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

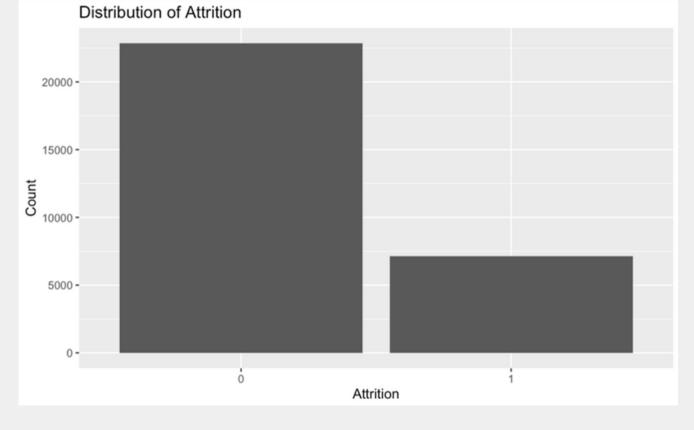Target /Binary

DATA PREPROCESSING

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

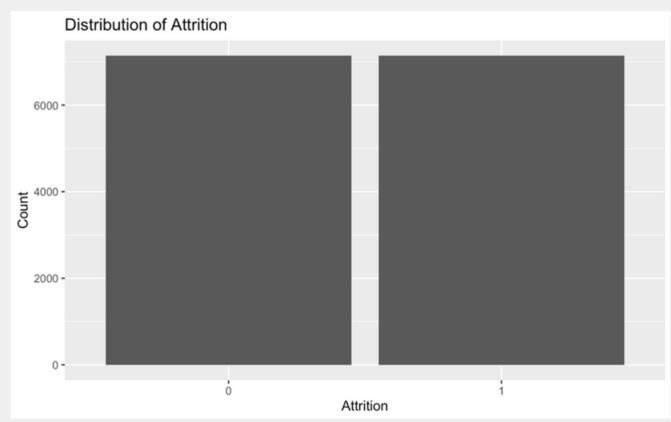- **CLASS IMBALANCE:**
  MORE "NO ATTRITION" (CLASS 0) THAN "ATTRITION" (CLASS 1).
  IMBALANCE CAN AFFECT MODEL ACCURACY.

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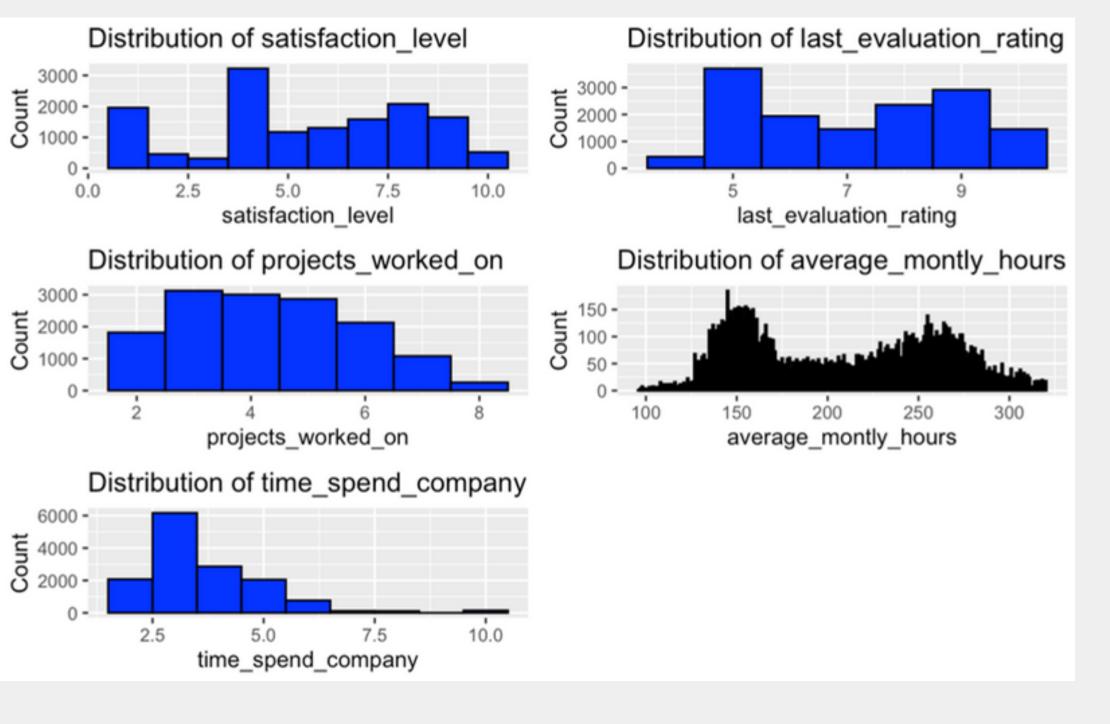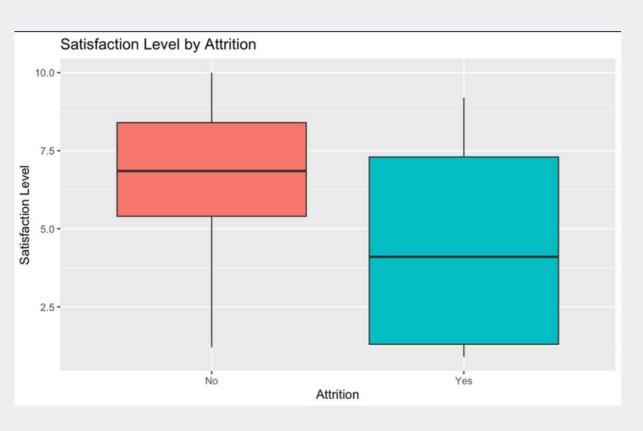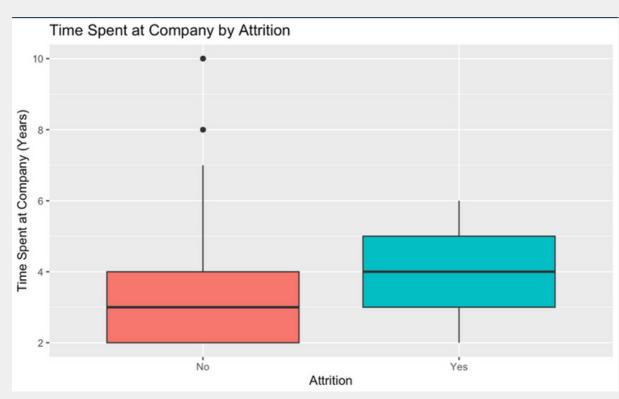 BALANCING CLASSES PREVENTS MODEL BIAS.

