COMPONENT-5

PREDICTIVE ANALYTICS ↘

Predicting Employee Attrition
(A machine learning approach)

PRESENTED BY:

NAME	PRN
Mohd. Amaan Khan	21021021177
Pragya Verma	21021021200
Priyanshi	21021021207
Priyanshu S. Kainth	21021021210
Udichi Pathak	21021021297

PRESENTED TO:

Ms. Kriti Priya Gupta

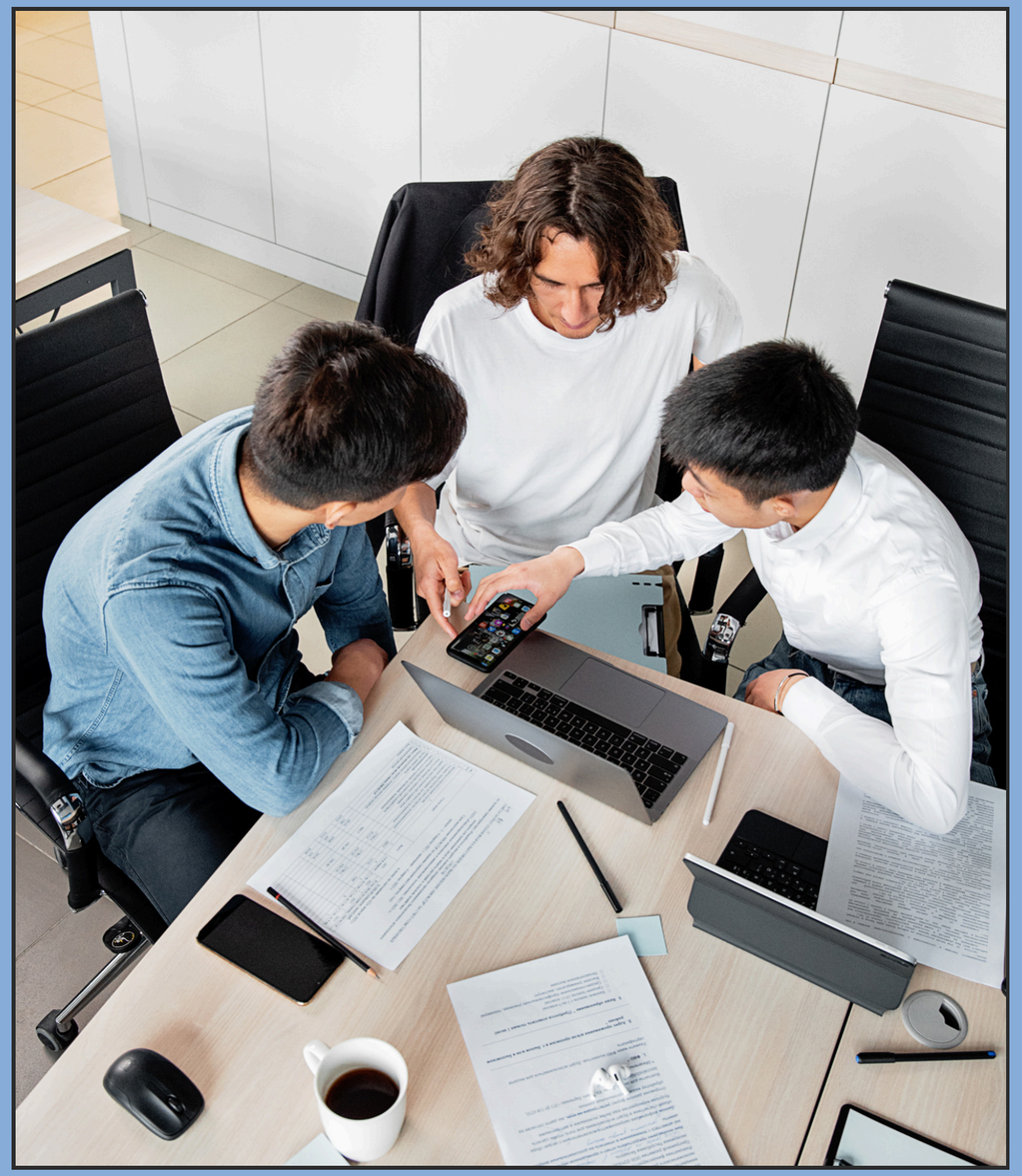
AGENDA

01	PROBLEM STATEMENT
02	VARIABLE DESCRIPTION
03	DATA CLEANING AND PROCESSING
04	MACHINE LEARNING MODELS 1,2,3,4
05	COMPARISION BETWEEN MODELS
06	RECOMMENDATION

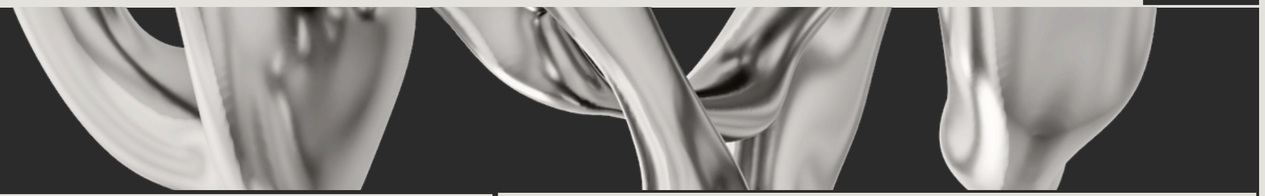
THE PROBLEM- HIGH EMPLOYEE ATTRITION

ABC Company: **Valuing Talent**

- **ABC Company** is a leading organization that prioritizes employee success.
- They believe their talented workforce is the key to achieving their goals. They are the ones who meet deadlines, drive sales, and build our brand through positive customer interactions.
- Unfortunately, they are facing a challenge of high employee attrition. When valuable employees leave, it disrupts their operations, leads to a loss of institutional knowledge, and incurs significant costs in recruitment and training new staff.
- To address this challenge, they are committed to using data science to predict which employees are at risk of leaving. By proactively identifying these "flight risks," they can take steps to retain our top talent and foster a thriving work environment.



VARIABLE DESCRIPTION



CATEGORICAL VARIABLES:

- **Department:** This indicates the employee's department within the company
- **EducationField:** This represents the employee's educational background
- **Gender:** This identifies the employee's gender (e.g., Male, Female).
- **MaritalStatus**: This indicates the employee's marital status

NUMERICAL VARIABLES:

- **Age**: Employee's current age (years).
- **Distance:** Distance from workplace
- **Job Level**: Employee's seniority level .
- **Monthly Income:** Employee's monthly salary/wages.
- **Companies Worked**: Number of previous employers.
- **Salary Hike (%):** Percentage increase in most recent salary review.
- **Total Work Experience (Years):** Total years of professional work.
- **Years at Company:** Years employed at ABC Company.
- **Years Since Last Promotion:** Years since last promotion at ABC Company.
- **Years with Current Manager:** Years working under current manager.

TARGET VARIABLE:

- **Attrition (Yes/No)**: This is the binary variable you're trying to predict. It indicates whether the employee has left the company (Yes) or is still employed (No).

