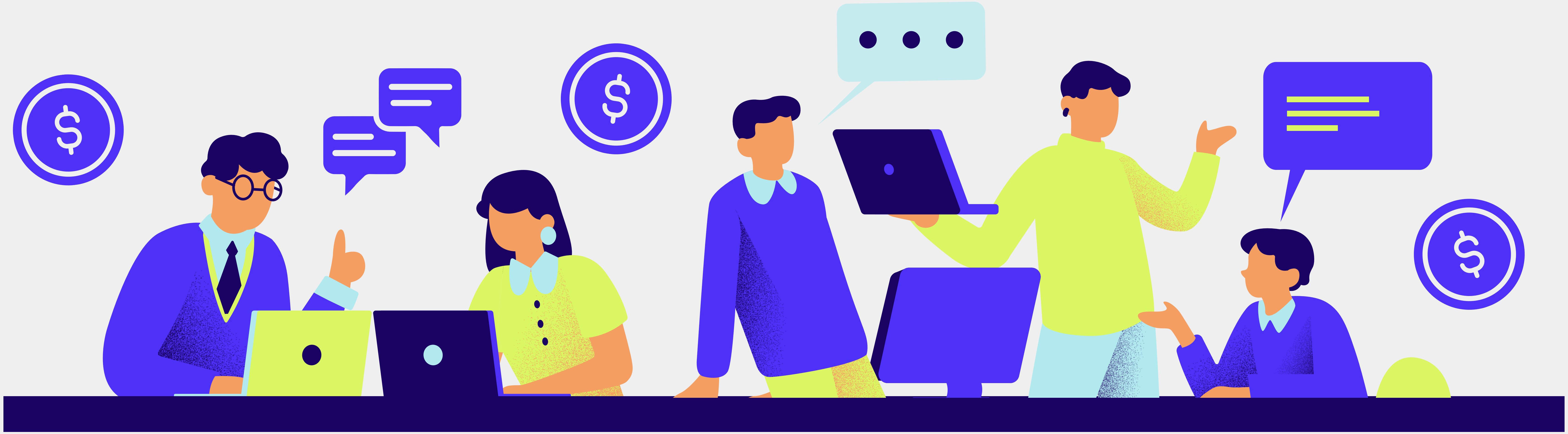
STATISTICAL LEARNING

Understanding Employee Attrition



EMILE RAHAL

Attrition refers to the voluntary departure of employees from their jobs within an organization.

What are the main factors contributing to employee attrition?

How can organizations keep employees in their jobs while enhancing the overall work environment?

WHAT'S IN THE DATASET?

Satisfaction Level	Numerical
Last Evaluation Rating	Numerical
Projects Worked On	Numerical
Average Monthly Hours	Numerical
Work Accident	Binary
Salary Level	Categorical
Department	Categorical
Time Spent with Comp.	Numerical
Promo in the Last 5Y	Binary
Attrition	Target /Binary

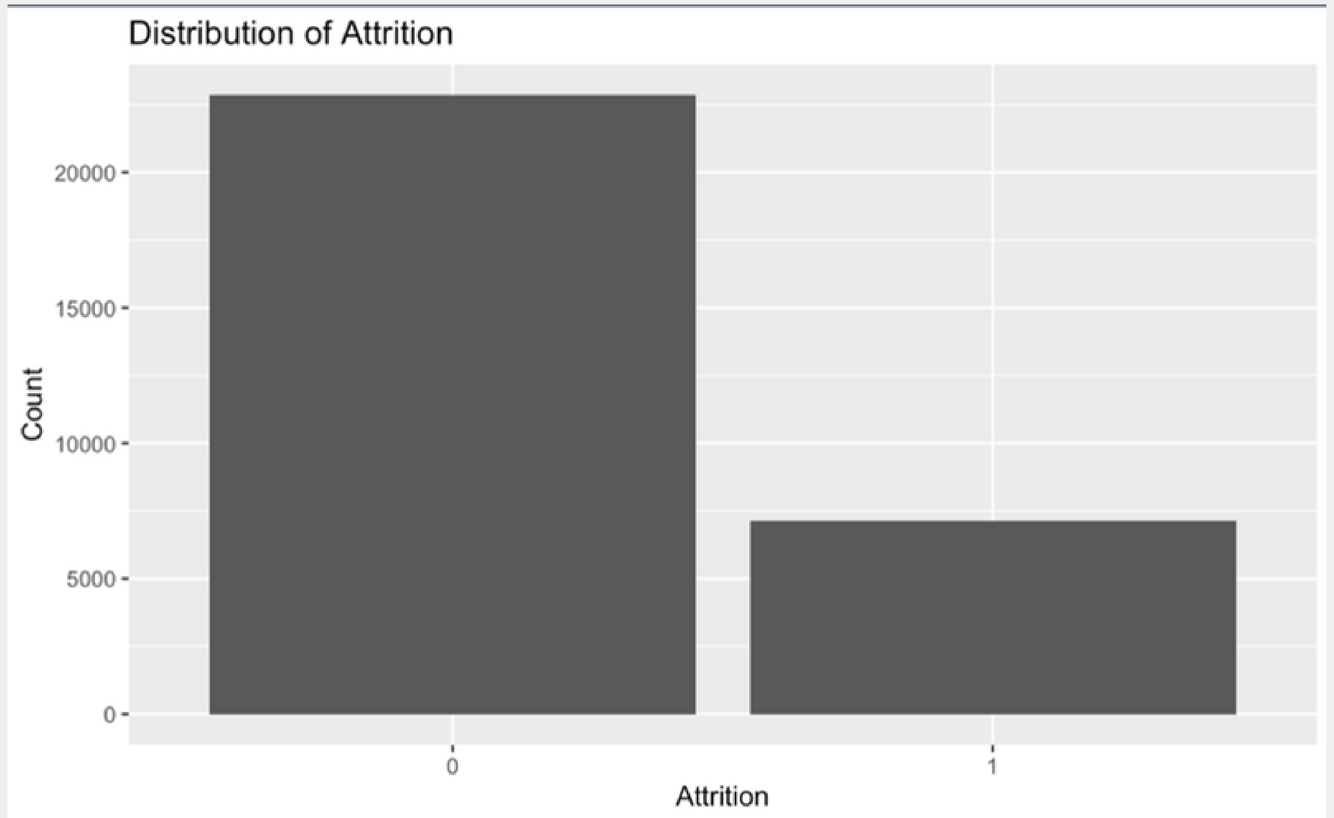
DATA PREPROCESSING

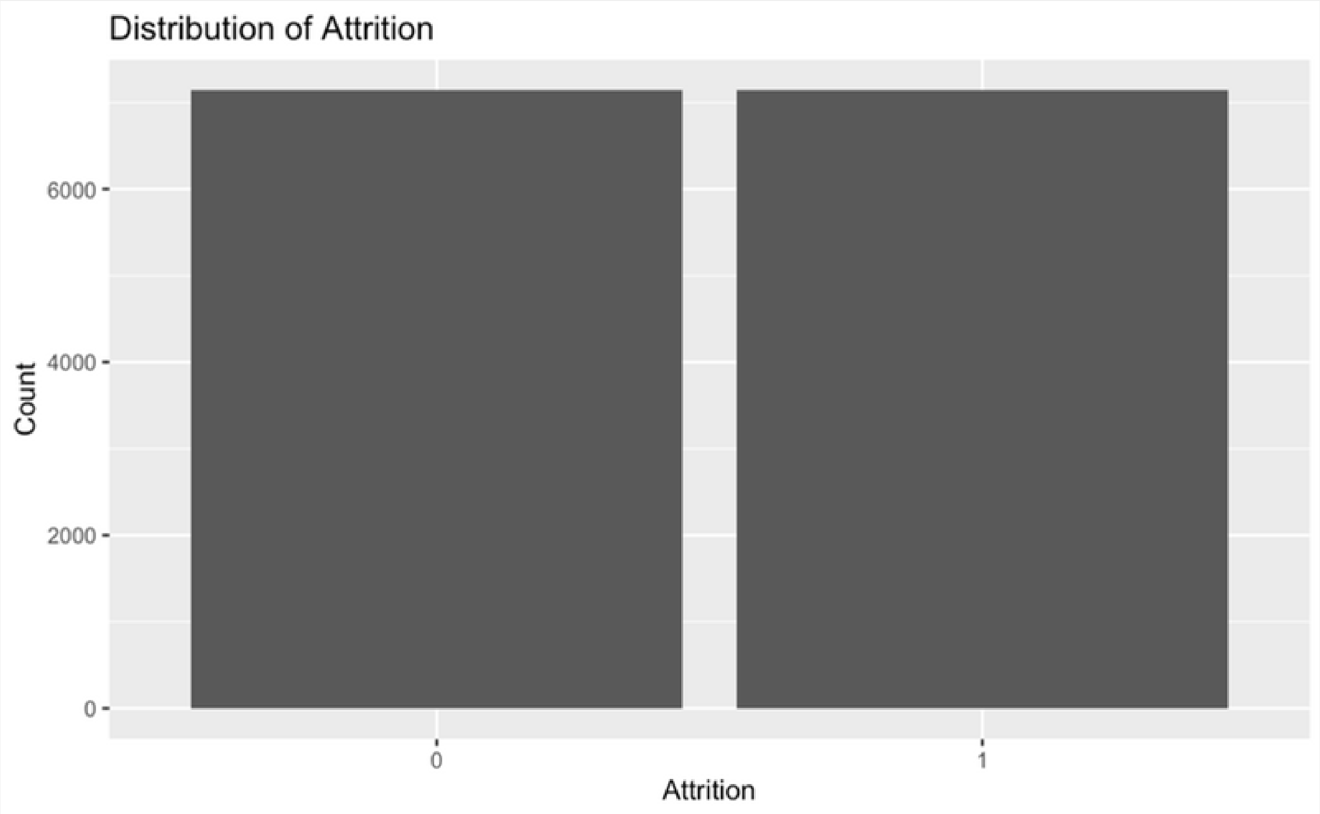
- **MERGED DATASETS:** COMBINED **TRAINING** AND **TEST** SETS TO EXPAND THE **DATASET**.
- **NO MISSING VALUES:** NO **MISSING** DATA, SO NO **IMPUTATION** IS NEEDED.

- **CLASS IMBALANCE:**
  MORE "NO ATTRITION" (**CLASS 0**) THAN "ATTRITION" (CLASS 1).
  IMBALANCE CAN AFFECT MODEL **ACCURACY**.

- **DOWNSAMPLING:**
  REDUCED CLASS 0 TO **MATCH** CLASS 1 SIZE.
  BALANCING CLASSES PREVENTS MODEL **BIAS**.