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)
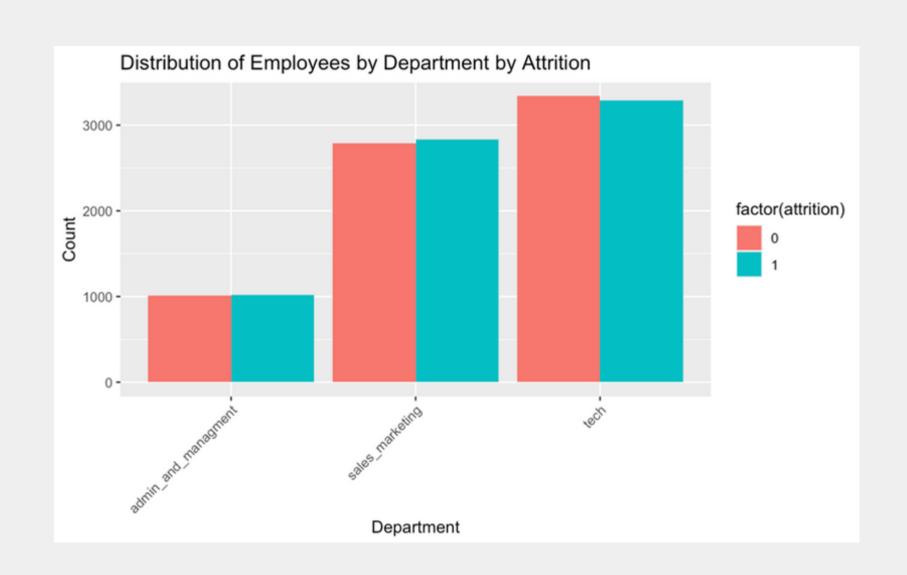
CATEGORICAL VARIABLES

Distribution of Employees by Department by Attrition

Distribution of Salary Levels by Attrition

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

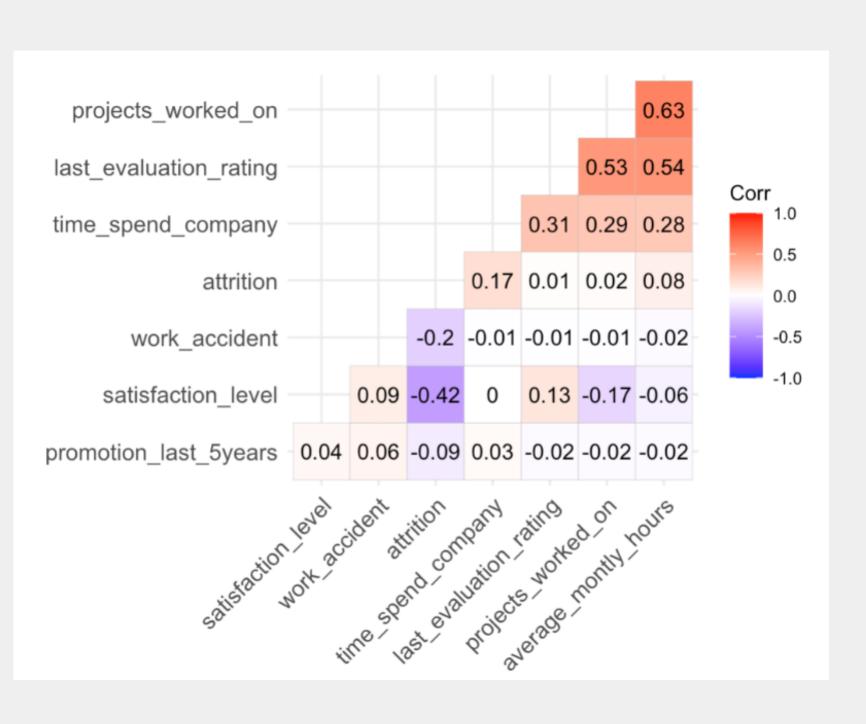- **KEY FACTOR:** "SATISFACTION LEVEL" STRONGLY INFLUENCES ATTRITION; LOWER SATISFACTION CORRELATES WITH HIGHER ATTRITION.

