3SIGMA SQUAD

PREDICTIVE ANALYTICS FOR EMPLOYEE ATTRITION

Mentor: Muhammad Hanif Fajari

Final Project by 3Sigma Squad
Data Science Batch 54
Rakamin Academy

July 19, 2025

3SIGMA SQUAD

Employee Attrition Rate 2015:
A Warning Sign We Shouldn’t Ignore

16.1%

Target rate varies by industry, aim below 10% (shrm.org)

One in six left this year – are we listening?

Direct Financial • Intangible Capital • Operational Disruptions

Problem & Objectives

Problem Statement

XYZ Company experiences **high employee attrition rates**, running at around **16.1% per year**, which **negatively impacts** project flow, recruitment costs, and training efficiency

Goal

To reduce the annual employee attrition rate at XYZ Company from 16.1% **to 10%** within the next **12 months** by implementing a **predictive attrition system** that enables proactive interventions by the HR team and management

Objectives

- Analyze historical data to **identify the most significant factors** contributing to employee attrition
- Develop a **classification model** to **predict employee attrition** with **high accuracy** using metrics suited for imbalanced data
- Translate model predictions into **clear reports** and **dashboards** for HR and management to use for proactive intervention

SUCCESS METRICS

3SIGMA SQUAD

-30%

Attrition Rate

Saving ~200 employees annually

175%

Return on Invesment (ROI)

Every $1 invested yields $1.75 in return

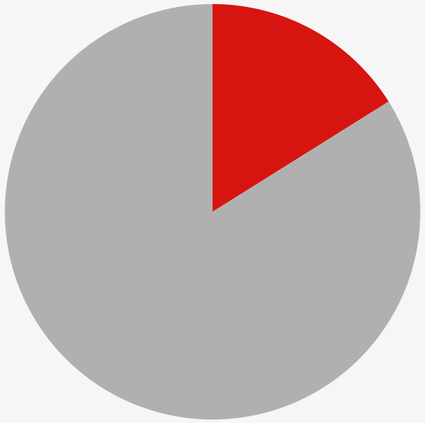
-30%

Key Talent Attrition

Protecting critical talent

Industry Relevance

Employee attrition is a significant and universal business challenge, **impacting costs**, **productivity**, and **morale** across various sectors. This project is crucial for **maintaining operational stability**, **retaining key talent**, and **enhancing** the **company's competitiveness**.