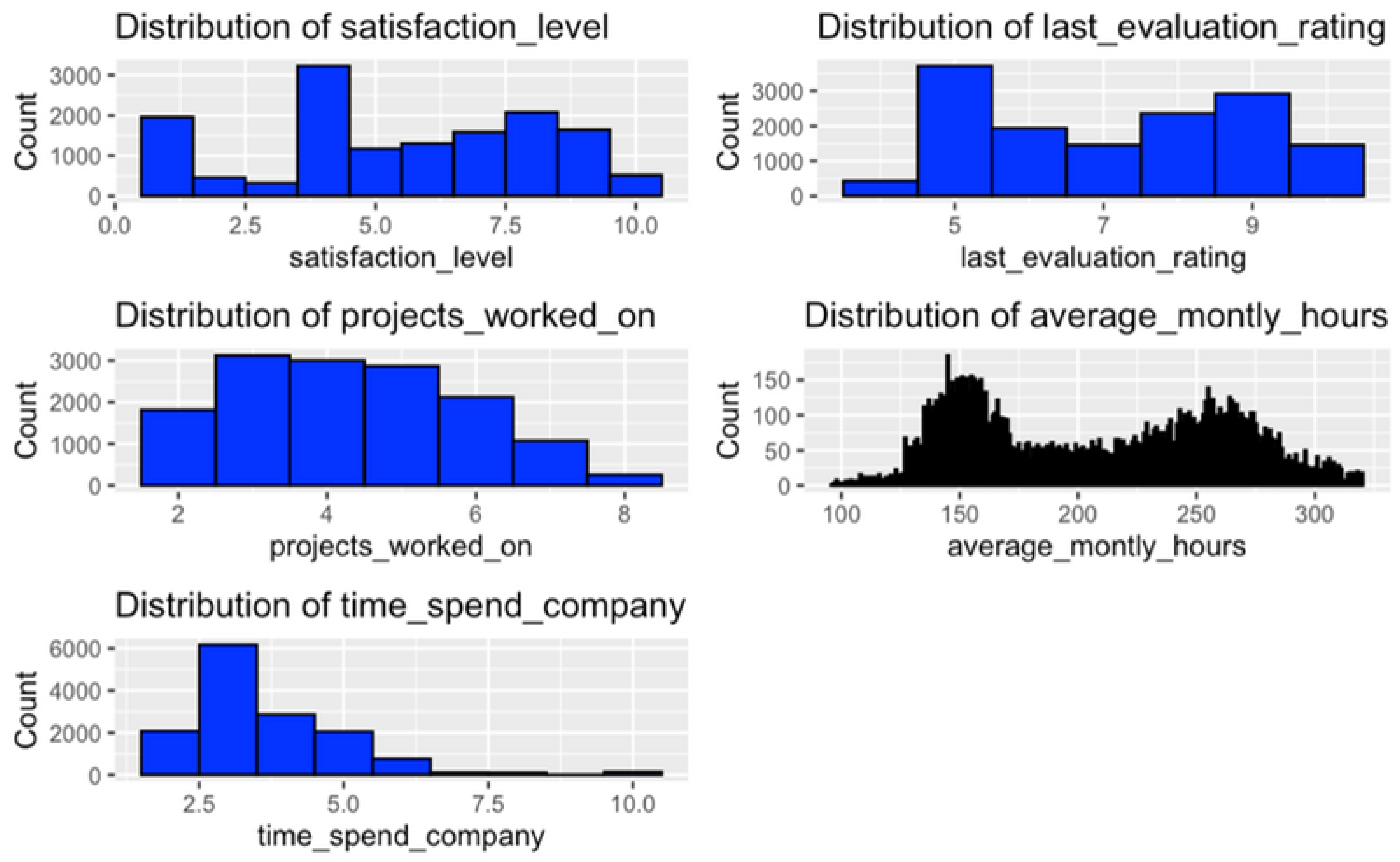
BEFORE





AFTER

EXPLANATORY DATA ANALYSIS



- **SATISFACTION LEVEL:** ~5.52/10 (MODERATE)

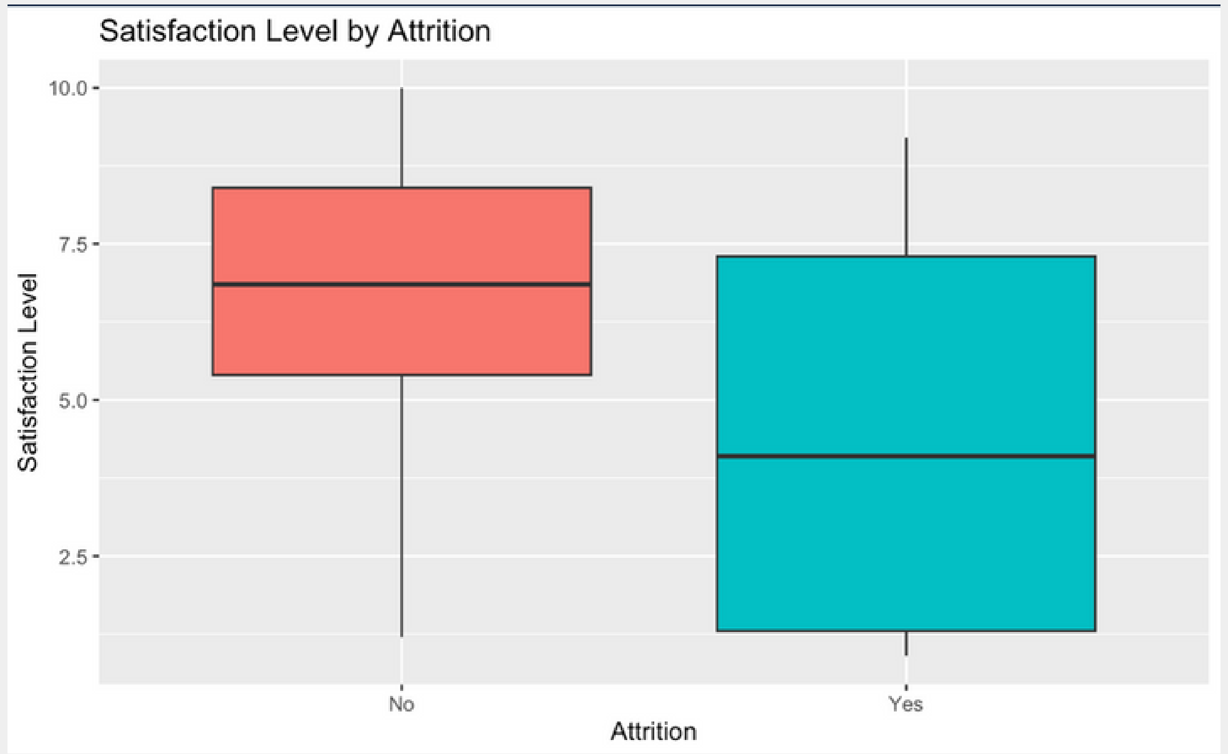
- **LAST EVALUATION RATING:** ~7.17/10 (RELATIVELY HIGH)

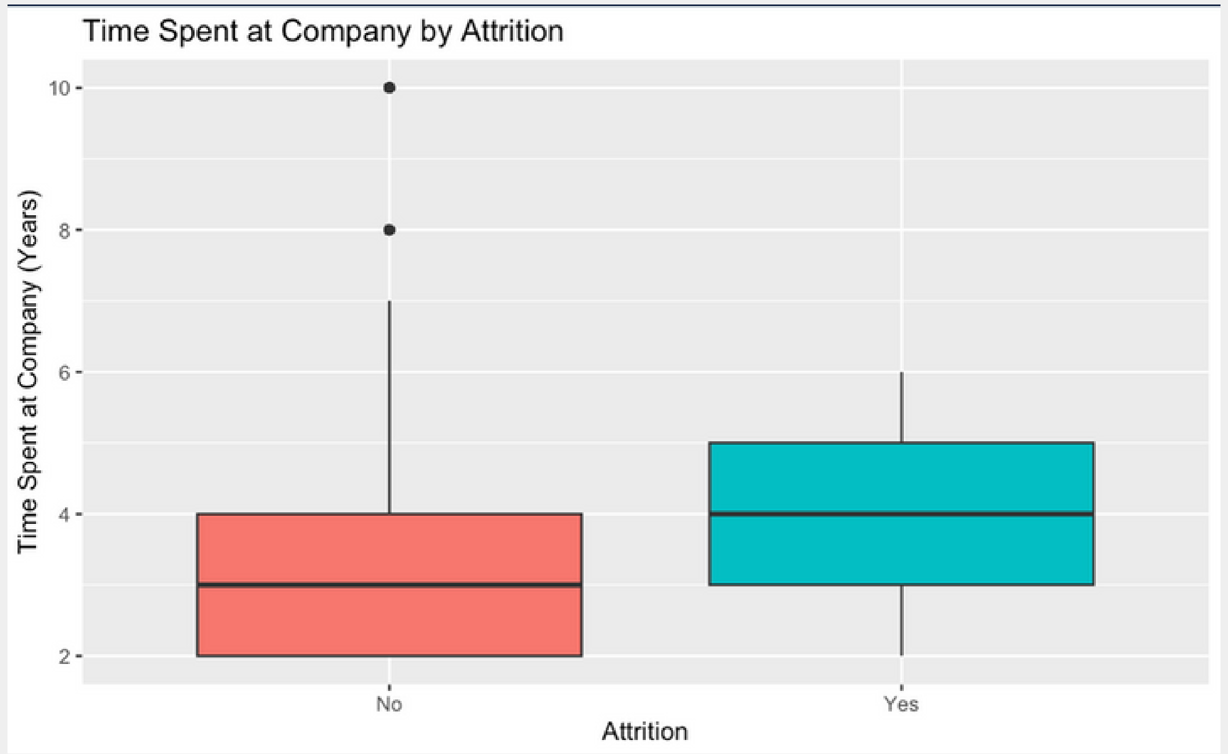
- **PROJECTS WORKED ON:** ~4.32 (MULTIPLE PROJECTS)

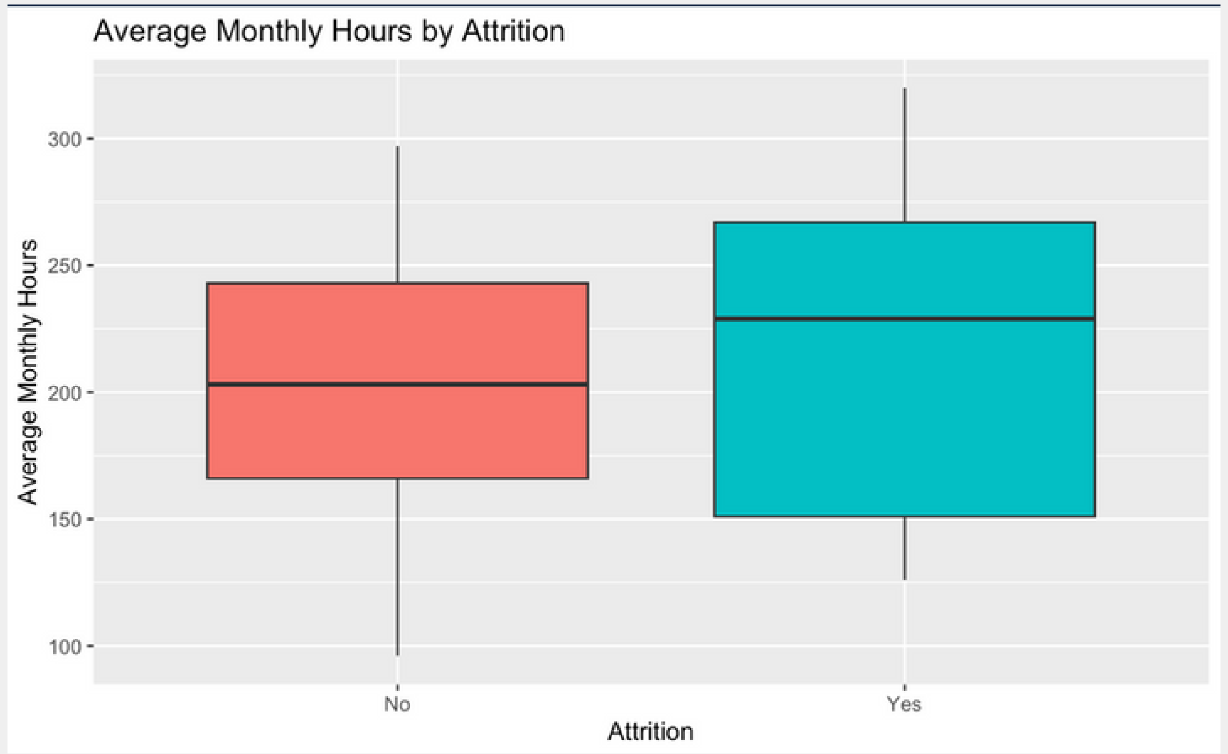
- **MONTHLY HOURS:** ~208.08 HOURS (MODERATE WORKLOAD)

- **TIME WITH COMPANY:** ~3.64 YEARS (SHORT TENURE)