- **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 SIGNIFICANT ISSUES OF MULTICOLLINEARITY

ALL VIF VALUES ARE NEAR 1

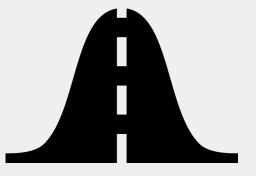# SUPERVISED LEARNING

DATA PREPROCESSING

CATEGORICAL ENCODING:

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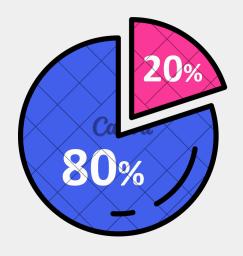
NUMERICAL SCALING:

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.

- EACH TREE CONSIDERS A RANDOM SUBSET OF THREE VARIABLES AT EACH SPLIT.

PERFORMANCE:

- **OUT-OF-BAG (OOB) ERROR RATE:** ~1.69%, INDICATING HIGH ACCURACY.

- CONFUSION MATRIX SHOWS LOW-CLASS ERROR RATES FOR BOTH ATTRITION AND NO ATTRITION CASES.

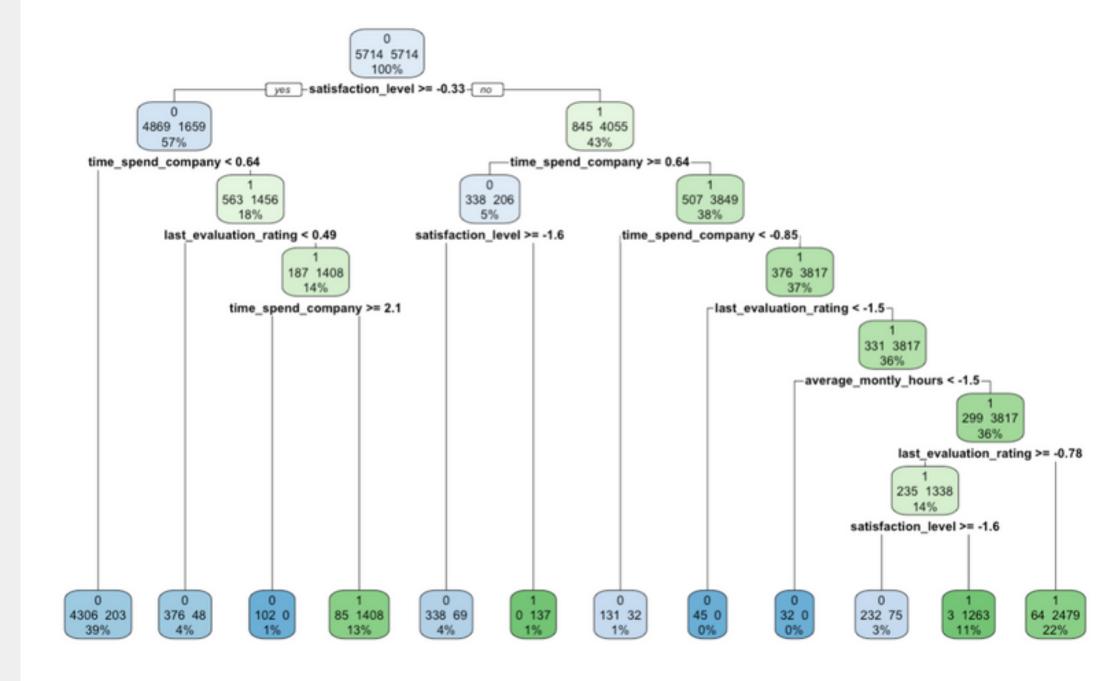- EFFECTIVE AT CLASSIFYING EMPLOYEES WHO LEAVE AND THOSE WHO STAY.