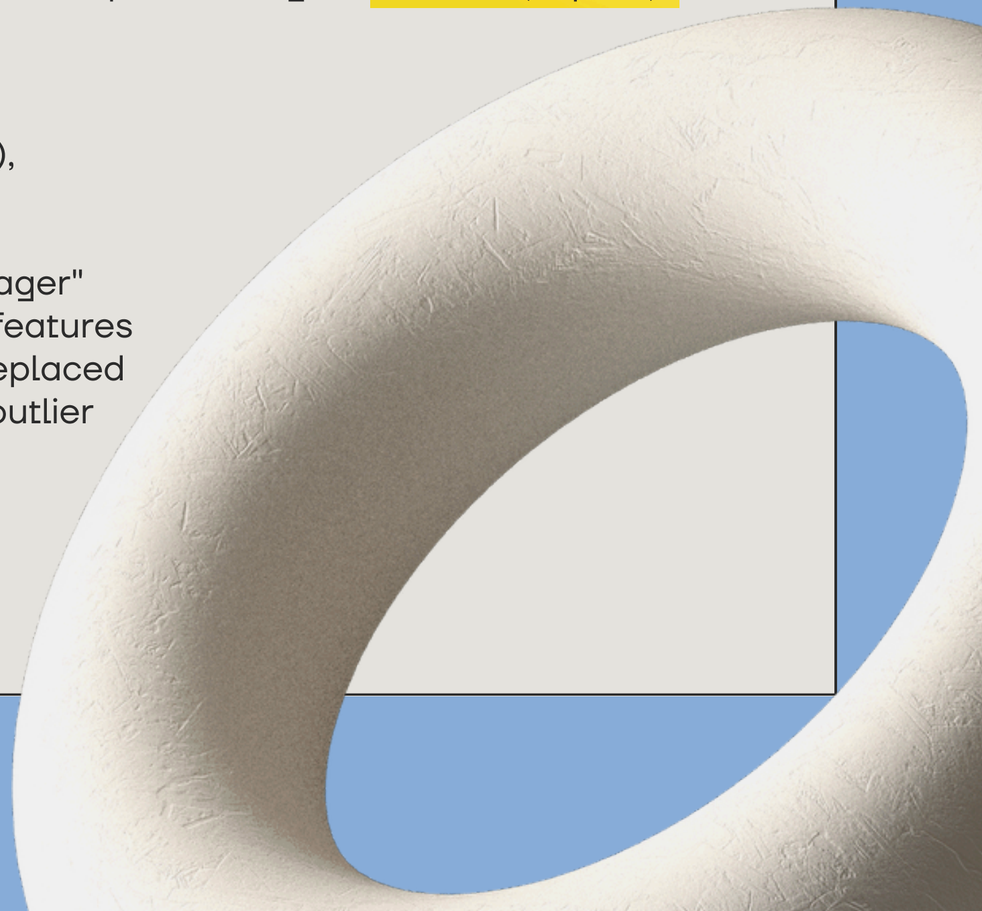
DATA CLEANING

This process involved cleaning the "EmployeeAttrition" dataset for use in machine learning models. Here's a summary of the key steps:

- **HANDLING MISSING VALUES:**
  We identified missing values using map(E,~sum(is.na(.))).
  Missing values in "YearsSinceLastPromotion" and "YearsWithCurrManager" were imputed using the median (impute).

- **TREATING OUTLIERS:**
  We identified outliers in several numerical features using boxplots.
  For positively skewed features (Monthly Income, NumCompaniesWorked, etc.), transformations were applied:
  - Square root for "Monthly Income" and "NumCompaniesWorked"
  - Logarithm with offset for "YearsSinceLastPromotion", "YearsWithCurrManager"
  - "YearsAtCompany", and "TotalWorkingYears" Outliers in the transformed features ("YAC", "TWY") were identified again using boxplots. These outliers were replaced with missing values (converted back to NA).Missing values introduced by outlier removal were again imputed using the median.

- **FEATURE ENGINEERING:**
  All categorical variables were converted to factors using lapply(E[,-c(1,5,7,11,12,13,14,15,16,17)],factor).



