16
DISTANCEFROMHOME	0.08
JOBROLE	0.07
DEPARTMENT	0.06
...	
YEARSINCURRENTROLE	-0.16
JOBLEVEL	-0.17
TOTALWORKINGYEARS	-0.17
EMPLOYEECOUNT	NA
OVER18	NA
STANDARDHOURS	NA


```
# Drop the column without correlation  
data <- data[, -which(names(data) == "EmployeeCount")]  
data <- data[, -which(names(data) == "Over18")]  
data <- data[, -which(names(data) == "StandardHours")]
```

Drop the column that have NA or zero correlation.

INCORPORATING N-FOLD CROSS VALIDATION

- We chose five-fold cross-validation since it has been shown to have low bias and moderate variance in predicting model performance.

□ Logistic Regression



```
# Train the model using N-Fold Cross Validation
control <- trainControl(method="cv", number=5)
lr <- train(Attrition~., data=trainData, method="glm", trControl=control)
```

□ Random Forest

```
# Train the model using N-Fold Cross Validation
control <- trainControl(method = "cv", number = 5)
rf <- train(Attrition~., data=trainData, method="rf", trControl=control, tuneLength=3, ntree=750, type='classification')
```

ALGORITHMS



LOGISTIC REGRESSION



RANDOM FOREST



LOGISTICS REGRESSION

LR commonly use for the prediction of a binary outcome or multiclass classification. The capabilities of LR are:

- The ability to identify the common traits of employee churning
 - E.g. the most important variables that are related with the binary outcome, which help in identifying the key variable for prediction or classification
- Predict whether an individual employee will resign or stay in the company

Process:

1. Data Cleaning → 2. Removal of Non-Correlation → 3. N-Fold Cross Validation → 4. Model Training & Evaluation

RANDOM FOREST

Random Forest uses ensemble learning methods. The benefits of using random forest are:

- Improve model accuracy and robustness
- Handle high-dimensional data well
- Perform feature selection by measuring feature importance.
- Less prone to overfitting

Process:

1. Data Cleaning → 2. Removal of Non-Correlation → 3. N-Fold Cross Validation

→ 4. Model Training & Evaluation → 5. Tuning of Hyperparameters → 6. Retrain with Hyperparameters

RANDOM FOREST IMPROVEMENT

PERFORMANCE	RF500	RF750	RF1000
ACCURACY	0.8163	0.8231	0.8197
95% CI	[0.7672 , 0.8589]	[0.7746 , 0.865]	[0.7709 , 0.862]
KAPPA	0.1687	0.1841	0.1763
F1	0.8958	0.9	0.8979
MSE	0.1837	0.1769	0.1803
MAE	0.1837	0.1769	0.1803
PRECISION RATE	0.9789	0.9873	0.9831
RECALL RATE	0.8256	0.8269	0.8262