DECISION TREE

- PREDICTS ATTRITION BASED ON ALL FEATURES THROUGH RECURSIVE PARTITIONING.
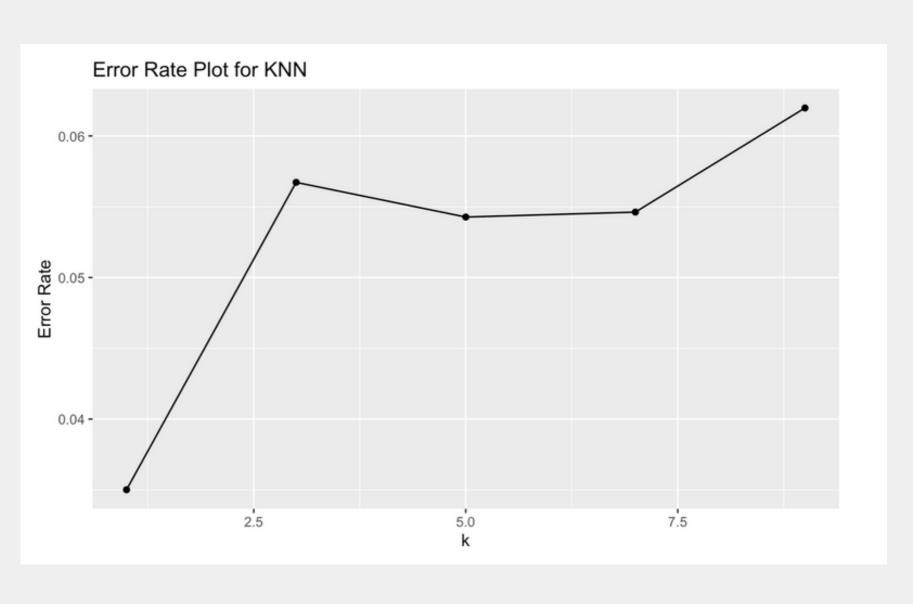
- EACH BRANCH REPRESENTS CONDITIONS LEADING TO ATTRITION OR NON-ATTRITION.

- **KEY FACTORS IN THE TREE STRUCTURE:**
- SATISFACTION LEVEL
- TIME SPENT IN COMPANY
- LAST EVALUATION RATING
- AVERAGE MONTHLY HOURS

K-NEAREST NEIGHBOR

Error Rate Plot for KNN

Error Rate

0.06

0.05

0.04

k

2.5   5.0   7.5

OPTIMAL "K" VALUE:

- K=1: ERROR RATE ≈ 0.0350, SENSITIVE TO NOISE OR OUTLIERS.
- K=3, 5, 7, 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, BALANCED SENSITIVITY, AND SPECIFICITY.

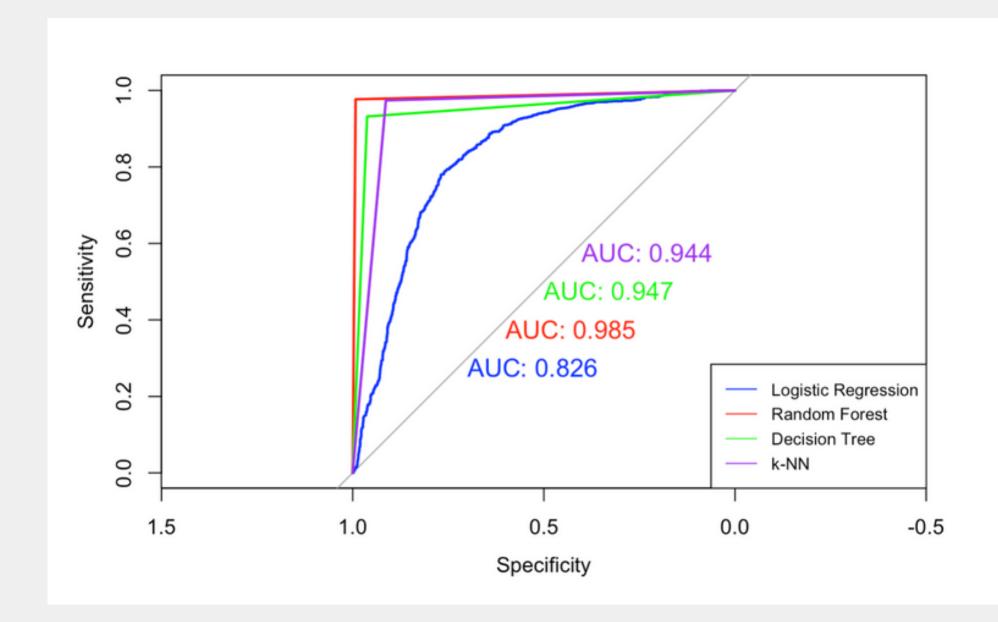GENERALIZES EFFECTIVELY TO NEW DATA, ROBUST FOR PREDICTING ATTRITION.

ROC-AUC

ROC-AUC IS A CRITICAL METRIC FOR BINARY CLASSIFICATION MODELS.

IT MEASURES A MODEL'S ABILITY TO DISTINGUISH POSITIVE AND NEGATIVE CLASSES**.**

- **LOGISTIC REGRESSION:**
  MODERATE DISCRIMINATIVE POWER.

- **RANDOM FOREST:**
  VERY HIGH DISCRIMINATIVE POWER.

- **DECISION TREE:**
  STRONG DISCRIMINATIVE ABILITY.

- **K-NN:**
  GOOD DISCRIMINATIVE POWER.