CHECKING ASSUMPTIONS

```
> skewness
$Age
[1] 0.4128645

$DistanceFromHome
[1] 0.95714

$PercentSalaryHike
[1] 0.8202899

$MI
[1] 0.8610798

$YSLP
[1] 0.7383711

$YWCM
[1] -0.3573206

$NCW
[1] 0.03315543

$YAC
[1] 0.03012916

$TWY
[1] 0.02771128
```

```
> kurtosis
Age
[1] 2.593149

DistanceFromHome
[1] 2.771852

PercentSalaryHike
[1] 2.696344

MI
[1] 2.882922

YSLP
[1] 2.652369

YWCM
[1] 2.090585

NCW
[1] 2.374395

YAC
[1] 2.492147

TWY
[1] 2.769275
```

All variables are approximately normally distributed as their skewness and kurtosis values are with the recommended criteria (**absolute skewness value ≤ 1 and absolute excess kurtosis ≤ 4)**

```
> x<-E[,c(1,5,13,18,19,20,21,22,23)]
> cor(x)
                          Age DistanceFromHome PercentSalaryHike           MI         YSLP         YWCM          NCW         YAC
Age                1.000000000      0.006963332      -0.033136611 -0.044540140  0.19672812  0.17551499  0.30967584  0.24242330
DistanceFromHome   0.006963332      1.000000000       0.038124615 -0.021893036  0.00618813  0.03529100 -0.02926698  0.02404565
PercentSalaryHike -0.033136611      0.038124615       1.000000000  0.006272912 -0.03232840 -0.02629922  0.01378349 -0.01155010
MI                -0.044540140     -0.021893036       0.006272912  1.000000000  0.06679718  0.02995299 -0.02620271  0.01564094
YSLP               0.196728119      0.006188130      -0.032328404  0.066797179  1.00000000  0.49364196 -0.05921330  0.54222015
YWCM               0.175514993      0.035291004      -0.026299218  0.029952992  0.49364196  1.00000000 -0.12787463  0.77947262
NCW                0.309675842     -0.029266982       0.013783491 -0.026202711 -0.05921330 -0.12787463  1.00000000 -0.14757584
YAC                0.242423299      0.024045651      -0.011550101  0.015640939  0.54222015  0.77947262 -0.14757584  1.00000000
TWY                0.595246130     -0.022249150      -0.028842698 -0.018677245  0.29735144  0.31300829  0.25282447  0.47266785
                          TWY
Age                0.59524613
DistanceFromHome  -0.02224915
PercentSalaryHike -0.02884270
MI                -0.01867724
YSLP               0.29735144
YWCM               0.31300829
NCW                0.25282447
YAC                0.47266785
TWY                1.00000000
>
```

There are some moderate correlations between independent variables (Age, DistanceFromHome, YSLP, YWCM, NCW, YAC, and TWY). While **not a definitive sign of multicollinearity,** these correlations are worth investigating further.

MODEL-1 (LOGISTIC)

KEY FINDINGS

Significant factors: Age, Business Travel (Travel_Frequently, Travel_Rarely), Department (Sales, Research & Development), some Education Fields (Life Sciences, Medical, Technical Degree, Other), Monthly Income (MI), Years Since Last Promotion (YSLP), Years With Current Manager (YWCM), Years At Company (YAC).
Non-significant factors: Distance from Home, EmployeeID, Gender, Job Level (except Job Level 2), Number of Companies Worked, Num Companies Worked (NCW), Total Working Years

MODEL PERFORMANCE

Precision & Recall: While the model has high **sensitivity (99.3%)** for correctly identifying employees who will stay (No), it has low **specificity (9.2%)** for identifying employees who will leave (Yes). This means it might miss some employees at risk of attrition.