Attrition

16.1%

Retention

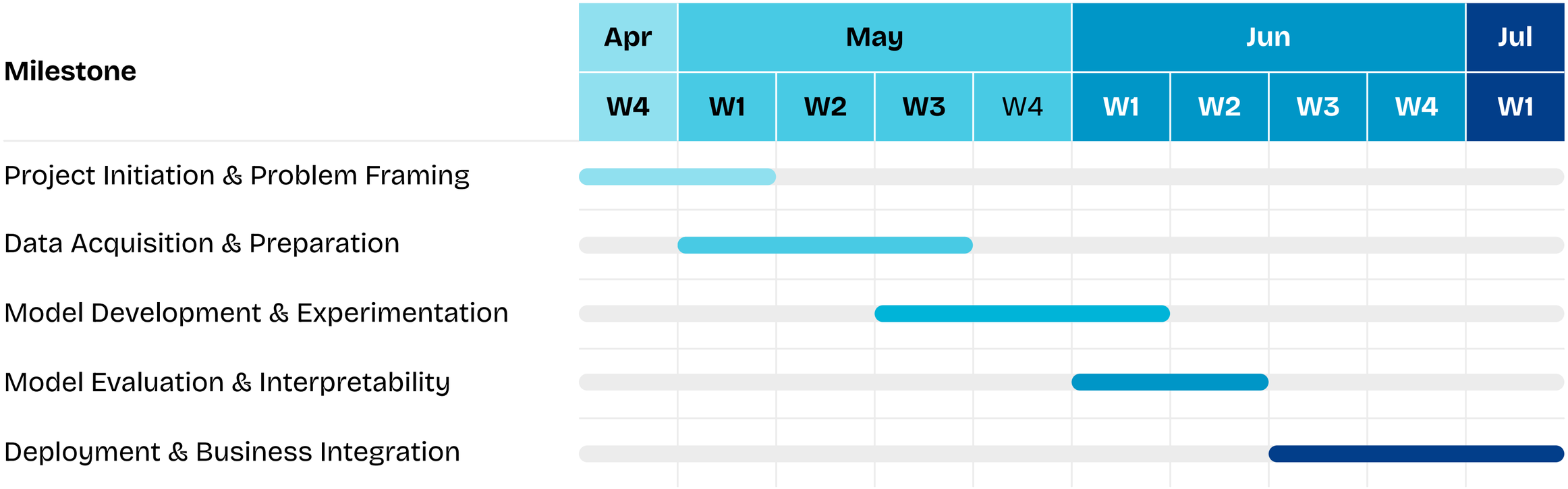
83.9%

PROJECT WORKFLOW & TIMELINE

3SIGMA SQUAD

This workflow and timeline are designed based on the **CRISP-DM** methodology, breaking down the entire project into **5 clear phases** ensuring that each stage is well-planned and executed within a structured timeframe.

Business Understanding → **Data Understanding** → **Data Preparation** → **Modeling** → **Evaluation** → **Deployment**



Complete Project Plan: [click here]

DATA OVERVIEW

3SIGMA SQUAD

Data Source
Internal Company Data

Scope
4410 employee

Target Variable
Attrition (Yes/No)

Data Period
2015 Snapshot

Data Quality Insight

- **Outliers: 60% detected by IQR**
- **Missing Values: Only 2.5%**
- **Duplicates: 0%**
- **Imbalance: 1:6 ratio**

Features

Demographic & Personal
Age, Gender, MaritalStatus Education, EducationField DistanceFromHome, Over18 NumCompaniesWorked

Career Tenure & History
TotalWorkingYears, YearsAtCompany YearsSinceLastPromotion, YearsWithCurrManager TrainingTimesLastYear

Satisfaction
JobSatisfaction, EnvironmentSatisfaction WorkLifeBalance, JobInvolvement

Job & Role Information
Department, JobRole, JobLevel, BusinessTravel, EmployeeCount, EmployeeNumber, StandardHours

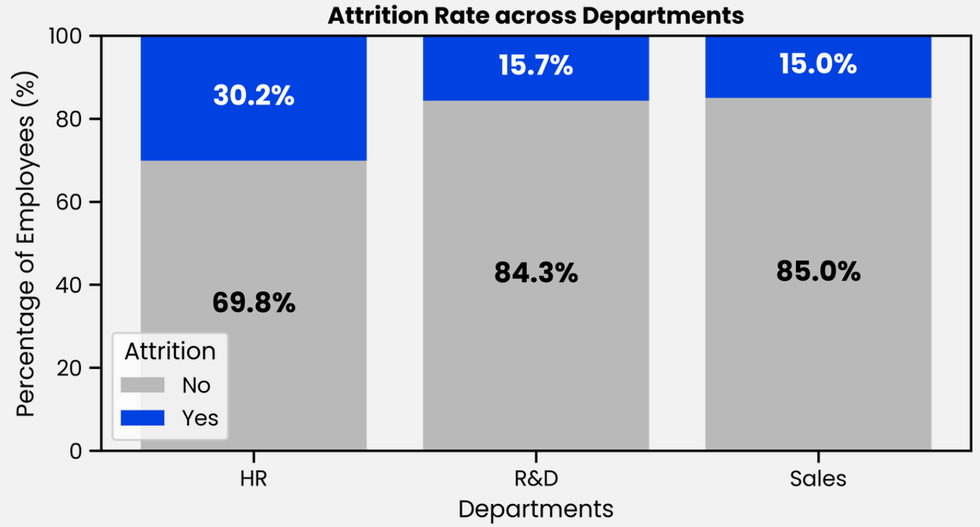
Compensation & Benefit
MonthlyIncome, PercentSalaryHike PerformanceRating, StockOptionLevel