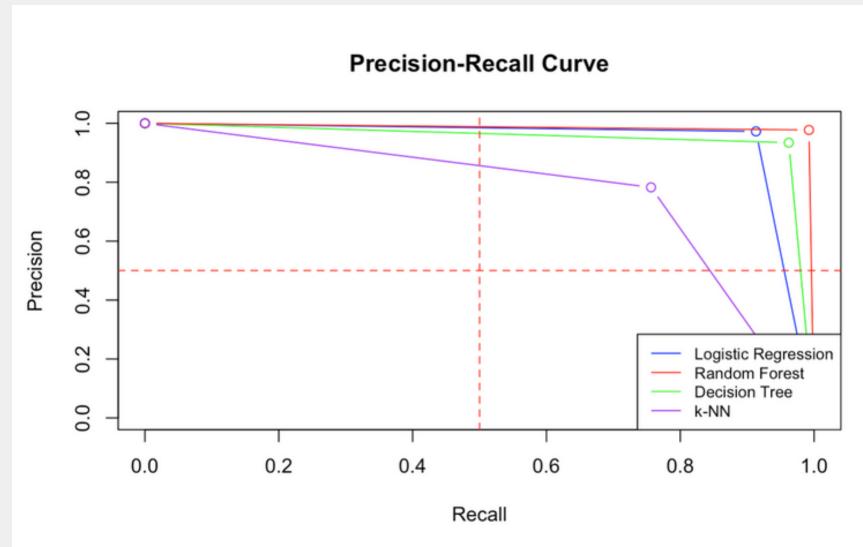**RANDOM FOREST** PERFORMED BEST IN DISTINGUISHING ATTRITION

PRECISION-RECALL

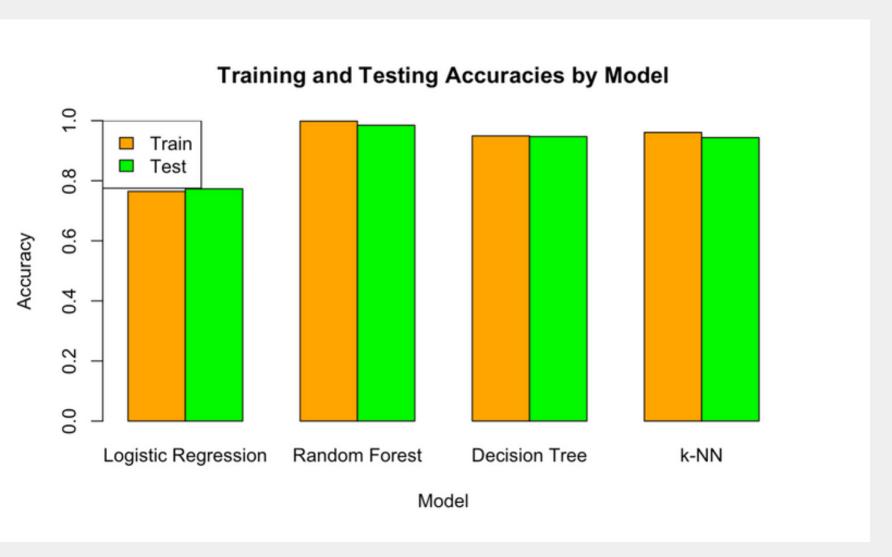**PRECISION** IS CONCERNED WITH HOW MANY OF THE PREDICTED POSITIVES ARE ACTUALLY CORRECT.

RECALL IS CONCERNED WITH HOW MANY OF THE ACTUAL POSITIVES WERE CORRECTLY PREDICTED.

- **RANDOM FOREST:**
HIGH PRECISION AND RECALL SCORES

- **DECISION TREE:**
GOOD BALANCE BETWEEN PRECISION AND RECALL.

- **K-NN :**
BALANCED PRECISION AND RECALL.