PERFORMANCE COMPARISON

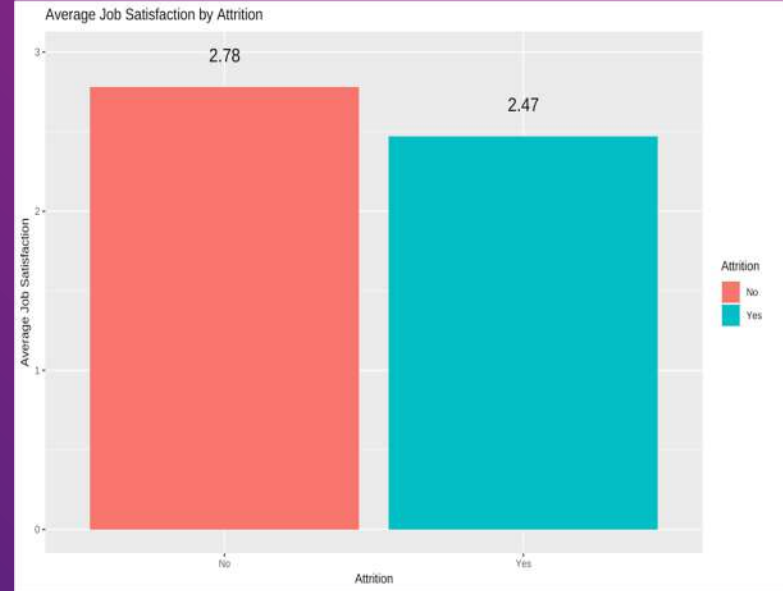
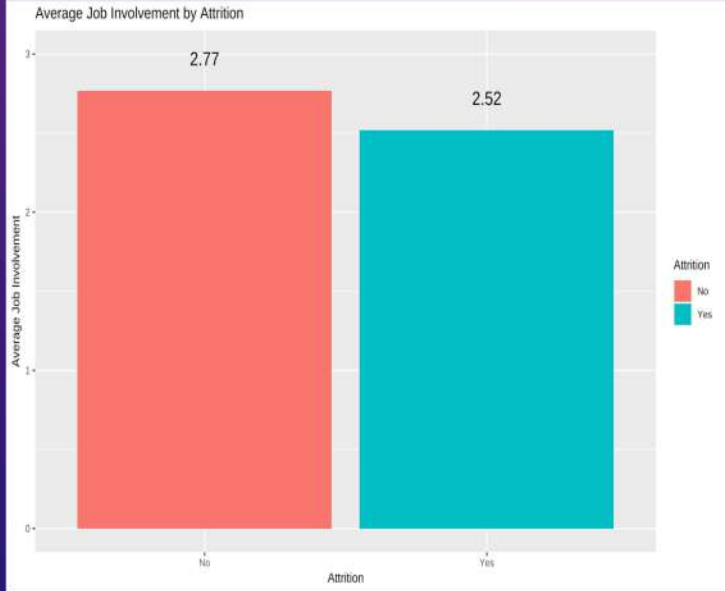


PERFORMANCE	LOGISTIC REGRESSION	RANDOM FOREST (750)
ACCURACY	0.8197	0.8231
95% CI	[0.7709 , 0.862]	[0.7746 , 0.865]
KAPPA	0.1265	0.1841
F1	0.899	0.9
MSE	0.1803	0.1769
MAE	0.1803	0.1769
PRECISION RATE	0.9958	0.9873
RECALL RATE	0.8194	0.8269