NUMERICAL VARIABLES


Satisfaction Level by Attrition


Time Spent at Company by Attrition


Average Monthly Hours by Attrition

SATISFACTION LEVEL

EMPLOYEES WHO **LEFT** HAD SIGNIFICANTLY **LOWER** SATISFACTION LEVELS

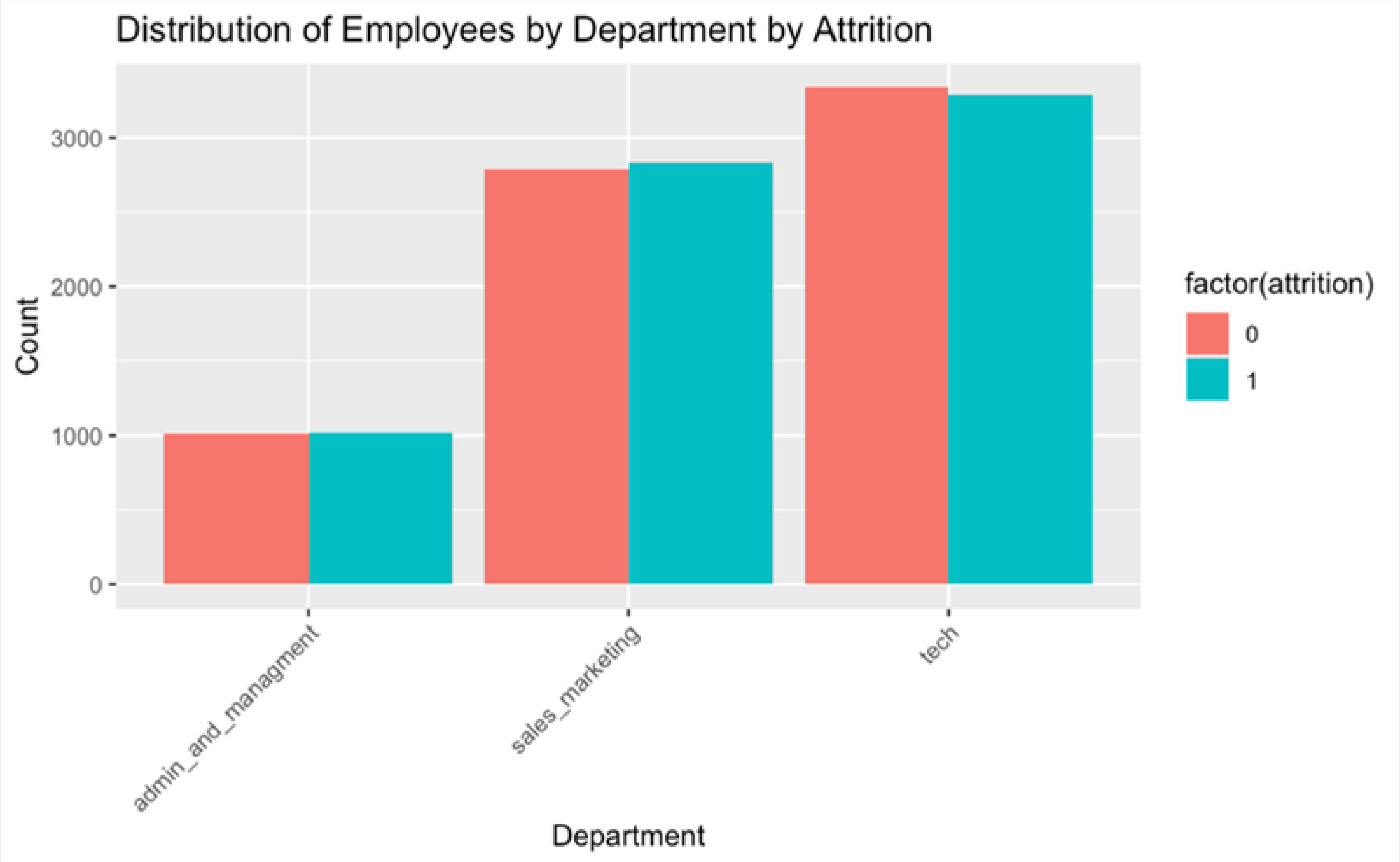
TIME WITH COMPANY

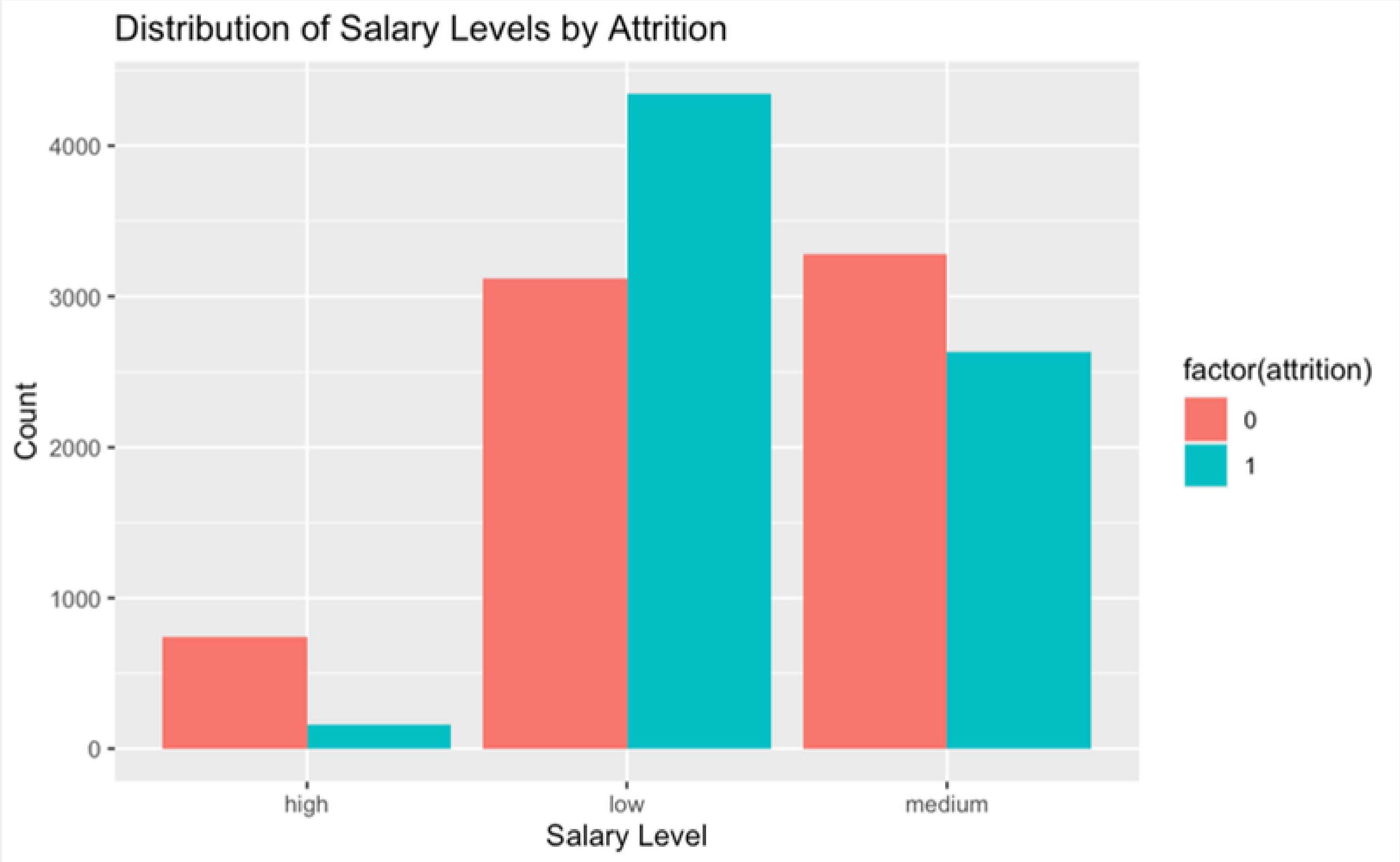
EMPLOYEES WHO **LEFT** HAD SLIGHTLY **MORE** TENURE.

MONTHLY HOURS

EMPLOYEES WHO **LEFT** WORKED **MORE** HOURS ON AVERAGE

CATEGORICAL VARIABLES





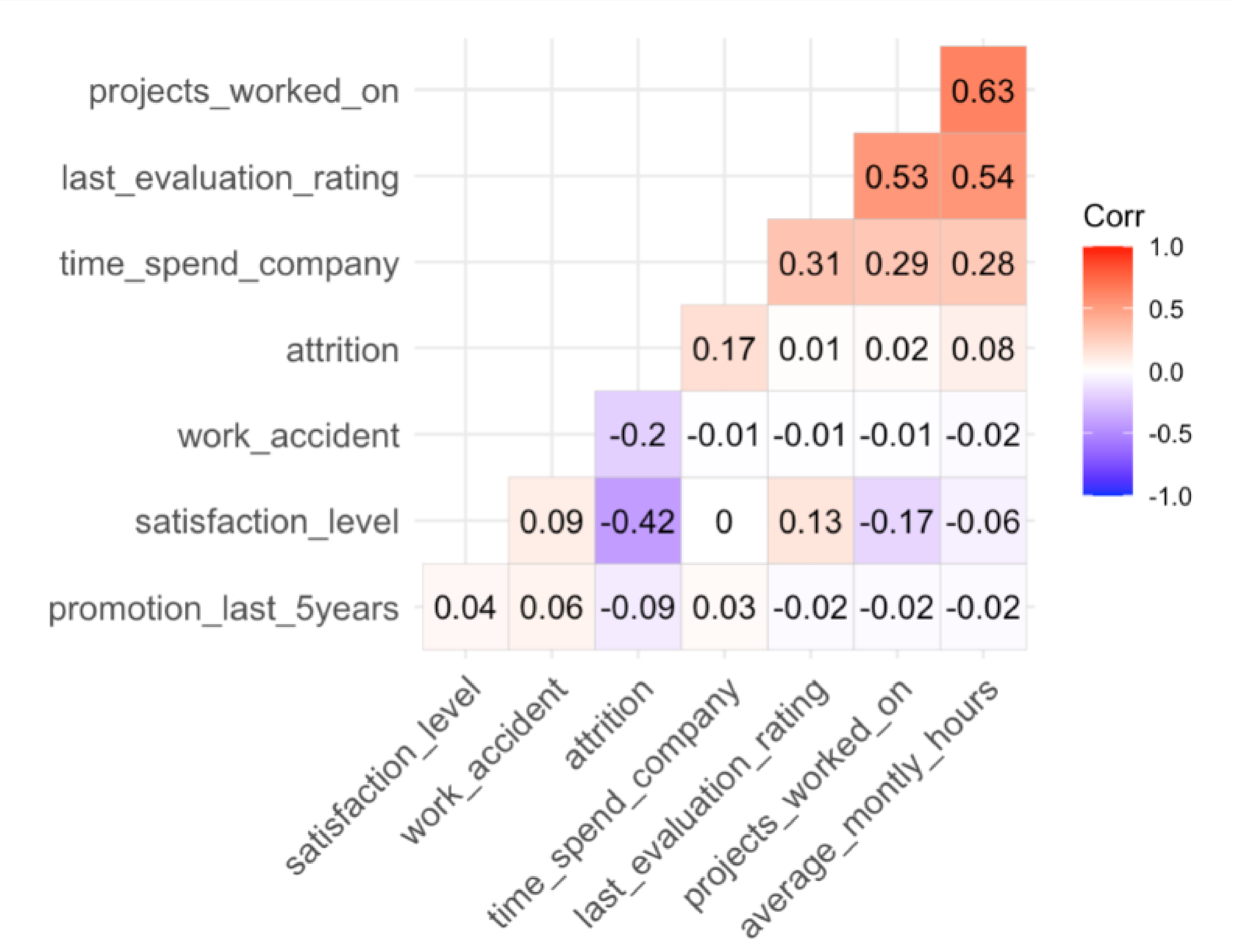
DEPARTMENT

DEPARTMENTS WERE **MAPPED** INTO **3 BROADER** CATEGORIES THAT ARE **UNIFORM** IN TERMS OF ATTRITION

SALARY LEVEL

HIGHER ATTRITION FOR **LOW** AND **MEDIUM** SALARIES COMPARED TO HIGH SALARIES.

CORRELATION ANALYSIS



- **KEY FACTOR:** "SATISFACTION LEVEL" STRONGLY **INFLUENCES** ATTRITION; **LOWER** SATISFACTION CORRELATES WITH **HIGHER** ATTRITION.
- **WEAKER ASSOCIATIONS:** "EVALUATION RATINGS," "PROJECTS WORKED ON," AND "AVERAGE MONTHLY HOURS" SHOW **LESS** PRONOUNCED **LINKS** TO ATTRITION.
- **NOTE: CORRELATION** DOES NOT IMPLY **CAUSATION**.