- **LOGISTIC REGRESSION:**
RESPECTABLE PRECISION AND RECALL

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

Precision-Recall Curve

ACCURACY

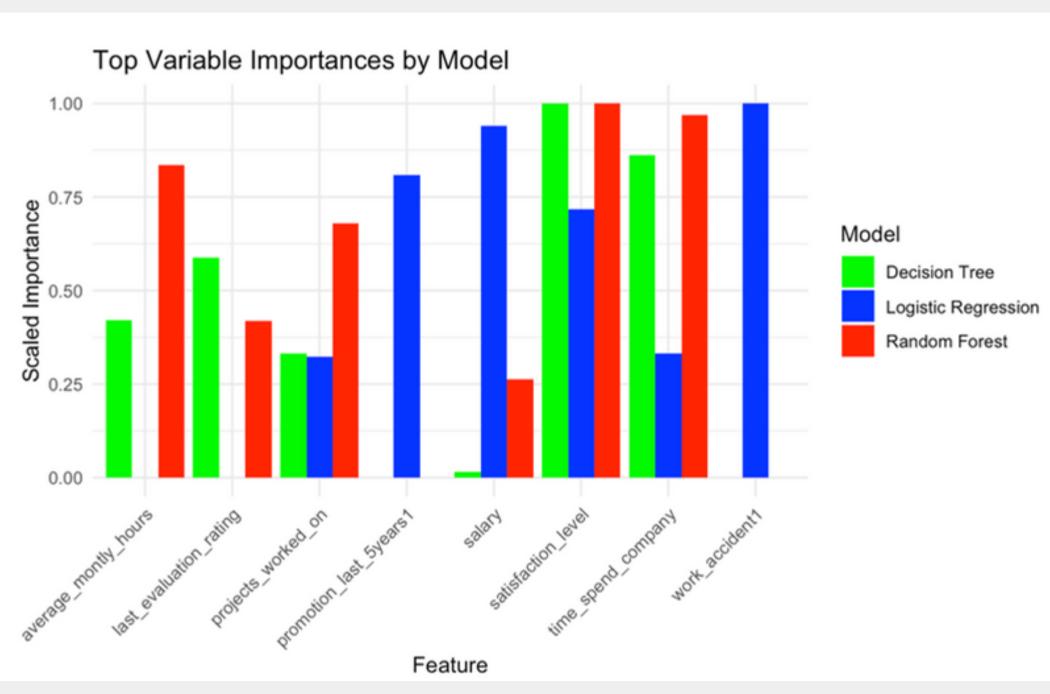Training and Testing Accuracies by Model

ACCURACY MEASURES THE OVERALL CORRECTNESS OF THE MODEL'S PREDICTIONS.

- **RANDOM FOREST:** REMARKABLE ACCURACY ON BOTH TRAINING AND TEST DATASETS.
- **DECISION TREE:** STRONG ACCURACY.
- **K-NN:** GOOD ACCURACY.
- **LOGISTIC REGRESSION:** RESPECTABLE ACCURACY.

RANDOM FOREST PERFORMS WITH THE HIGHEST ACCURACY

FEATURE IMPORTANCE
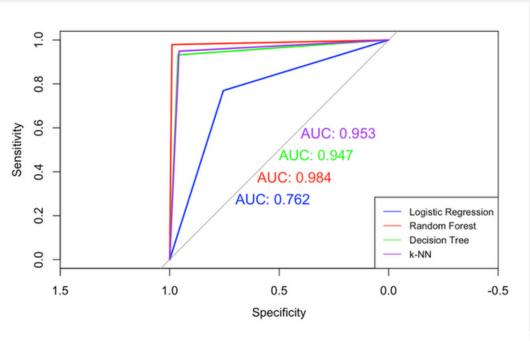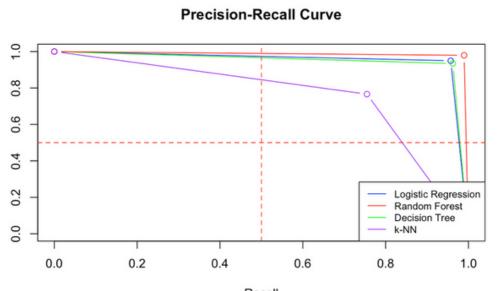

Top Variable Importances by Model

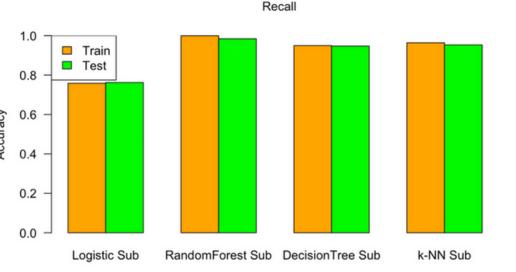UNDERSTAND **IMPORTANT** FEATURES FOR ATTRITION PREDICTION.

- **SATISFACTION LEVEL:** TOP PREDICTOR.

- **TIME SPENT:** SECOND MOST IMPORTANT.

- **SALARY:** VARIES ACROSS MODELS.

- **PROMOTIONS/LAST EVALUATION:** INCONSISTENT IMPACT.

- **OTHERS:** VARYING IMPORTANCE.

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

FEATURE SELECTION

FEATURE SELECTION IMPROVES MODEL EFFICIENCY AND INTERPRETABILITY.

FOCUS ON CRUCIAL FEATURES SHOWN IN THE "IMPORTANCE" PLOT.