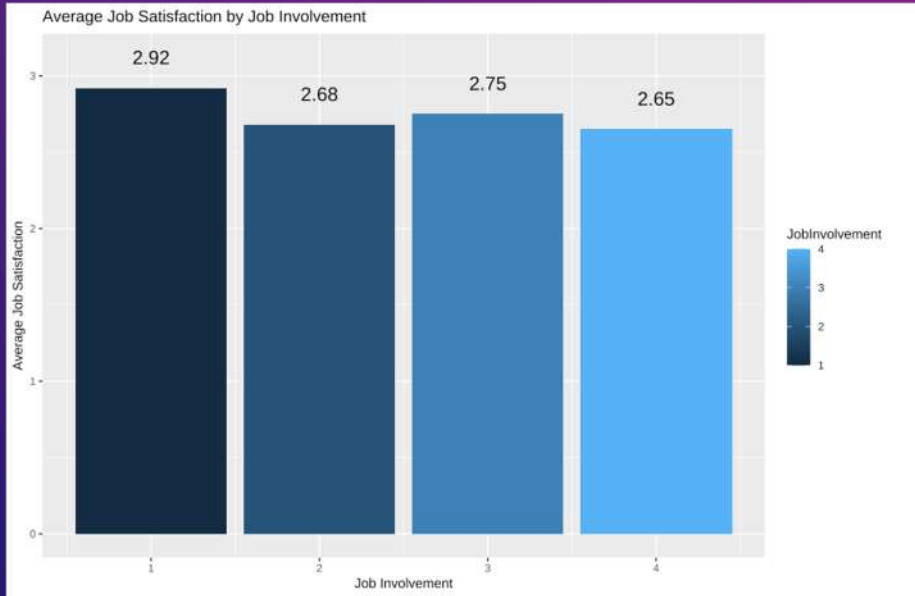
DATA EXPLORATORY



JOB SENTIMENT



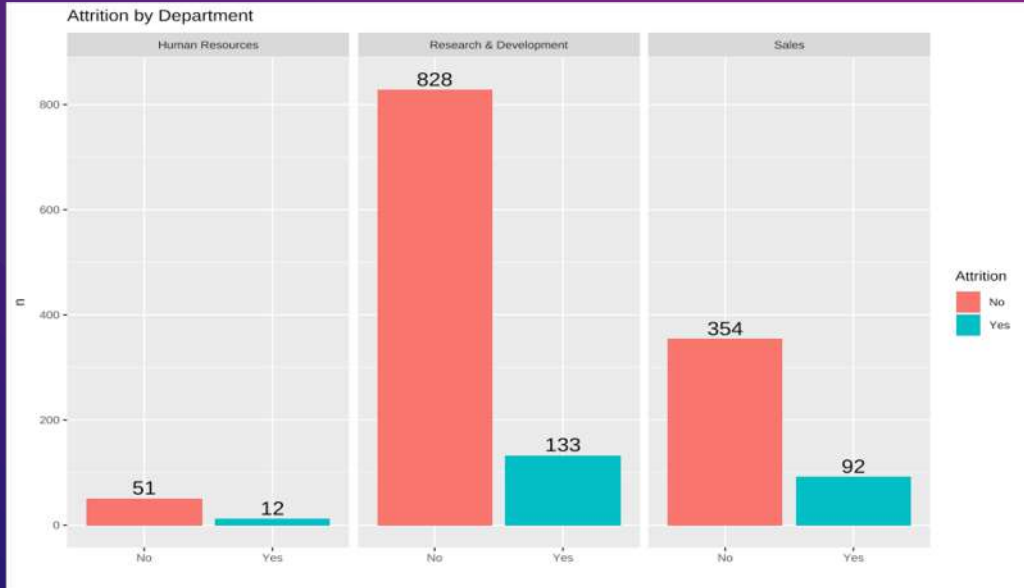
JOB SENTIMENT



Group 1 & 3 is deemed to be less prone to attrition

- Above the respective attrition threshold figure

DEPARTMENT BREAKDOWN



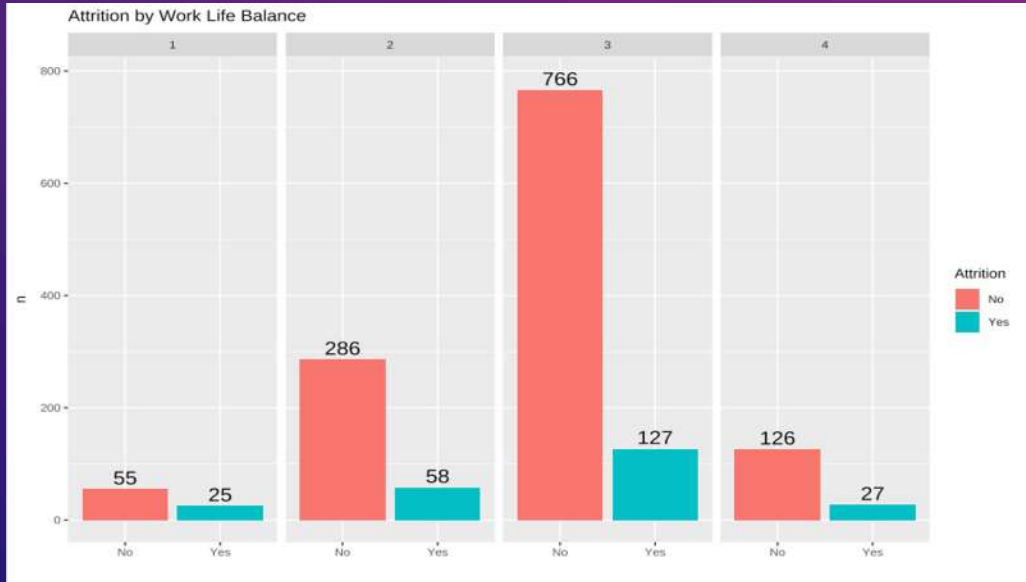
Research and Development have the highest attrition figure