SOME VARIABLES EXHIBIT **STRONG** CORRELATIONS

VIF ANALYSIS FOR THE VARIABLES REVEALS **NO MULTICOLLINEARITY**

ALL VIF VALUES ARE NEAR 1

SUPERVISED LEARNING

DATA PREPROCESSING

CATEGORICAL ENCODING:

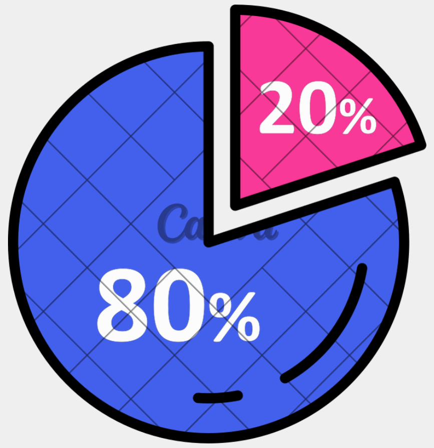
- **"SALARY": ORDINAL** ENCODING AS "LOW," "MEDIUM," AND "HIGH."
- **"DEPARTMENT": ONE-HOT** ENCODING WITH ONE DUMMY COLUMN REMOVED.

NUMERICAL SCALING:

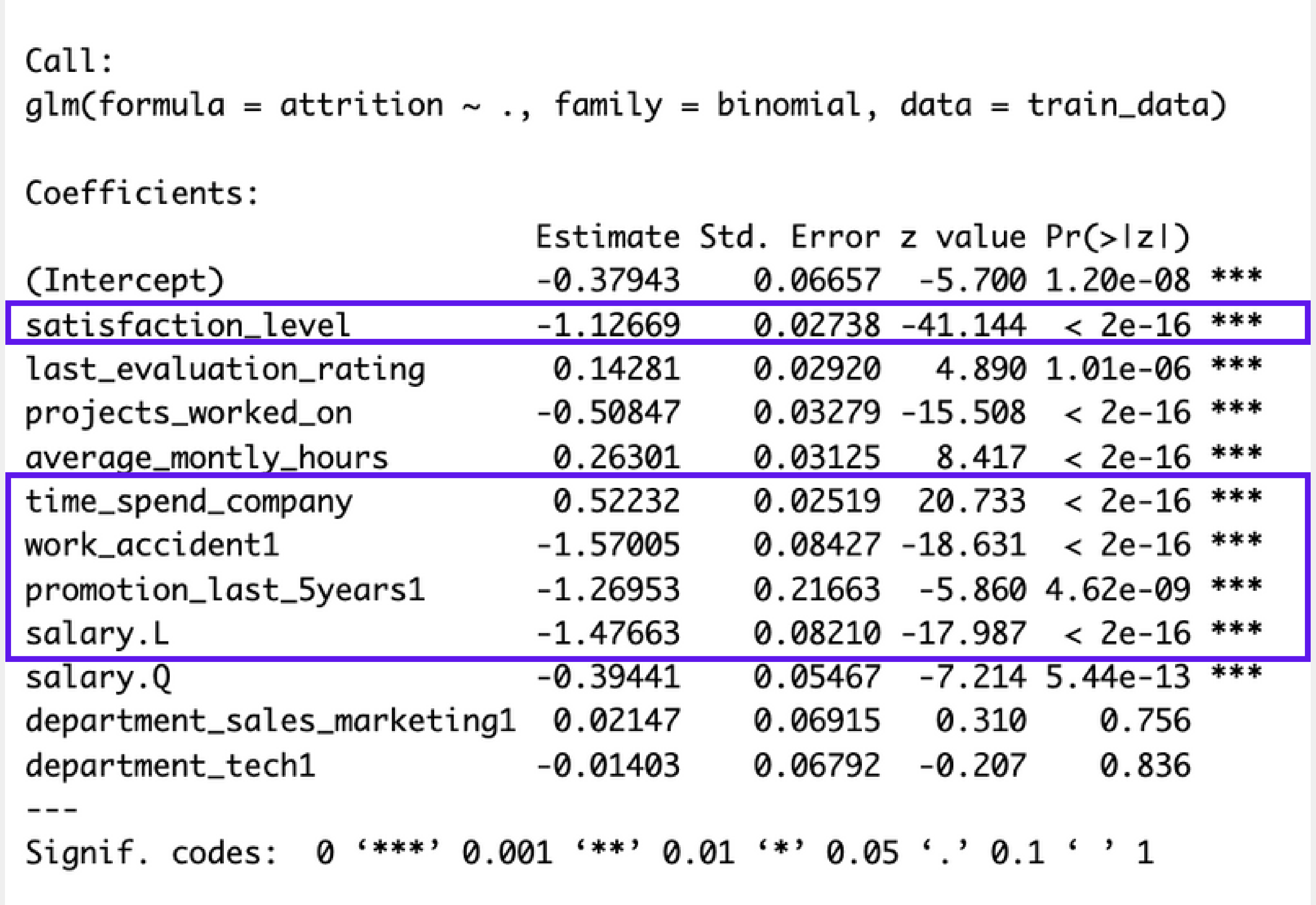
APPLIED **Z-SCORE** SCALING FOR NUMERICAL VARIABLES (MEAN 0, STD. 1)

DATA SPLITTING:

DATASET SPLIT INTO 80% **TRAINING** AND 20% **TESTING** SETS



LOGISTIC REGRESSION



```
Call:
glm(formula = attrition ~ ., family = binomial, data = train_data)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 -0.37943    0.06657  -5.700 1.20e-08 ***
satisfaction_level          -1.12669    0.02738 -41.144  < 2e-16 ***
last_evaluation_rating       0.14281    0.02920   4.890 1.01e-06 ***
projects_worked_on          -0.50847    0.03279 -15.508  < 2e-16 ***
average_montly_hours         0.26301    0.03125   8.417  < 2e-16 ***
time_spend_company           0.52232    0.02519  20.733  < 2e-16 ***
work_accident1              -1.57005    0.08427 -18.631  < 2e-16 ***
promotion_last_5years1      -1.26953    0.21663  -5.860 4.62e-09 ***
salary.L                    -1.47663    0.08210 -17.987  < 2e-16 ***
salary.Q                    -0.39441    0.05467  -7.214 5.44e-13 ***
department_sales_marketing1  0.02147    0.06915   0.310    0.756    
department_tech1            -0.01403    0.06792  -0.207    0.836    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **SATISFACTION LEVEL: DECREASE** LINKED TO **HIGHER** ATTRITION.
- **TIME SPEND IN COMPANY: MORE** TIME, **HIGHER** ATTRITION.
- **WORK ACCIDENT: DECREASED** LIKELIHOOD OF ATTRITION.
- **PROMOTION IN LAST 5 YEARS: DECREASED** LIKELIHOOD OF ATTRITION.
- **SALARY LEVEL: LOWER** SALARIES LINKED TO **HIGHER** ATTRITION.

RANDOM FOREST

	Attrition (1)	No Attrition (0)
Attrition (1)	5570	144
No Attrition (0)	49	5665
Class Error	0.025	0.009

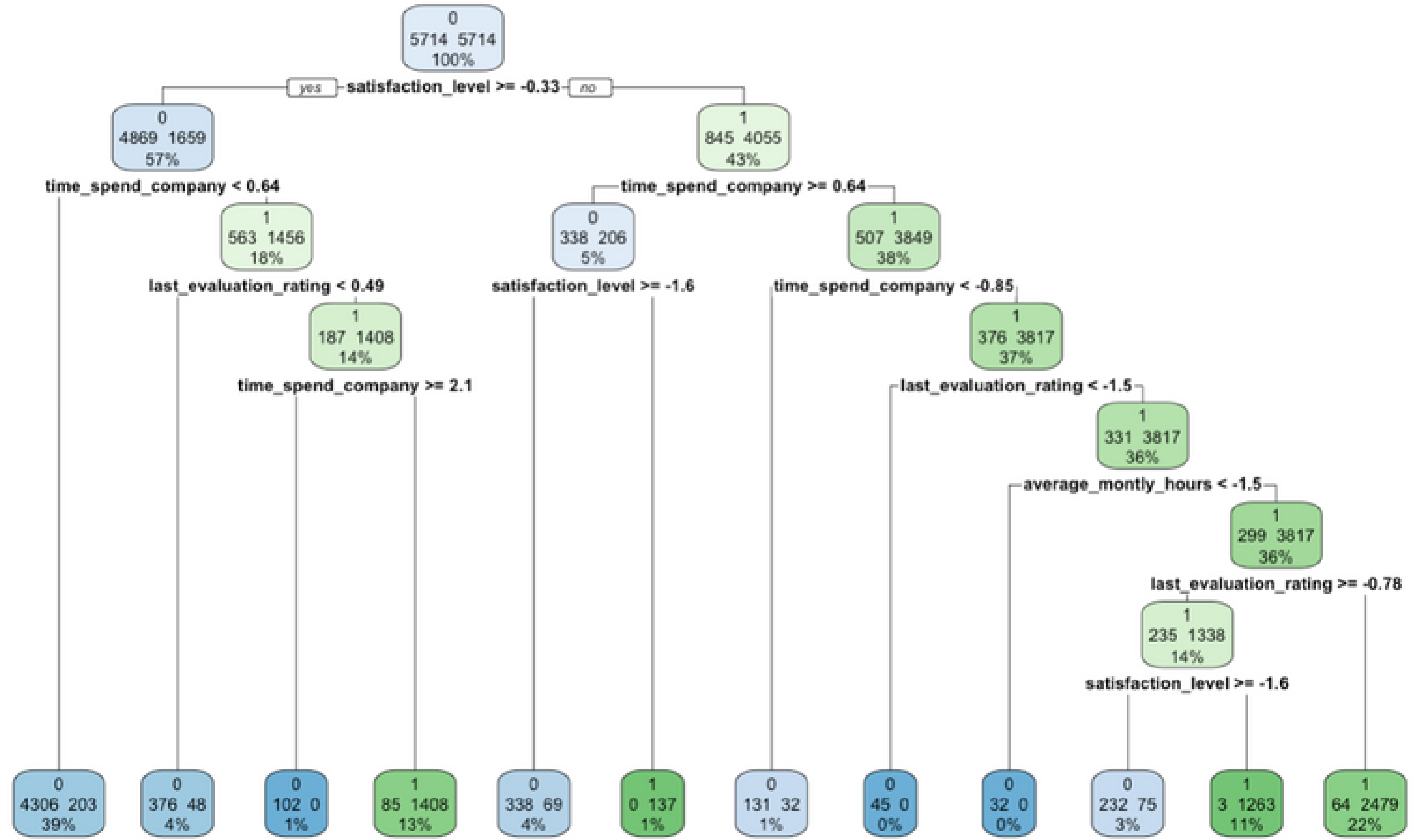
- USED TO PREDICT ATTRITION WITH **100 TREES**.
- A RANDOM SUBSET OF **THREE VARIABLES** AT EACH **SPLIT**.

PERFORMANCE:

- **OUT-OF-BAG (OOB) ERROR RATE:** ~1.69%, INDICATING **HIGH** ACCURACY.
- EFFECTIVE AT CLASSIFYING EMPLOYEES WHO LEAVE AND THOSE WHO STAY.

DECISION TREE

- PREDICTS ATTRITION THROUGH **RECURSIVE PARTITIONING**.
- EACH **BRANCH** REPRESENTS **CONDITIONS** LEADING TO ATTRITION OR NON-ATTRITION.

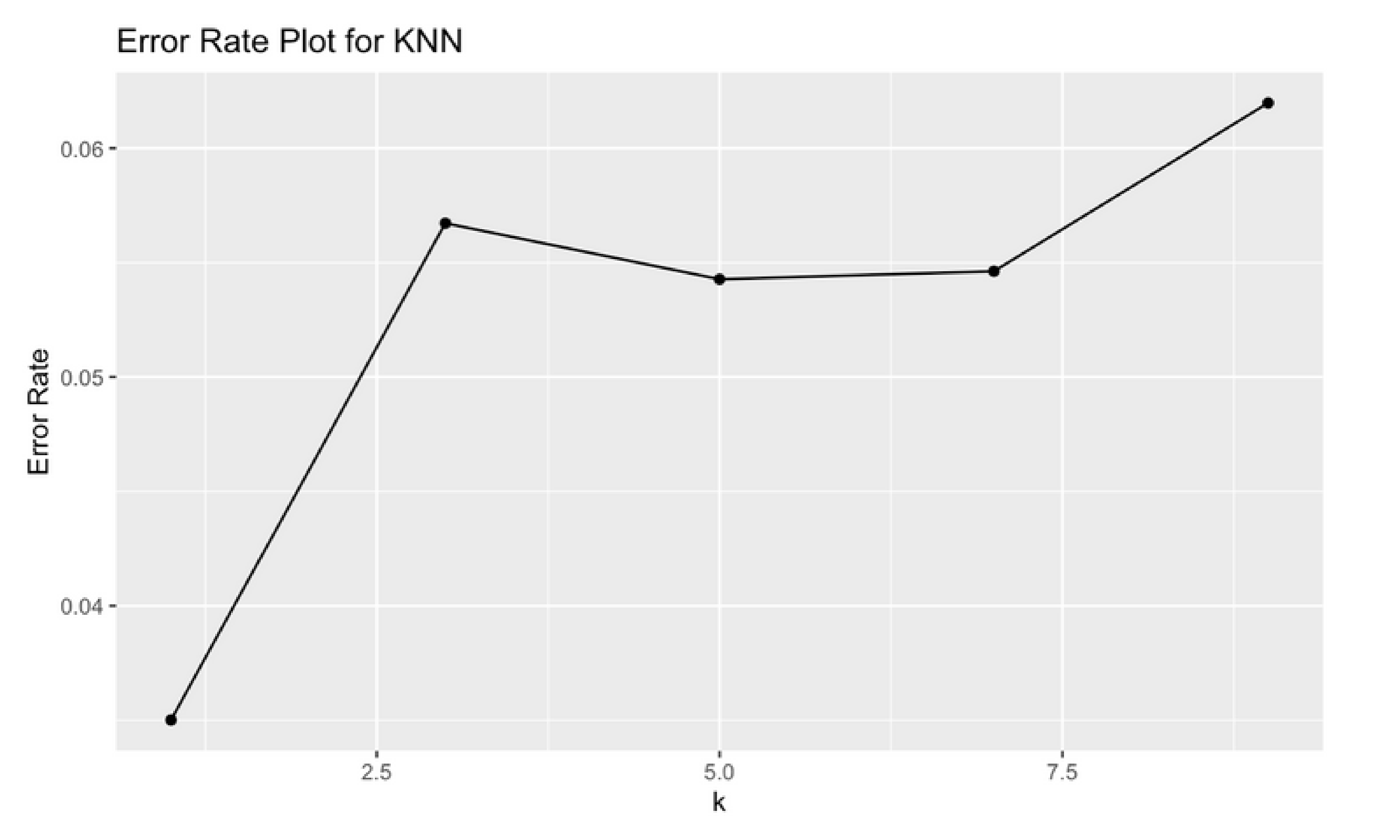


0
5714 5714
100%
yes
satisfaction_level >= -0.33
no
0
4869 1659
57%
1
845 4055
43%
time_spend_company < 0.64
time_spend_company >= 0.64
1
563 1456
18%
0
338 206
5%
1
507 3849
38%
last_evaluation_rating < 0.49
satisfaction_level >= -1.6
time_spend_company < -0.85
1
187 1408
14%
1
376 3817
37%
time_spend_company >= 2.1
last_evaluation_rating < -1.5
1
331 3817
36%
average_montly_hours < -1.5
1
299 3817
36%
last_evaluation_rating >= -0.78
1
235 1338
14%
satisfaction_level >= -1.6
0
4306 203
39%
0
376 48
4%
0
102 0
1%
1
85 1408
13%
0
338 69
4%
1
0 137
1%
0
131 32
1%
0
45 0
0%
0
32 0
0%
0
232 75
3%
1
3 1263
11%
1
64 2479
22%

KEY FACTORS IN THE TREE STRUCTURE:

- SATISFACTION LEVEL
- TIME SPENT IN COMPANY
- LAST EVALUATION RATING
- AVERAGE MONTHLY HOURS

K-NEAREST NEIGHBOR



OPTIMAL "K" VALUE:

- K=1: ERROR RATE ≈ 0.0350, SENSITIVE TO NOISE **(OVERFITTING)**
- **K=3, TO 9: ERROR RATE INCREASES WITH LARGER "K," LESS SENSITIVE TO INDIVIDUAL DATA POINTS.**
- **AFTER CAREFUL EVALUATION, "K=5" SELECTED.**

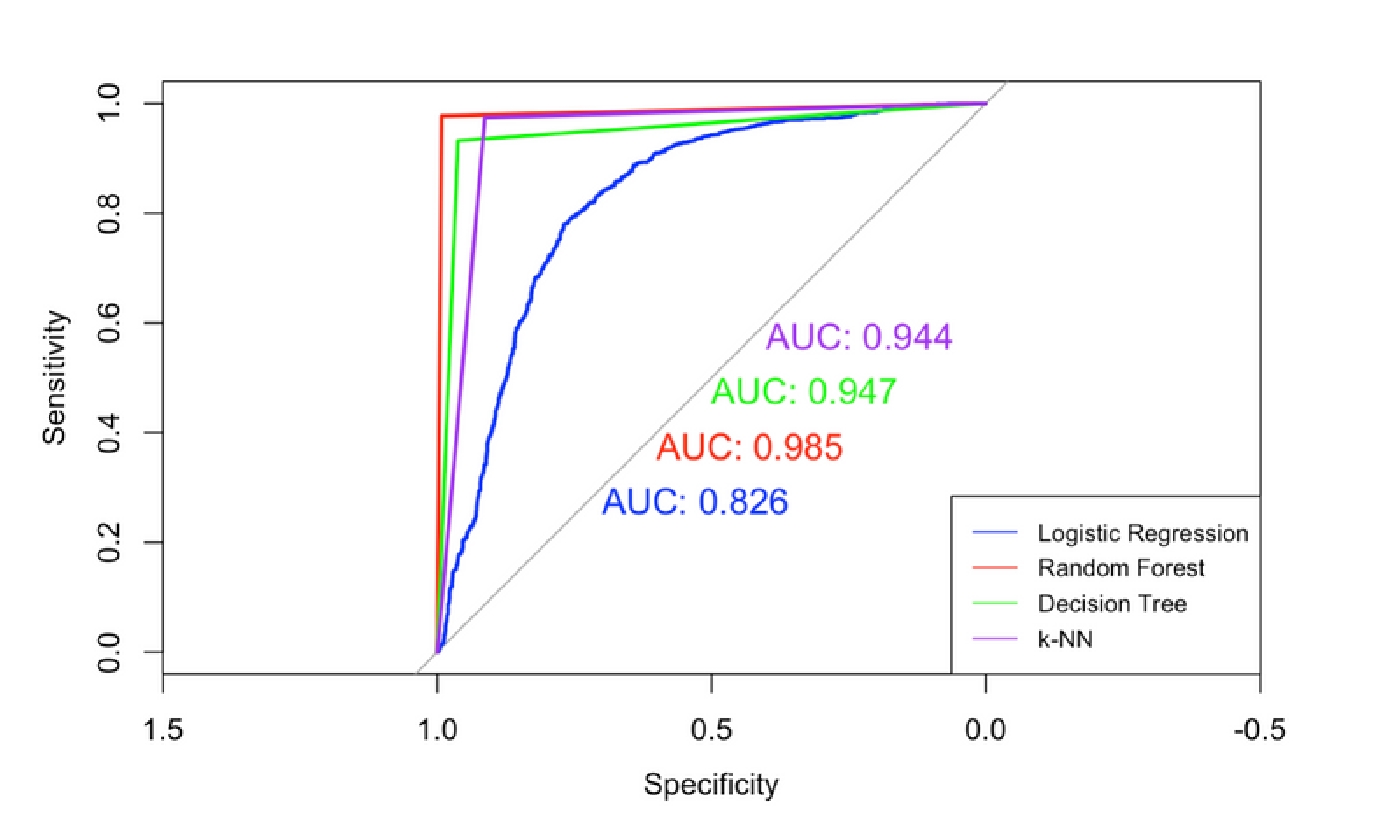
5-NN MODEL PERFORMS WELL, HIGH ACCURACY.

GENERALIZES EFFECTIVELY TO NEW DATA, GOOD FOR PREDICTING ATTRITION.

ROC-AUC

- **LOGISTIC REGRESSION:** MODERATE
- **RANDOM FOREST:** VERY HIGH
- **DECISION TREE:** STRONG
- **K-NN:** GOOD

RANDOM FOREST PERFORMED BEST IN DISTINGUISHING ATTRITION



PRECISION-RECALL