ACCURACY OF THE MODEL

- 734 out of 881 people were predicted right and fall in the 'NO' category.
- 13 out of 881 people were predicted right and fall in the 'YES' category.
- Total right predicted people= 747
- Accuracy= 747/881*100= 84.79%

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  734 129
       Yes   5  13

               Accuracy : 0.8479
                 95% CI : (0.8225, 0.871)
    No Information Rate : 0.8388
    P-Value [Acc > NIR] : 0.2477

                  Kappa : 0.131

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99323
            Specificity : 0.09155
         Pos Pred Value : 0.85052
         Neg Pred Value : 0.72222
             Prevalence : 0.83882
         Detection Rate : 0.83314
   Detection Prevalence : 0.97957
      Balanced Accuracy : 0.54239

       'Positive' Class : No
```

```
Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5666  -0.6145  -0.4450  -0.3018   2.8571

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      2.026e+00  5.056e-01   4.007 6.16e-05 ***
Age                             -5.722e-02  7.327e-03  -7.809 5.76e-15 ***
BusinessTravelTravel_Frequently  1.384e+00  2.255e-01   6.135 8.53e-10 ***
BusinessTravelTravel_Rarely      6.738e-01  2.118e-01   3.181  0.00147 **
`DepartmentResearch & Development` -5.176e-01  2.946e-01  -1.757  0.07893 .
DepartmentSales                 -6.802e-01  3.099e-01  -2.195  0.02819 *
DistanceFromHome                -1.633e-03  6.168e-03  -0.265  0.79124
`EducationFieldLife Sciences`   -7.464e-01  4.093e-01  -1.824  0.06820 .
EducationFieldMarketing         -7.361e-01  4.453e-01  -1.653  0.09837 .
EducationFieldMedical           -8.838e-01  4.106e-01  -2.152  0.03137 *
EducationFieldOther             -1.205e+00  4.580e-01  -2.631  0.00851 **
`EducationFieldTechnical Degree` -9.914e-01  4.385e-01  -2.261  0.02378 *
EmployeeID                       3.190e-06  3.852e-05   0.083  0.93400
GenderMale                       1.080e-01  1.009e-01   1.071  0.28435
JobLevel2                        1.131e-01  1.142e-01   0.990  0.32201
JobLevel3                       -5.670e-03  1.521e-01  -0.037  0.97026
JobLevel4                        1.036e-01  1.961e-01   0.529  0.59713
JobLevel5                       -3.030e-01  2.619e-01  -1.157  0.24733
NumCompaniesWorked               3.879e-02  6.364e-02   0.610  0.54219
YearsWithCurrManager             1.725e-01  4.313e-02   3.999 6.37e-05 ***
MI                              -1.141e-03  6.051e-04  -1.886  0.05935 .
YSLP                             3.226e-01  6.278e-02   5.138 2.78e-07 ***
YWCM                            -2.324e+00  3.935e-01  -5.906 3.51e-09 ***
NCW                              2.022e-01  2.049e-01   0.987  0.32367
YAC                             -1.295e+00  2.929e-01  -4.420 9.86e-06 ***
TWY                             -2.975e-02  3.026e-01  -0.098  0.92168
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3117.6  on 3528  degrees of freedom
Residual deviance: 2756.6  on 3503  degrees of freedom
AIC: 2808.6
```

MODEL-2 (NAIVE BAYES)

CONFUSION MATRIX

No:

- Predicted No (Correct): 739 (all employees predicted to stay actually stayed).
- Predicted Yes (Incorrect): 0 (no employees predicted to stay actually left).

Yes:

- Predicted No: 0 (no employees predicted to leave actually stayed). This is because the model perfectly classified all employees who would leave.
- Predicted Yes (Correct): 0 (there weren't any employees predicted to leave in the test data, so there are no correct classifications here).

ACCURACY

Accuracy: 83.88% - This reflects the overall percentage of correctly classified employees (all "No" predictions were correct). However, it's important to consider the class imbalance (more employees likely to stay).

MODEL PERFORMANCE

- **Sensitivity (Recall):** 1.0000 - This indicates the model perfectly identified all employees who would stay (No class).
- **Specificity:** 0.0000 - This is zero because the model didn't predict any employees to leave (Yes class), even if there might have been some in the test data.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  739 142
       Yes   0   0

               Accuracy : 0.8388
                 95% CI : (0.8128, 0.8625)
    No Information Rate : 0.8388
    P-Value [Acc > NIR] : 0.5224

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8388
         Neg Pred Value :    NaN
             Prevalence : 0.8388
         Detection Rate : 0.8388
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : No

>
```