- Substantial turnover
- Underlying factors

Sales have the highest attrition rate of 21%

- Unobservable from raw figure

WORK LIFE BALANCE



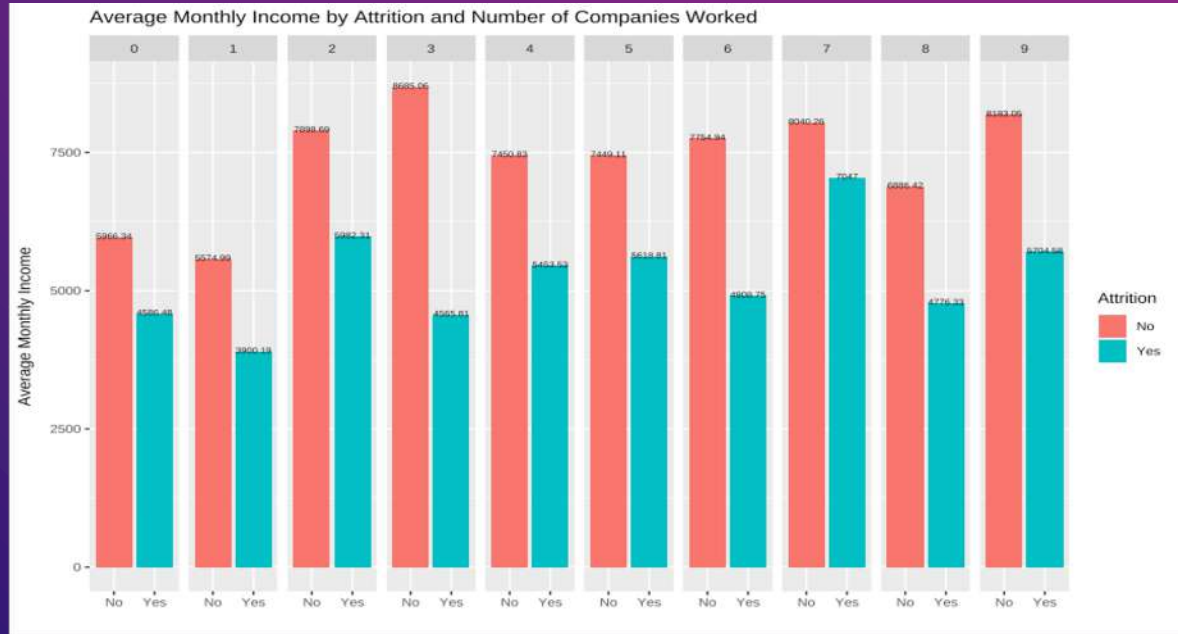
Group 3 have the most significant attrition by count

- Lowest by ratio (14%)

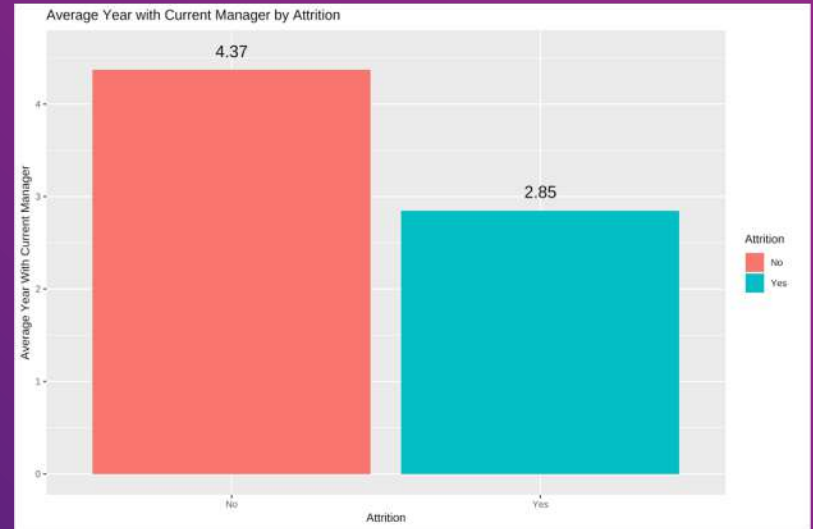
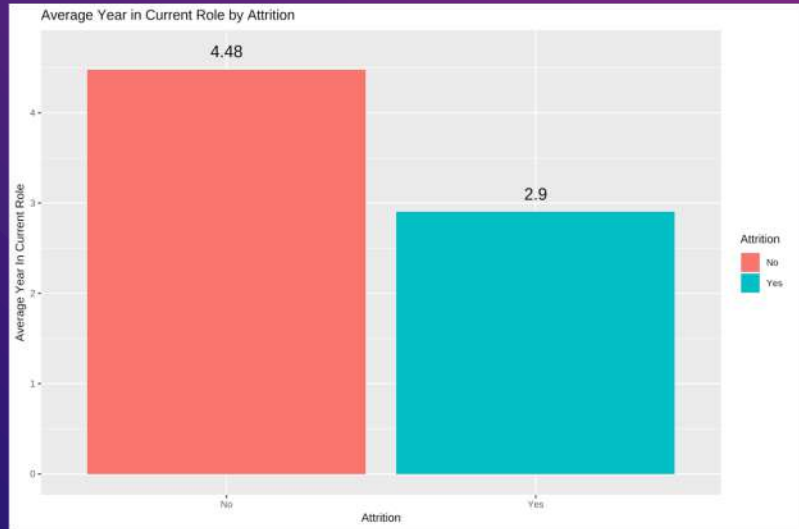
Group 1 has the highest attrition