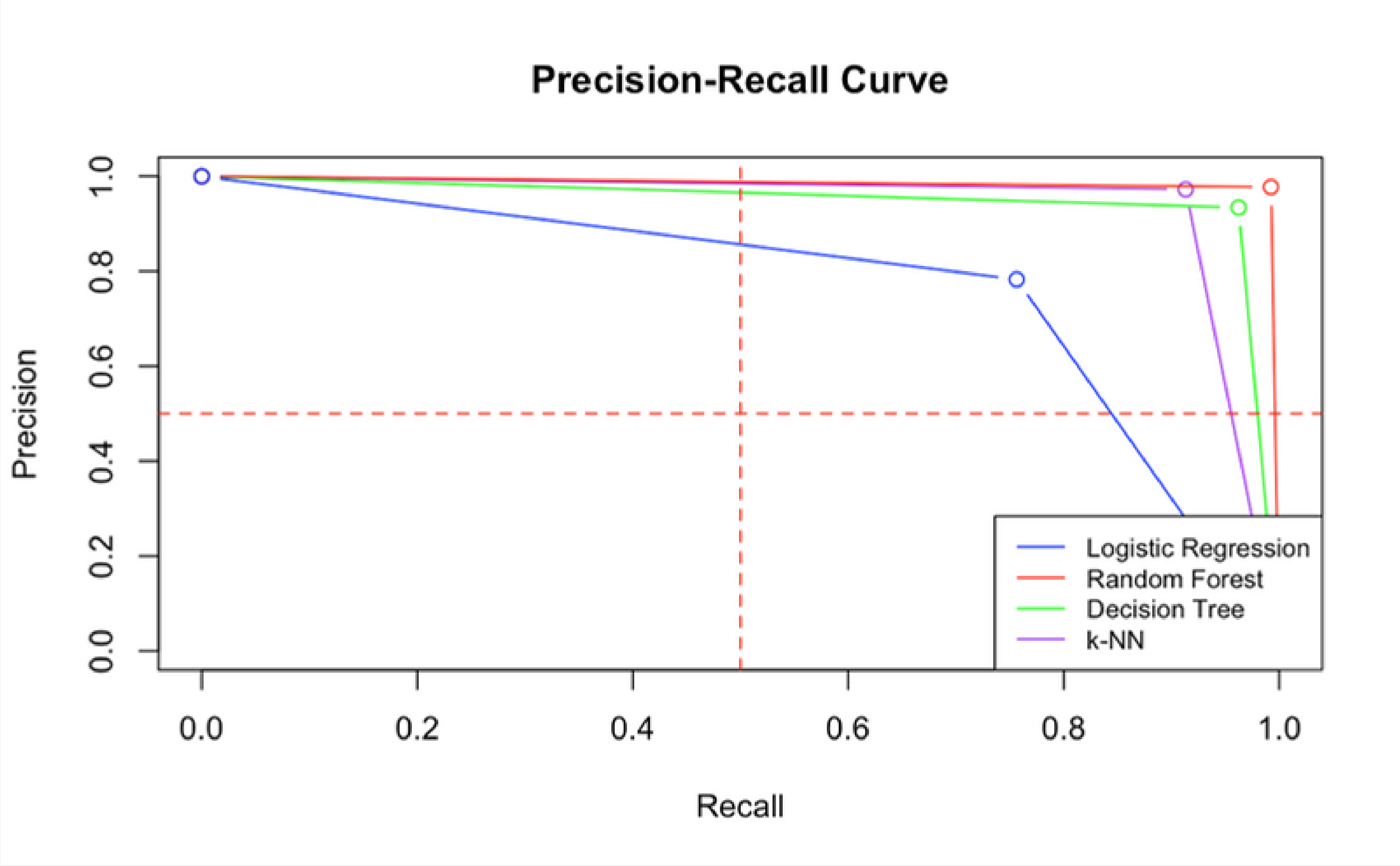
- **RANDOM FOREST:**
HIGH SCORES

- **DECISION TREE:**
GOOD BALANCE

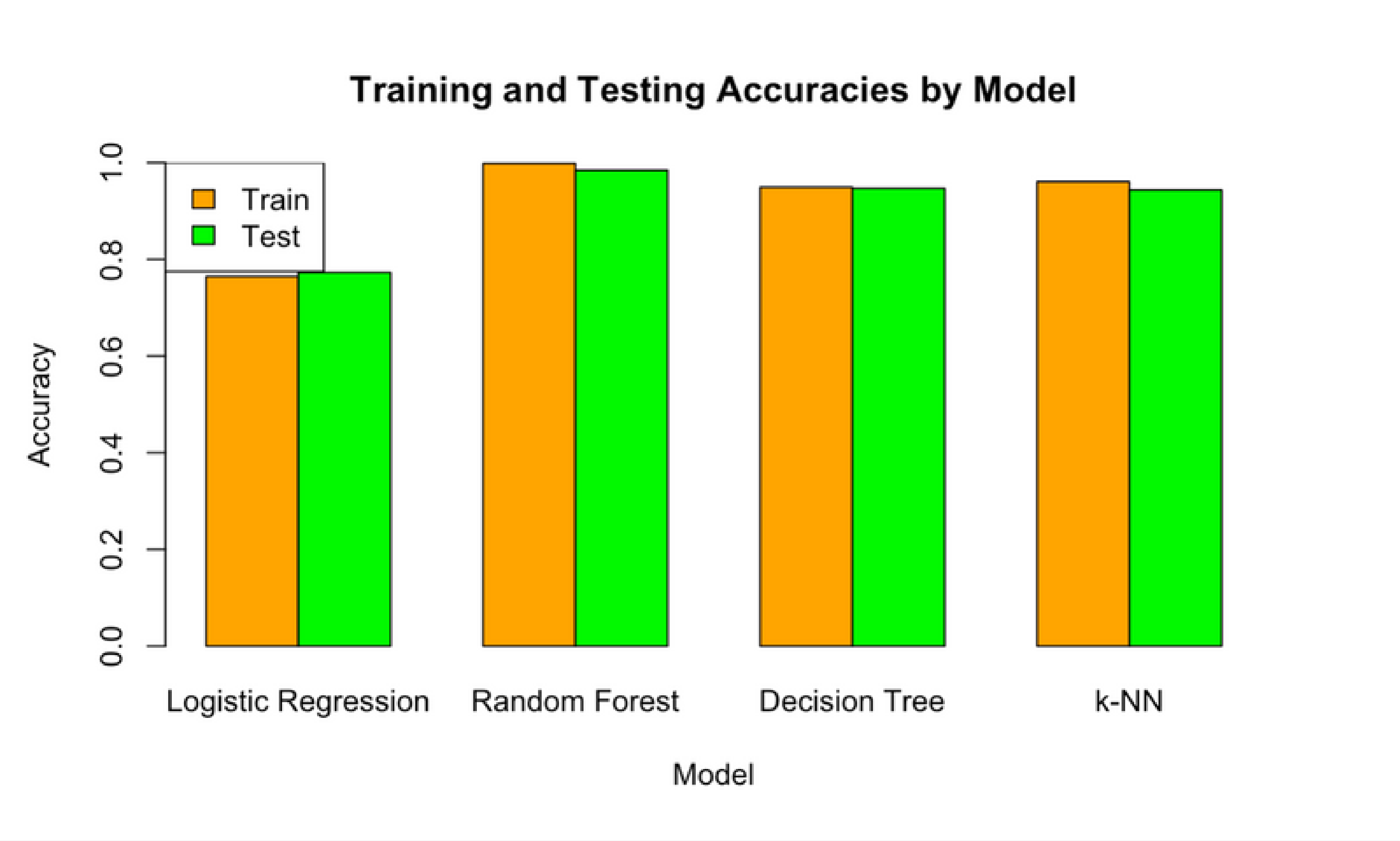
- **K-NN :**
BALANCED

- **LOGISTIC REGRESSION:**
RESPECTABLE

RANDOM FOREST SHOWS EXCEPTIONAL PRECISION AND RECALL, MAKING IT A ROBUST PREDICTOR.



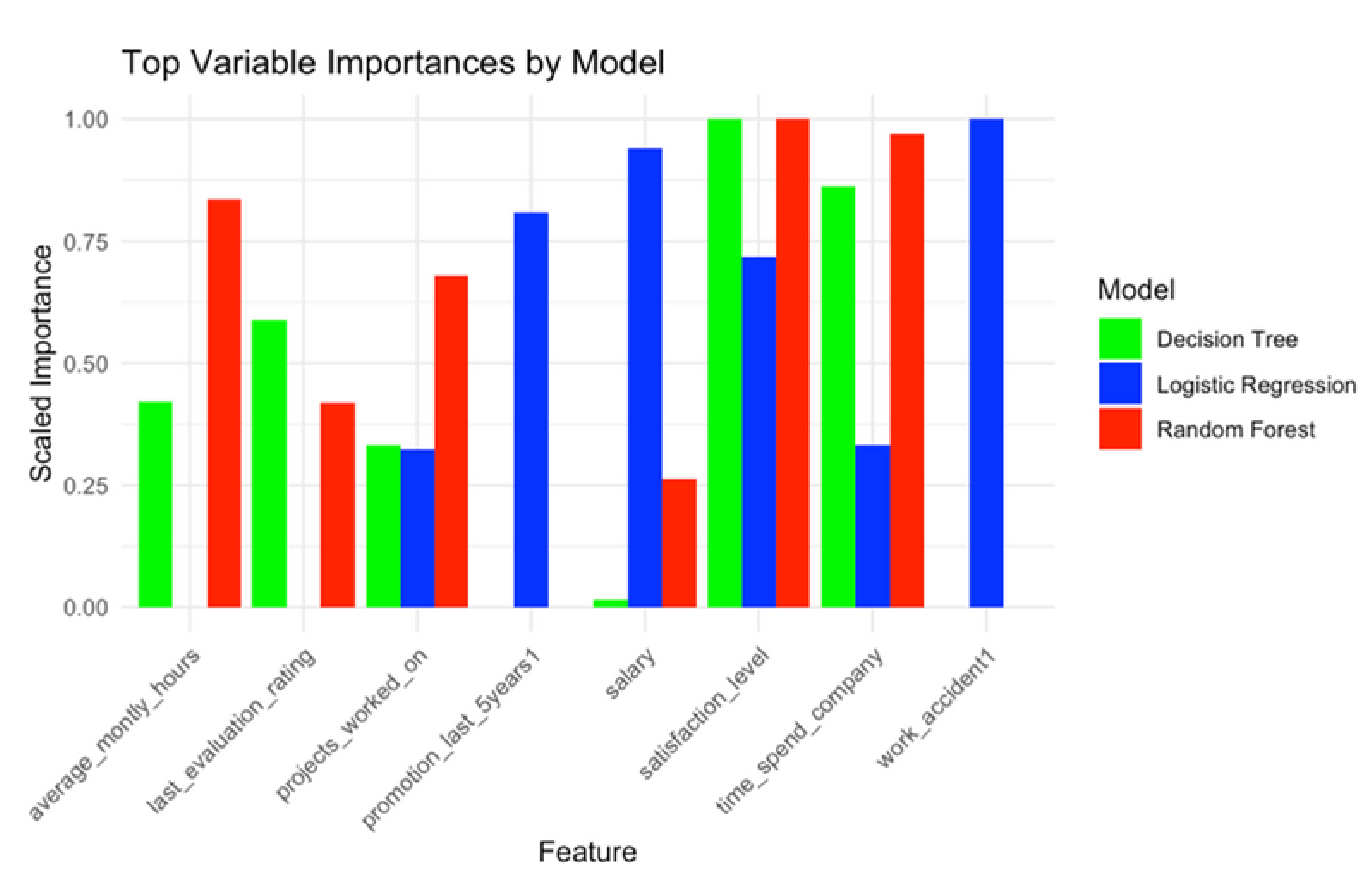
ACCURACY



- **RANDOM FOREST:** REMARKABLE
- **DECISION TREE:** STRONG
- **K-NN:** GOOD ACCURACY
- **LOGISTIC REGRESSION:** RESPECTABLE

RANDOM FOREST PERFORMS WITH THE HIGHEST ACCURACY

FEATURE **IMPORTANCE**

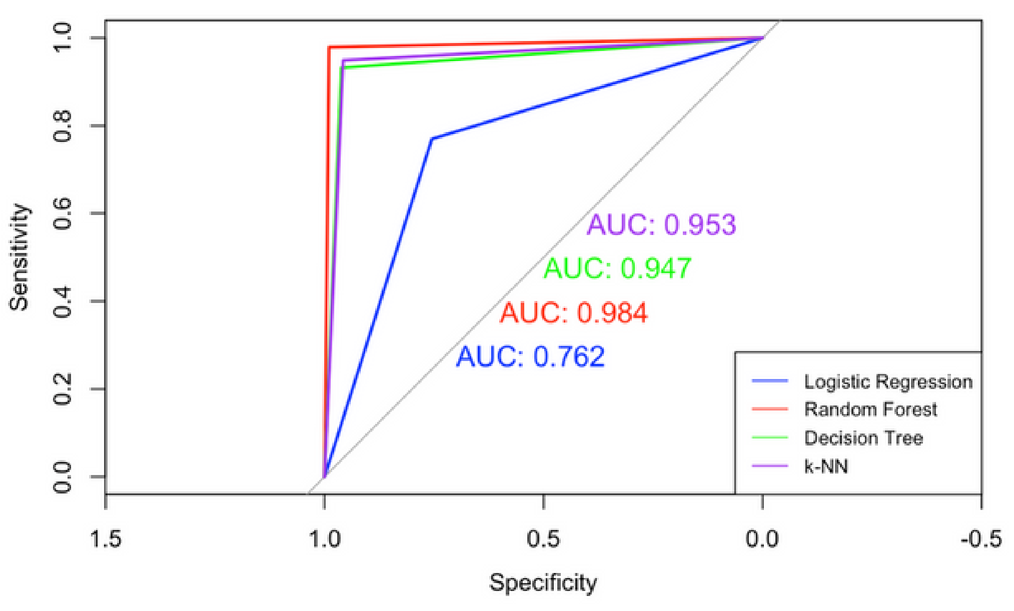


UNDERSTAND **IMPORTANT** FEATURES FOR ATTRITION PREDICTION.

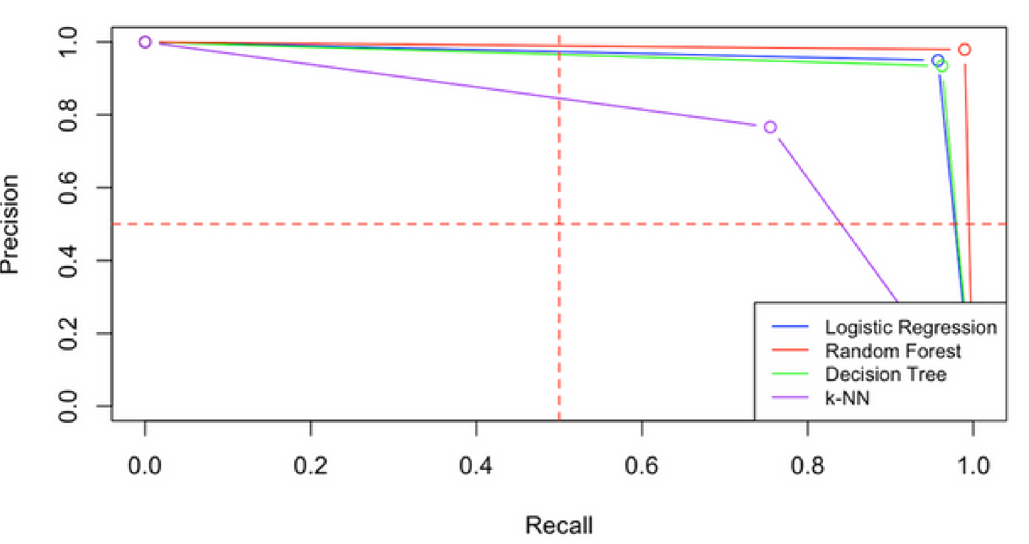
- **SATISFACTION LEVEL:** TOP PREDICTOR.
- **TIME SPENT:** SECOND MOST IMPORTANT.
- **SALARY:** VARIES ACROSS MODELS.
- **PROMOTIONS/LAST EVALUATION:** INCONSISTENT IMPACT.
- **PROJECTS:** IMPORTANT.

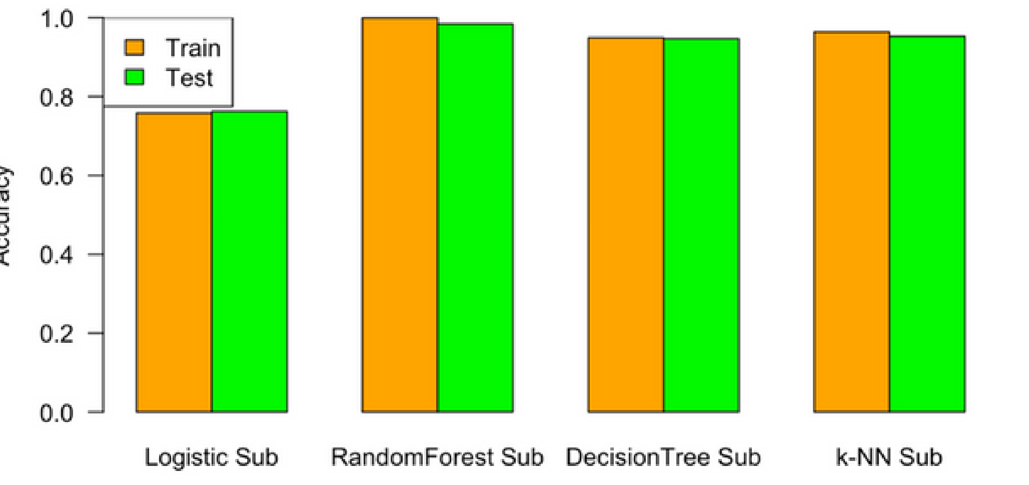
WE'LL USE THE **KEY FEATURES** SHOWN TO BUILD A **FOCUSED** MODEL AND **ASSESS** ITS PREDICTIVE PERFORMANCE.

FEATURE SELECTION



Precision-Recall Curve





FEATURE SELECTION IMPROVES THE MODEL.

FOCUS ON IMPORTANT FEATURES.

SUBSET MODELS **RANDOM FOREST** AND **K-NN KEPT THE HIGH ACCURACY**

DECISION TREE REMAINS GOOD.

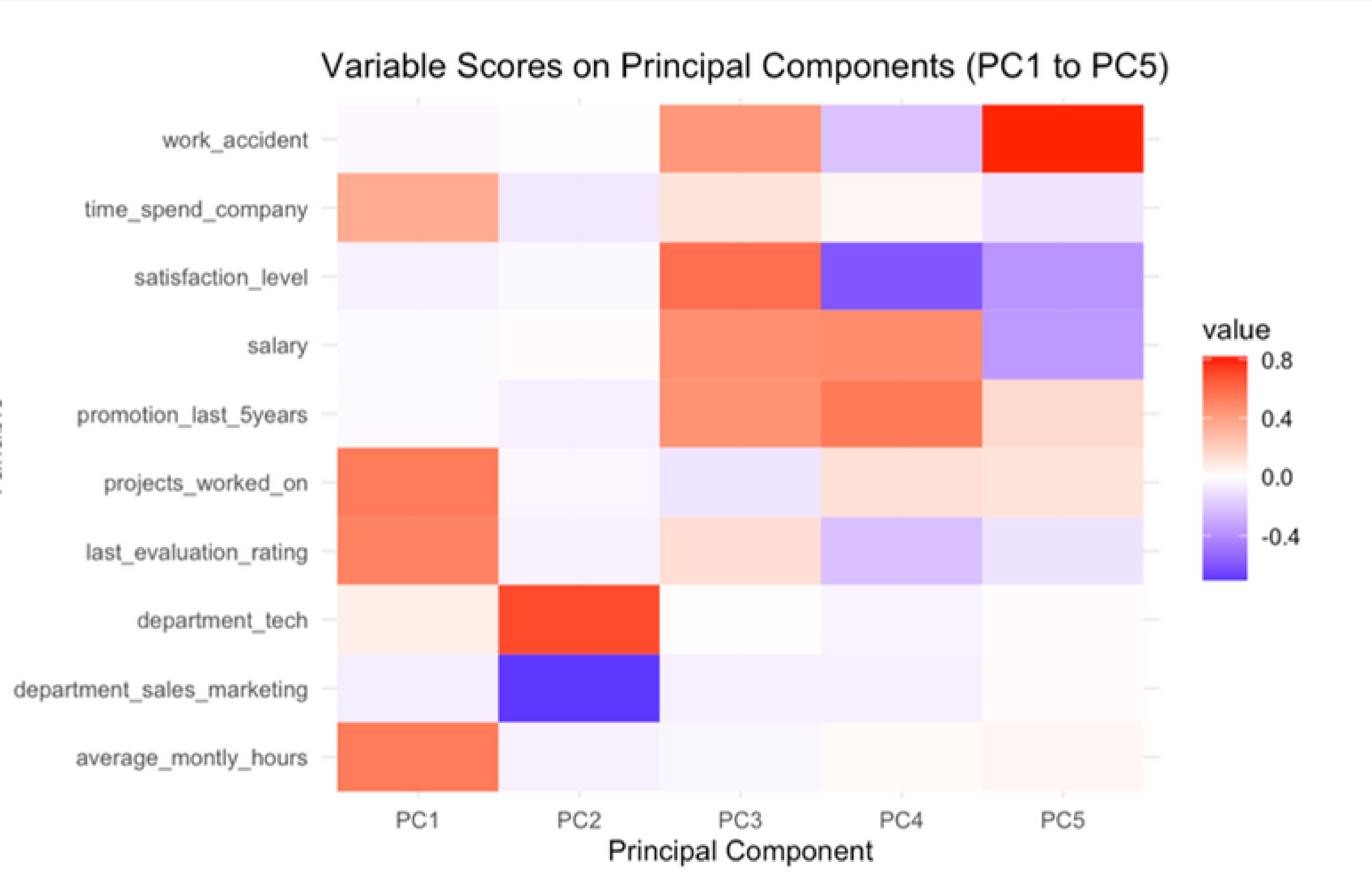
LOGISTIC REGRESSION SEES SLIGHT CHANGES (**LOWER AUC / HIGHER PR**)

- REDUCING FEATURES **OPTIMIZES** PERFORMANCE.

THIS HIGHLIGHTS THE SIGNIFICANCE OF
EMPLOYEE ENGAGEMENT,
WORK-LIFE BALANCE,
COMPENSATION,
AND CAREER GROWTH
IN REDUCING **ATTRITION** AND IMPROVING THE EMPLOYEE EXPERIENCE.

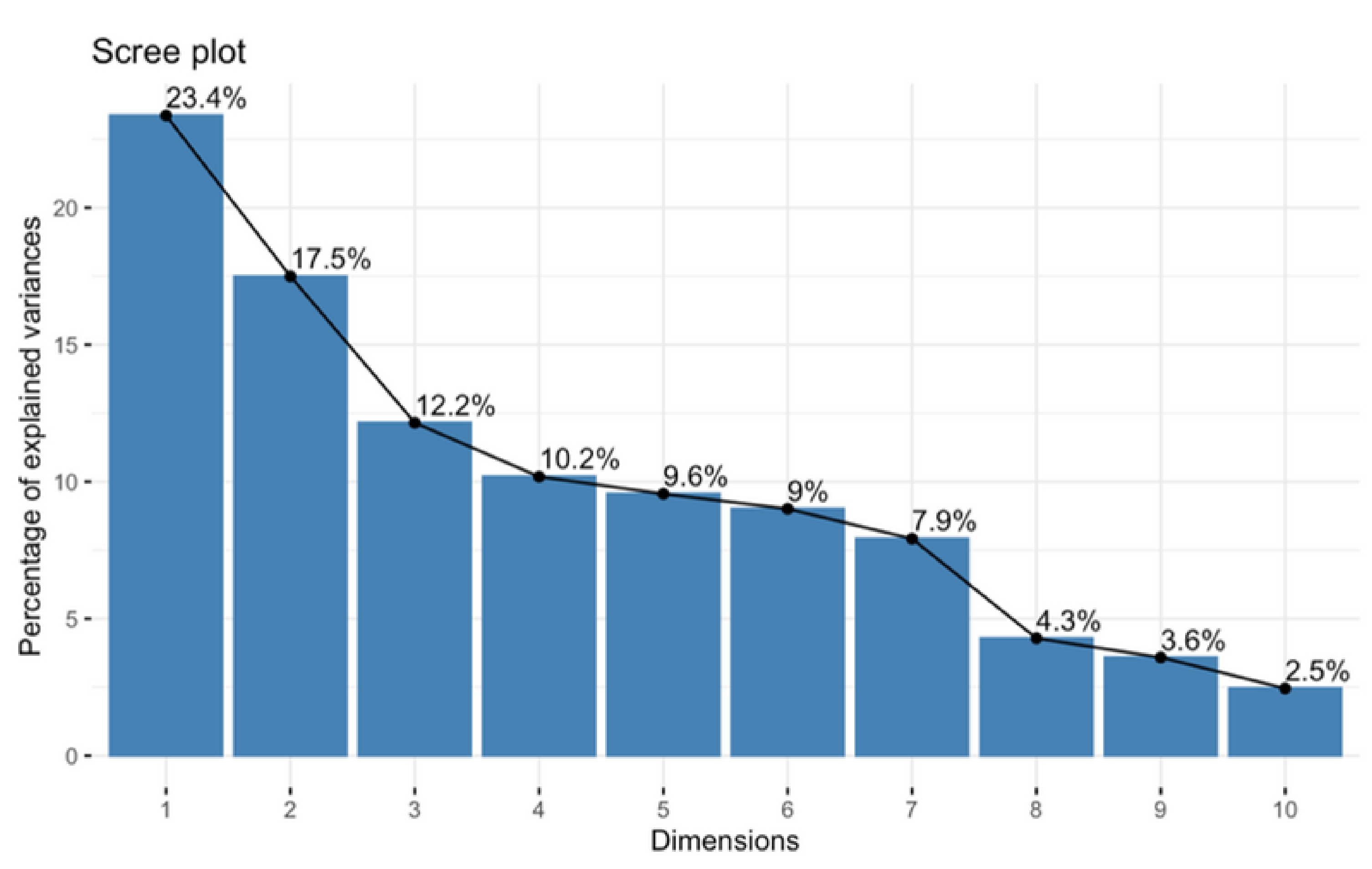
UNSUPERVISED LEARNING

PRINCIPAL COMPONENT ANALYSIS