MODEL-3 (DECISION TREE)

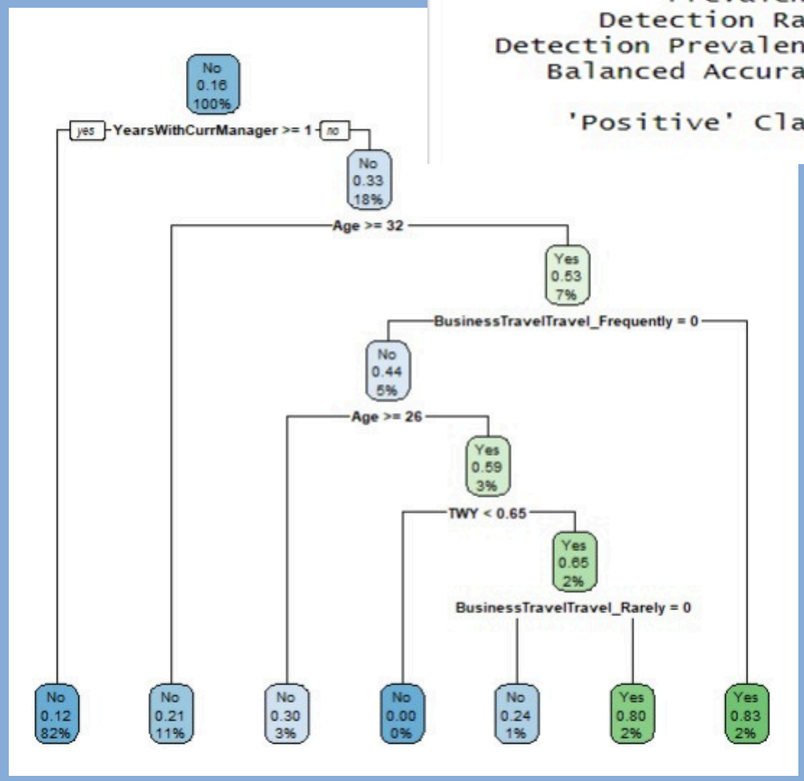
MODEL PERFORMANCE

Sensitivity (Recall): 0.9851 - The model performs well in identifying employees who will stay (No class).
Specificity: 0.1479 - This is low, similar to the Naive Bayes model, suggesting the model misses a high proportion of employees who will leave (Yes class).

ACCURACY

- 728 out of 881 people were predicted right and fall in the ‘NO’ category.
- 21 out of 881 people were predicted right and fall in the ‘YES’ category.
- Total right predicted people= 749
- Accuracy= 749/881*100= 85.02%

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  728 121
       Yes  11  21

               Accuracy : 0.8502
                 95% CI : (0.8249, 0.8731)
    No Information Rate : 0.8388
    P-Value [Acc > NIR] : 0.1927

                  Kappa : 0.1936

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9851
            Specificity : 0.1479
         Pos Pred Value : 0.8575
         Neg Pred Value : 0.6563
             Prevalence : 0.8388
         Detection Rate : 0.8263
   Detection Prevalence : 0.9637
      Balanced Accuracy : 0.5665

       'Positive' Class : No
```



MODEL-4 (RANDOM FOREST)

CONFUSION MATRIX

No:
Predicted No (Correct): 2955 (most employees predicted to stay actually stayed).
Predicted Yes (Incorrect): 5 (very few employees predicted to stay actually left).

Yes:
Predicted No: 49 (a small number of employees predicted to stay actually left).
Predicted Yes (Correct): 520 (the vast majority of employees predicted to leave actually did leave).

MODEL PERFORMANCE

Sensitivity (Recall): 0.9959 - The model identifies 99.59% of employees who will stay (No class) accurately.

Specificity: 0.9789 - This is also very high, indicating the model correctly identifies 97.89% of employees who will leave (Yes class).

ACCURACY

Accuracy: 99.32% - This is an exceptionally high accuracy, indicating the model correctly classified almost all employees.