- Attrition rate of 31%

WORK LIFE BALANCE



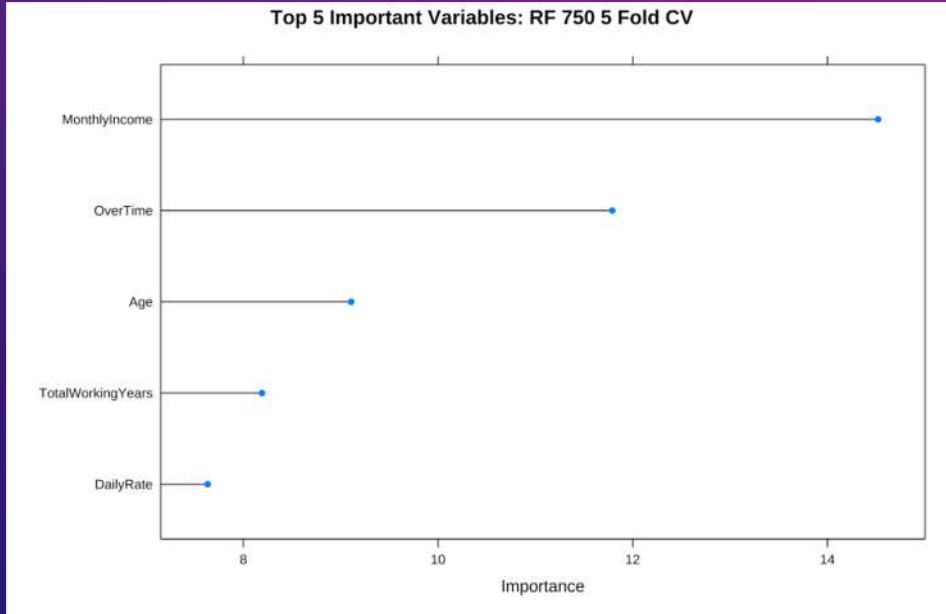
WORK RELATIONS



FINDINGS

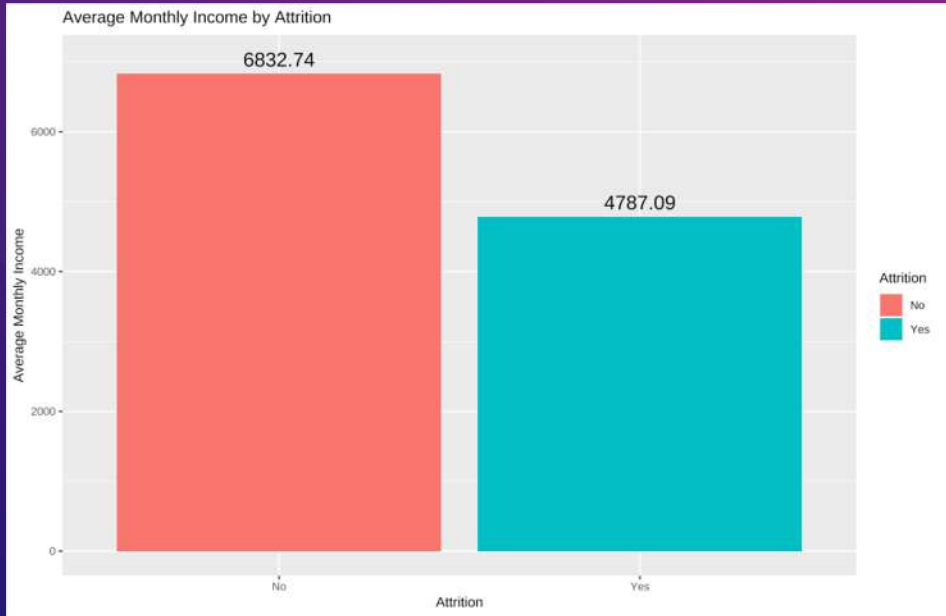


TOP 5 IMPORTANT VARIABLE OF RANDOM FOREST



1. Monthly Income
2. Overtime
3. Age
4. Total Working Years (Experience)
5. Daily Rate

MONTHLY INCOME

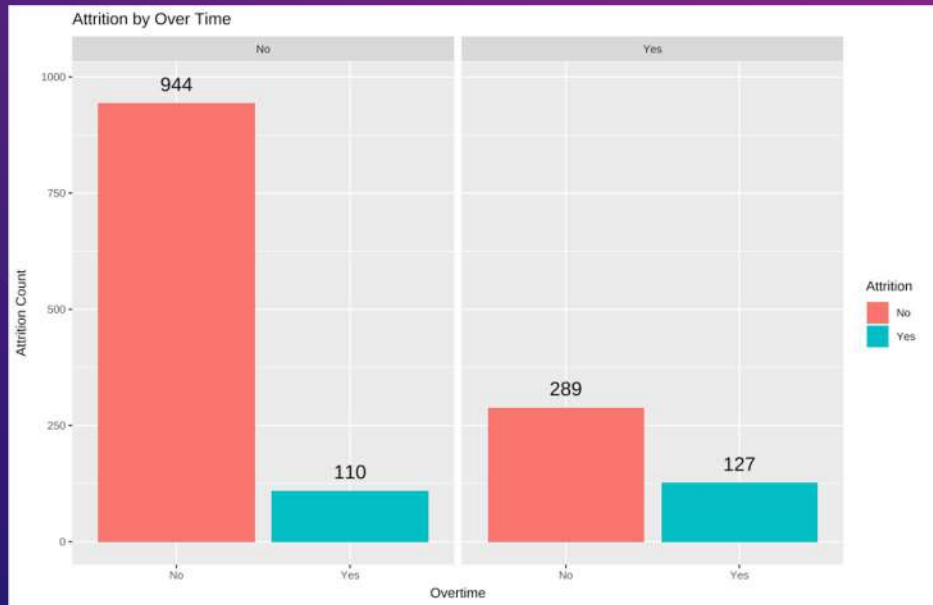


Salary has and will always be a key factor

Average: \$4,787.09

Median: \$3,202

OVERTIME



Experienced OT / overworked

- People (Stay): 23.44%
- People (Leave): 53.59%

AGE