IT REVEALS **PATTERNS** AND **DEPENDENCIES**.

THE FIRST **5 COMPONENTS** EXPLAIN AROUND **73%** OF THE DATASET **VARIANCE.**

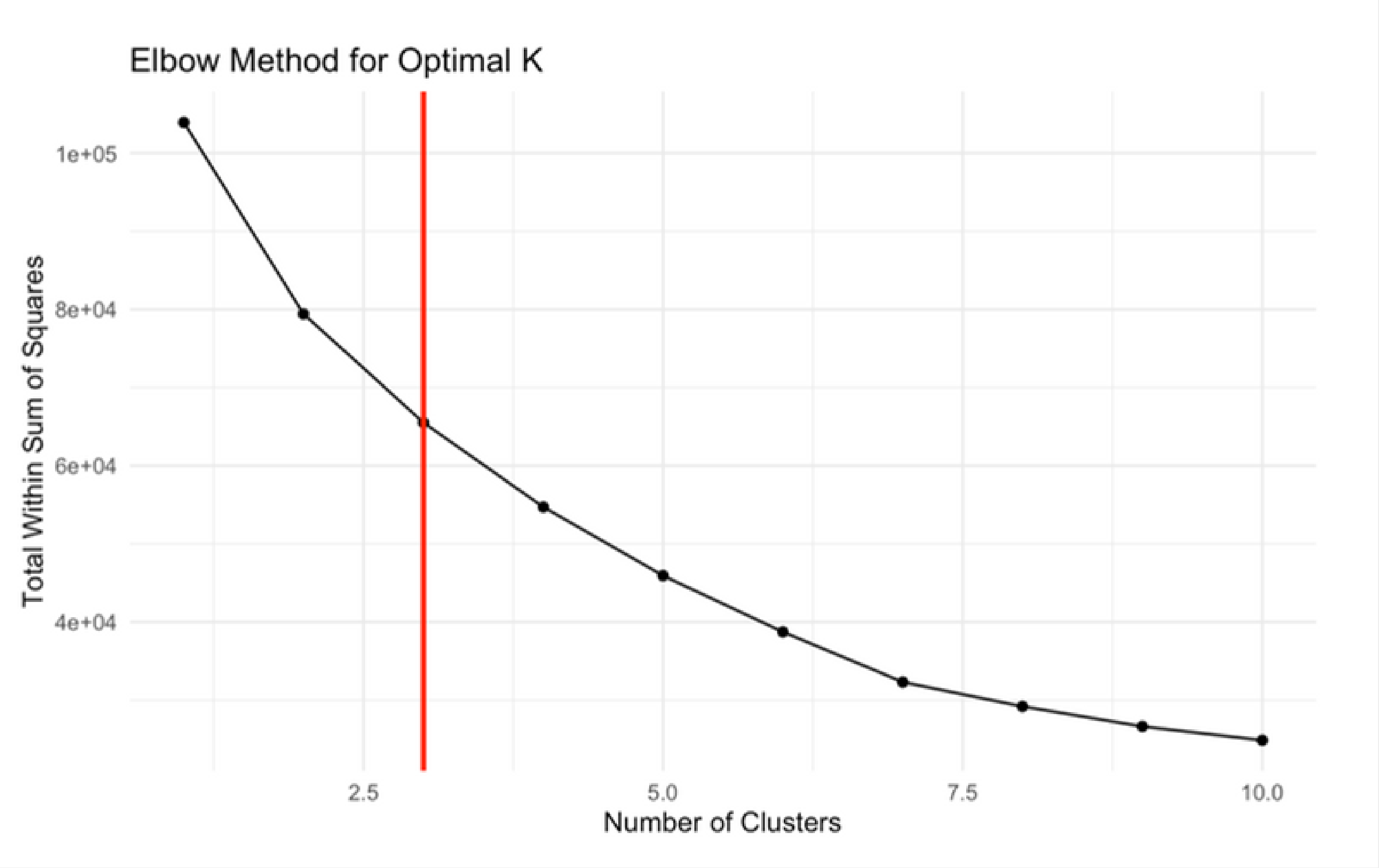


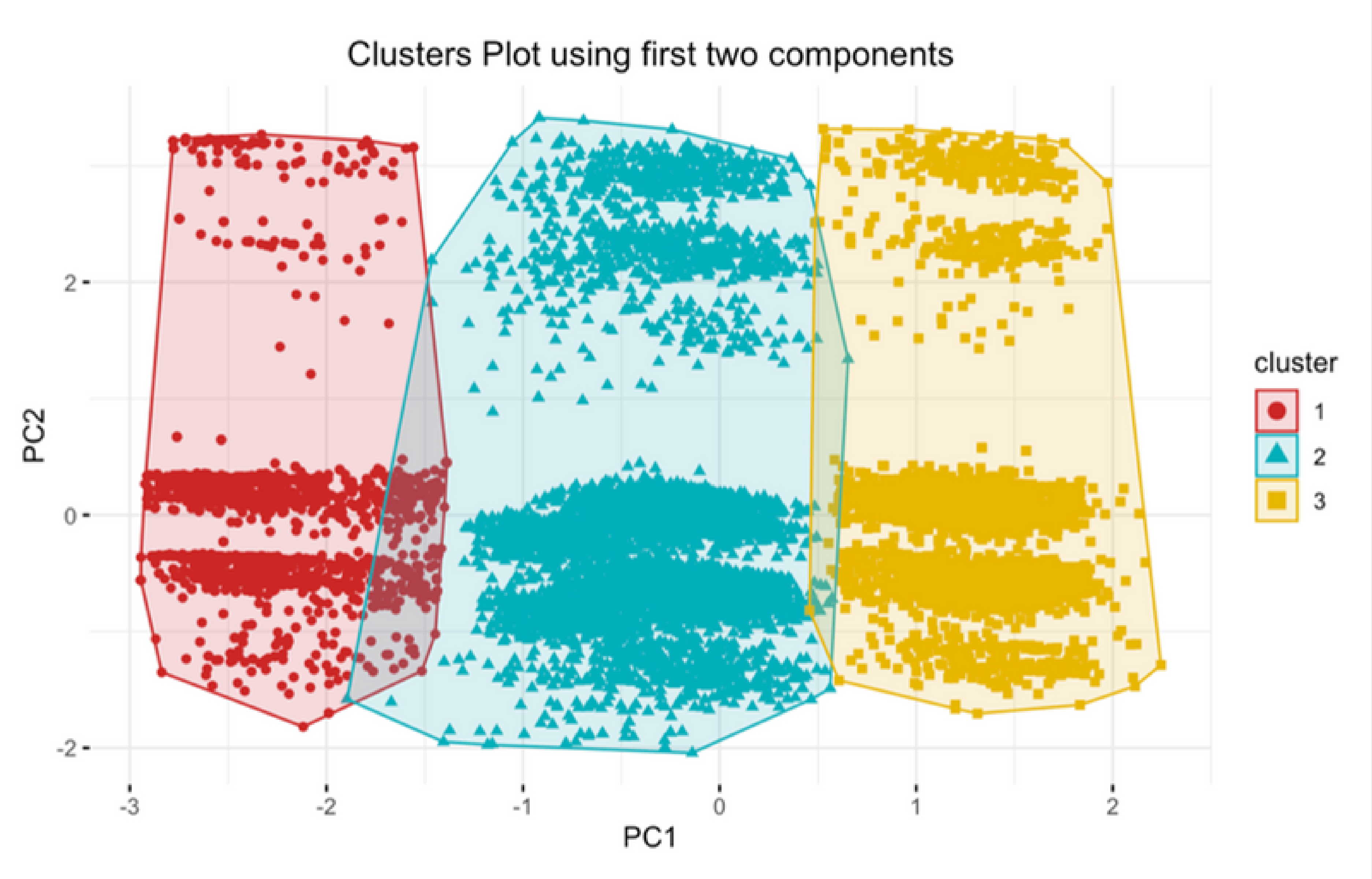
PC1: WORK ENGAGEMENT.
PC2: DEPARTMENTAL VARIATION.
PC3: JOB SATISFACTION.
PC4: CAREER PROGRESSION.
PC5: SAFETY-SATISFACTION TRADE-OFF.

CLUSTERING ANALYSIS:

- **K-MEANS** CLUSTERING ON THE FIRST **5 PCS.**
- VISUALIZED CLUSTERS IN **PC1** AND **PC2**.
- **ELBOW METHOD**.

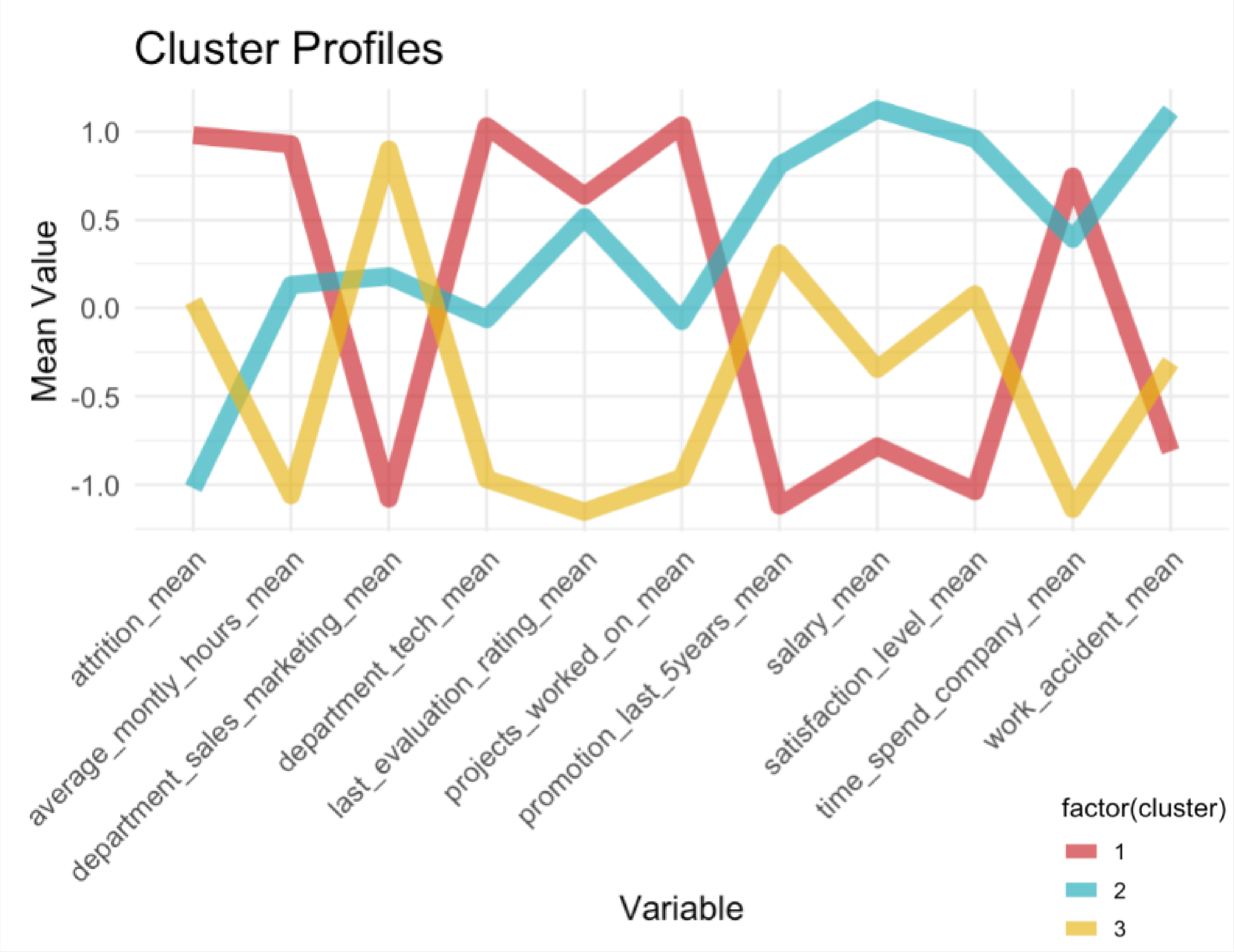
  OPTIMAL CLUSTERS DETERMINED AS **K = 3.**





- SHOWS **DISTRIBUTION OF EMPLOYEES** WITHIN EACH **CLUSTER**.
- ILLUSTRATES HOW CLUSTERS RELATE TO **WORK COMMITMENT** (PC1) AND **DEPARTMENT** (PC2).

CLUSTERS PROFILES



- CLUSTER 1 - "HIGH POTENTIAL, HIGH ATTRITION"

ACTIONS:
SURVEYS,
SALARY ADJUSTMENTS,
CAREER PROGRESSION OPPORTUNITIES.

- CLUSTER 2 - "LOYAL HIGH PERFORMERS"

ACTIONS:
RECOGNITION,
CAREER GROWTH OPPORTUNITIES,
MENTORING.

- CLUSTER 3 - "NEW TALENT"

ACTIONS:
TRAINING,
CAREER ADVANCEMENT PATHS,
MONITORING PROGRESS.

CONCLUSION...

SUMMARY OF FINDINGS

- **Key Factors:** Employee **satisfaction**, **evaluations**, **salary**, **promotions**, and **time spent** in the company influence attrition.
- **Model: Random Forest** excelled with fewer features, while **K-NN** and **Decision Tree** adapted well.
- **Clustering:** Identified three key employee clusters: "**New Talent**," "**High Potential, High Attrition,**" and "**Loyal High Performers**."

Actions should be done:

Surveys and Feedback

Fix Salaries

More Opportunities

Training Programs

Promotion and Rewards

Happy employees lead to better productivity

THANK YOU