SUBSET MODELS **MAINTAIN STRENGTH**, ESPECIALLY **RANDOM FOREST** AND **K-NN.**
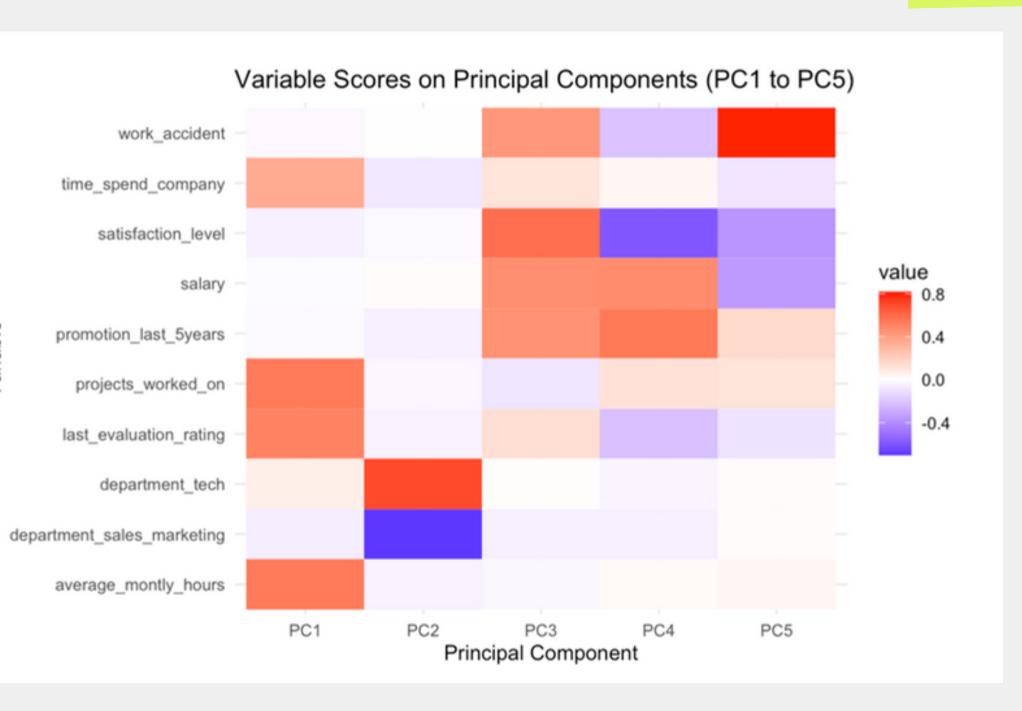
DECISION TREE REMAINS ROBUST.

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

Variable Scores on Principal Components (PC1 to PC5)

PCA IS A POWERFUL TOOL FOR EXPLORING DYNAMICS.

IT REVEALS **PATTERNS** AND **DEPENDENCIES.**
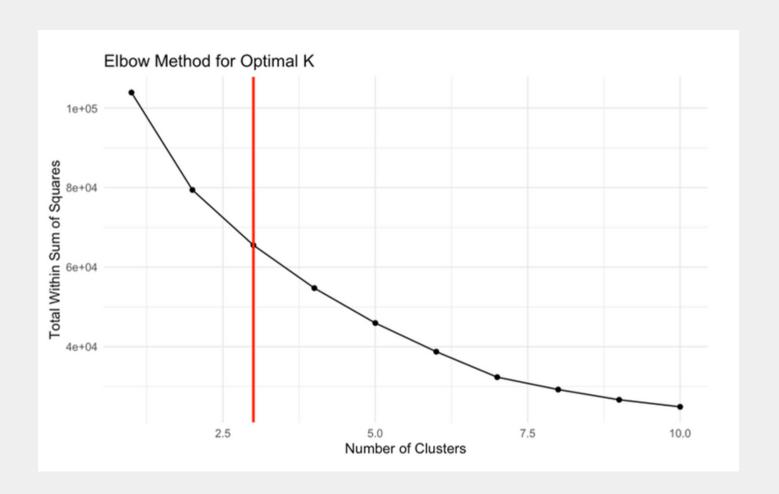
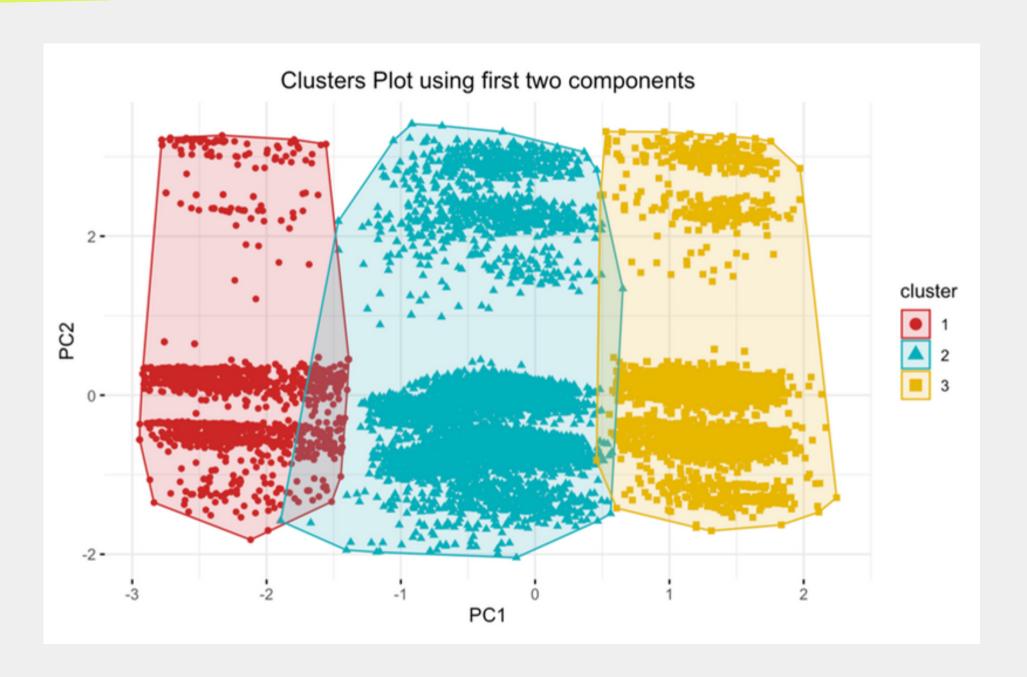THE FIRST **5 COMPONENTS** EXPLAIN AROUND **73%** OF THE DATASET **VARIANCE.**