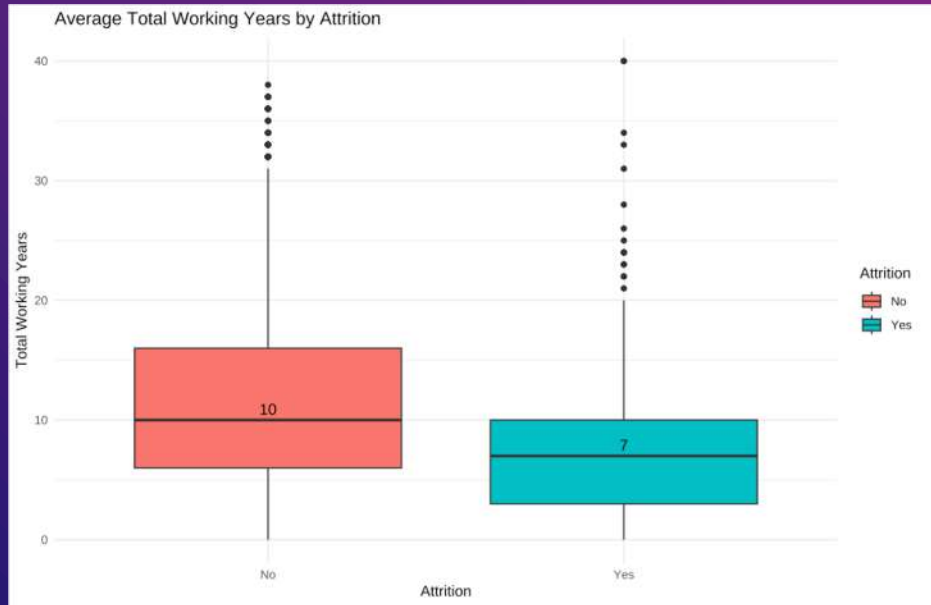
People are more likely to switch jobs

Younger Talent Pool

- Average: 33
- Median: 32

Lower maximum range

TOTAL WORKING YEARS



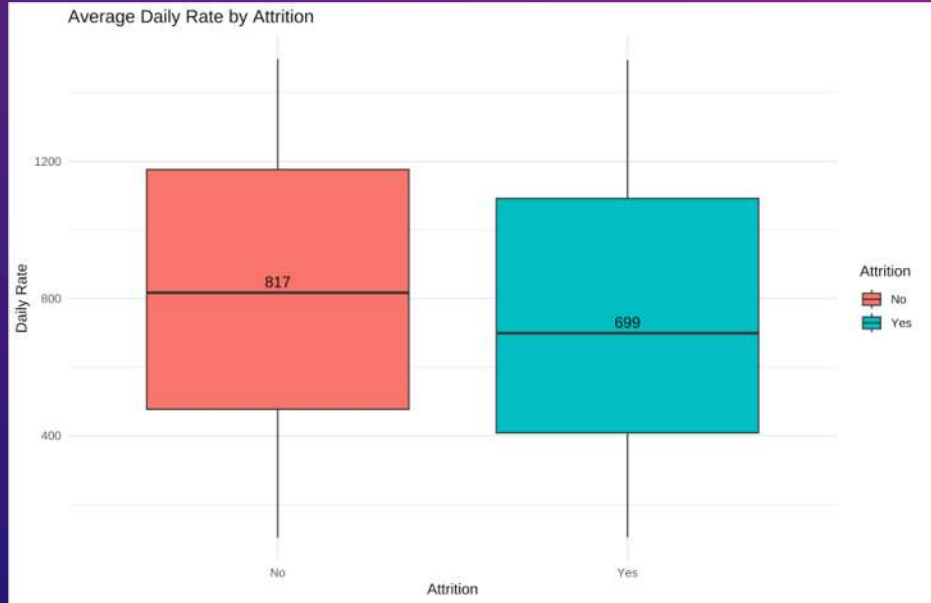
Total Working Years of the employee plays an significant impact

Average: 8

Median: 7

Much lower cap

DAILY RATE



Financial Concern is a repeating key factor

Resignees tend to have a lower compensation

- Average: \$750
- Median: \$699

Differentiator: Efficient Rate (Subjective)



CONCLUSION

Essential Traits Identification through
Logistic Regression vs Random Forest

CONCLUSION



VS



LOGISTIC REGRESSION

- Both able to identify traits and predict binary outcomes
- Logistics Regression assumes Linearity; Random Forest uses ensemble learning
- Random Forest would be better suited for larger and more complex data

RANDOM FOREST

APPLICATION

- Successfully Identification of top traits
- Assist organisations in retaining key valuable employees and reduce employee churn & hiring cost

* Note: Algorithm is highly dependent on unique requirement

DO YOU HAVE ANY QUESTIONS?