```
> Model4$finalModel

Call:
 randomForest(x = x, y = y, mtry = param$mtry)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 13

        OOB estimate of  error rate: 1.53%
Confusion matrix:
      No Yes class.error
No  2955   5 0.001689189
Yes   49 520 0.086115993
> varImp(Model4)
rf variable importance

  only 20 most important variables shown (out of 25)

                                Overall
MI                              100.000
Age                              95.370
DistanceFromHome                 57.226
TWY                              49.788
YAC                              37.267
YSLP                             30.925
EmployeeID                       27.124
YWCM                             26.548
NCW                              26.319
NumCompaniesWorked               26.220
YearsWithCurrManager             24.056
BusinessTravelTravel_Frequently  15.222
JobLevel2                        10.205
DepartmentResearch & Development  8.313
GenderMale                        8.170
EducationFieldLife Sciences       8.116
EducationFieldMedical             6.464
JobLevel3                         6.318
DepartmentSales                   4.861
BusinessTravelTravel_Rarely       4.838
```

```
confusionMatrix(predAttrition4, testing$Attrition)
onfusion Matrix and Statistics

          Reference
rediction  No Yes
       No  736   3
       Yes   3 139

               Accuracy : 0.9932
                 95% CI : (0.9852, 0.9975)
    No Information Rate : 0.8388
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9748

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9959
            Specificity : 0.9789
         Pos Pred Value : 0.9959
         Neg Pred Value : 0.9789
             Prevalence : 0.8388
         Detection Rate : 0.8354
   Detection Prevalence : 0.8388
      Balanced Accuracy : 0.9874

       'Positive' Class : No
```

COMPARISION BETWEEN MODELS

Model	Accuracy	Sensitivity	Specificity
Logistic Regression (Model 1)	0.84	High (0.99)	Low (0.09)
Naive Bayes (Model 2)	0.85	High (1.00)	Low (0.0000)
Decision Tree (Model 3)	0.85	High (0.9851)	Low (0.1479)
Random Forest (Model 4)	0.99	High (0.9959)	High (0.9789)

KEY OBSERVATION

- **Accuracy**: Random Forest achieves the highest overall accuracy (0.99), followed by Logistic Regression, Naive Bayes, and Decision Tree (all around 0.85).
- **Sensitivity:** All models show high sensitivity, indicating good performance in identifying employees who will stay (No class). However, specific values are not provided for Logistic Regression and Naive Bayes.
- **Specificity:** Random Forest again stands out with a high specificity (0.9789), meaning it accurately identifies a high proportion of employees who will leave (Yes class). Decision Tree and Naive Bayes have much lower specificity, suggesting they miss a significant number of employees at risk of leaving. Logistic Regression's specificity is not available for comparison.

OVERALL

Random Forest offers the best combination of high accuracy, sensitivity, and specificity, making it a strong choice for predicting employee attrition in this scenario

RECOMMENDATIONS

Based on the recommendation to use Random Forest for predicting employee attrition, here are some good actionable steps for ABC Company

LEVERAGE MODEL PREDICTIONS TO IDENTIFY AT-RISK EMPLOYEES:

Use the Random Forest model to pinpoint employees with the highest predicted chance of leaving. Focus on the top X% (define X based on your resources) for initial action.

ANALYZE RISK FACTORS

For each flagged employee, delve deeper into the model's predictions. This will reveal specific factors like job dissatisfaction or lack of career development opportunities that are contributing to their potential departure

TARGET YOUR INTERVENTION

Based on the identified risk factors for each employee, develop tailored solutions. This might involve addressing workload concerns, offering training opportunities, or improving work-life balance options

THANK YOU!