Scree plot

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.**
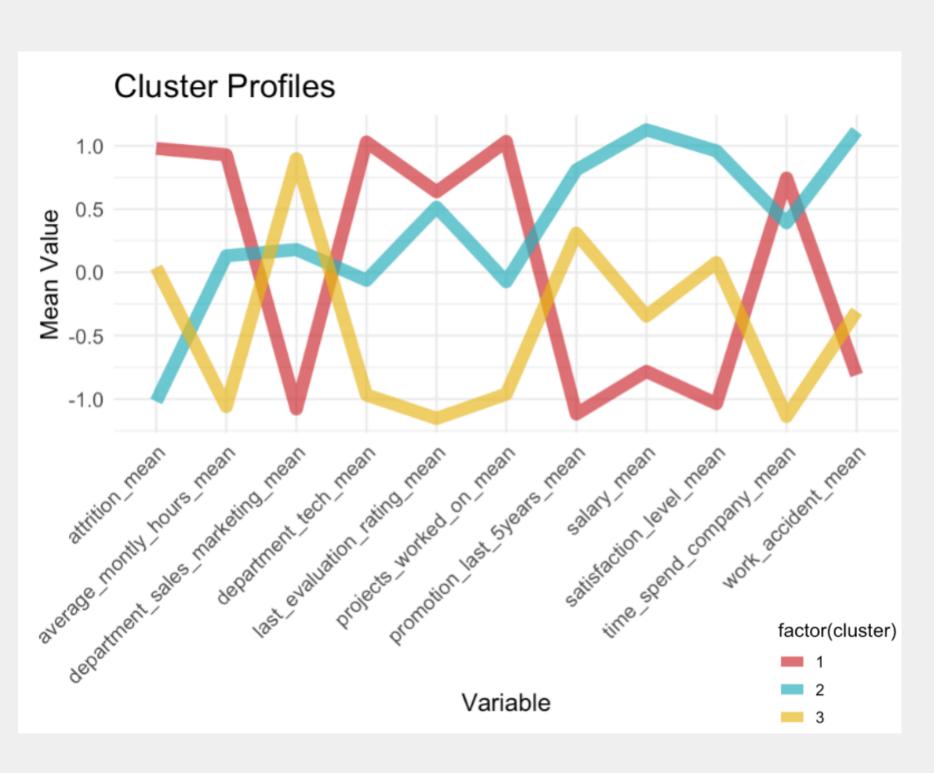

Clusters Plot using first two components

Elbow Method for Optimal K

- 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:
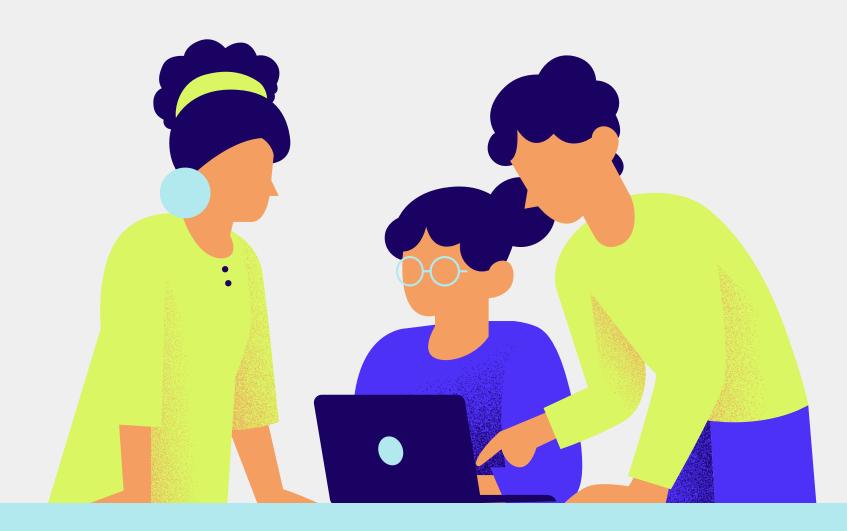TRAINING, CAREER ADVANCEMENT PATHS, MONITORING PROGRESS.

CONCLUSION...

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

Actions should be done:
Regular Surveys and Feedback
Salary Adjustments
Career Progression Opportunities
Training and Development Programs
Recognition and Rewards

Happy employees lead to better productivity

THANK YOU