Casestudy

Shaik Roshan Baba

1. Employee Attrition Analysis

A company aims to analyze and predict which employees are likely to leave (attrition) using data on their **age**, **salary**, and **years at the job**.
The attrition status is recorded as a binary variable ("Yes" for leaving, "No" for staying).

---

a. Suitable Regression Model and Justification

Selected Model: Logistic Regression

Justification: - **Binary Outcome:** The target variable, *Attrition*, is binary ("Yes" or "No"). Logistic regression is ideal for this type of problem.
- **Probability Prediction:** Logistic regression predicts the *probability* (between 0 and 1) that an employee will leave, unlike linear regression which predicts continuous values.
- **Interpretability:** The coefficients can be interpreted in terms of *odds ratios*, helping understand how each variable affects attrition likelihood.

---

b. R Code Implementation

The R code below: 1. Creates a sample employee dataset.
2. Fits a logistic regression model.
3. Generates predictions (both probabilities and Yes/No classifications).

```r
# 1. Data Preparation
set.seed(123)
n <- 300

employees <- data.frame(
  Age = sample(22:60, n, replace = TRUE),
  Salary = round(runif(n, 40000, 150000), -2),
  YearsOnJob = sample(1:30, n, replace = TRUE)
)

prob <- 1 / (1 + exp(-(1 - (employees$Salary * 0.00001)
                       - (employees$YearsOnJob * 0.05)
                       + (employees$Age * 0.01))))
employees$Attrition <- ifelse(runif(n) < prob, "Yes", "No")
```

```r
employees$Attrition <- as.factor(employees$Attrition)

print("--- Sample Data (First 6 Rows) ---")
```

```
## [1] "--- Sample Data (First 6 Rows) ---"
```

```r
head(employees)
```

```
##   Age Salary YearsOnJob Attrition
## 1  52 129500          9        No
## 2  36  73900         23        No
## 3  35 115700          7       Yes
## 4  24 142600         13        No
## 5  58  52700         29        No
## 6  35  54000         20        No
```

```r
# 2. Model Fitting (Logistic Regression)
attrition_model <- glm(Attrition ~ Age + Salary + YearsOnJob,
                       data = employees,
                       family = "binomial")

print("--- Model Summary ---")
```

```
## [1] "--- Model Summary ---"
```

```r
summary(attrition_model)
```

```
##
## Call:
## glm(formula = Attrition ~ Age + Salary + YearsOnJob, family = "binomial",
##     data = employees)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.104e+00  6.367e-01   1.734   0.0830 .
## Age          1.348e-02  1.181e-02   1.141   0.2538
## Salary      -8.610e-06  3.805e-06  -2.263   0.0236 *
## YearsOnJob  -7.292e-02  1.556e-02  -4.687 2.77e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 413.27  on 299  degrees of freedom
## Residual deviance: 382.05  on 296  degrees of freedom
## AIC: 390.05
##
## Number of Fisher Scoring iterations: 4
```

```r
# 3. Prediction
probabilities <- predict(attrition_model, newdata = employees, type = "response")
predictions <- ifelse(probabilities > 0.5, "Yes", "No")

print("--- Model Predictions (First 6) ---")
```

```
## [1] "--- Model Predictions (First 6) ---"
```

```r
head(predictions)
```

```
##     1     2     3     4     5     6
## "Yes"  "No" "Yes"  "No"  "No"  "No"
```

c. Model Evaluation

To evaluate the model's accuracy, I would use the Train/Test Split validation technique. 1. Split Data: Randomly divide the data into a training set (e.g., 70%) and a testing set (e.g., 30%). 2. Fit on Training Data: Build the glm model using only the training data. 3. Predict on Test Data: Use the trained model to make predictions on the testing data. 4. Evaluate: Create a Confusion Matrix to calculate metrics: * Accuracy: Overall percentage of correct predictions. * Precision: Percentage of predicted leaves that were correct. * Recall (Sensitivity): Percentage of actual leaves that were correctly identified.

---

d. Appropriateness of the Regression Approach

The logistic regression approach is highly appropriate for this task because: * It is purpose-built for binary outcomes ("Yes/No"). * It provides a probabilistic output, allowing the company to see how likely an employeeis to leave. * It is computationally efficient and highly interpretable.

---

2. R Markdown Syntax

a. Inline Mathematical Expression

The "code" to show the formula in a sentence is: The formula for the area of a rectangle is $A = l \times w$, where l is the length and w is the width.

The ... signs are what tell R Markdown to treat it as a formula.

---

b. Inserting an Image

The “code” to show an image named logo.png is:

Figure 1: Company Logo