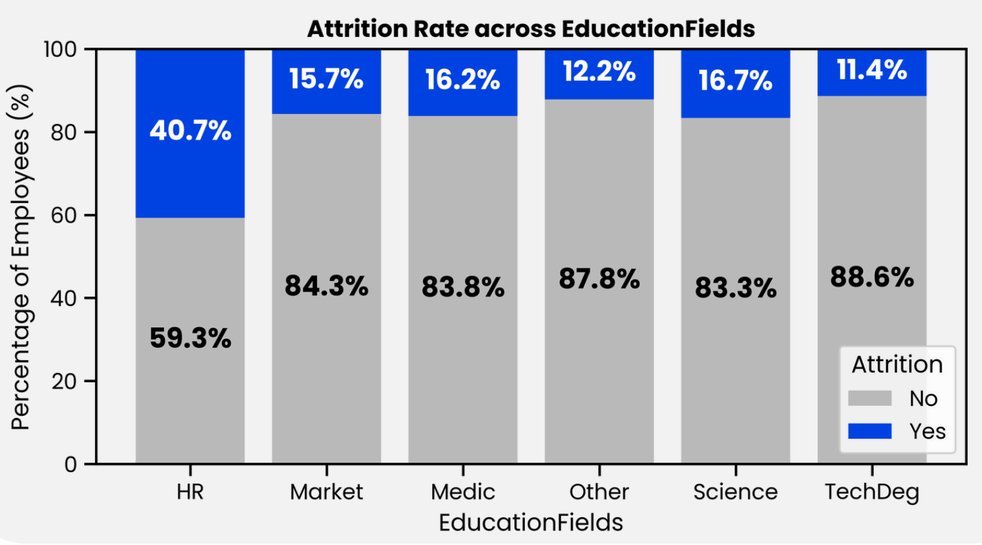
Attendance
in_time out_time

Dataset available at: [click here]

EXPLORATORY DATA ANALYSIS

3SIGMA SQUAD





Why is HR Leaving?

Attrition is **highly concentrated in HR**, with a **30.2%** turnover rate—double that of other departments—and **40.7%** among those with HR education. Even after controlling for key factors, being in HR **increases attrition odds by 2.22×**, pointing to potential systemic issues.

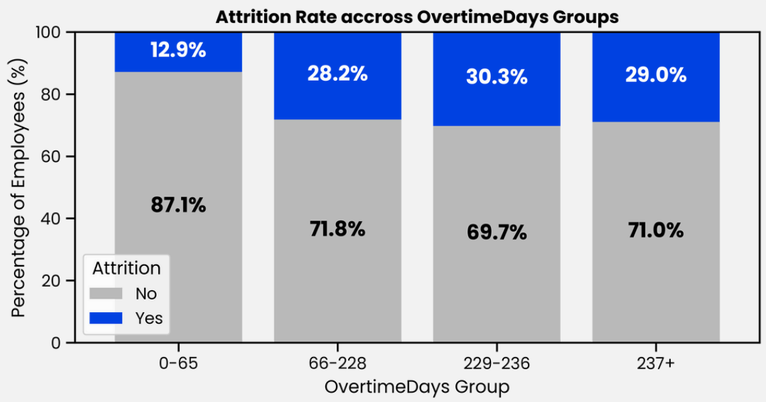
Statistical Findings

Department: significant effect

EducationField: significant effect

ANALYZING ROOT CAUSES

3SIGMA SQUAD

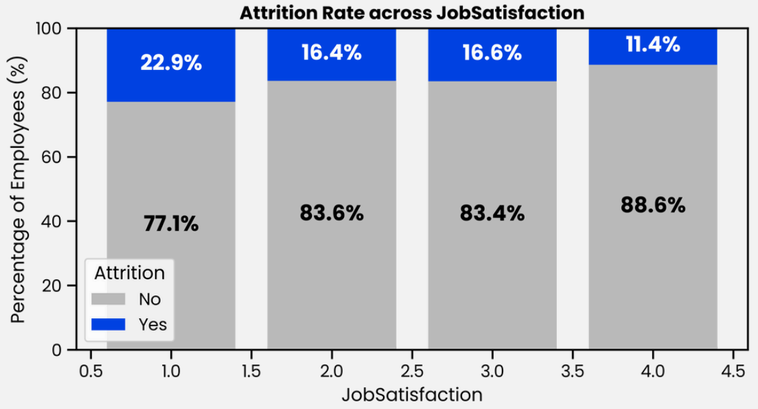


The Real Culprits: Employee Burnout

Our analysis reveals that factors directly related to workload and personal time show a strong and significant impact on an employee's decision to leave.

Statistical Findings

- **OvertimeDays: significant effect**
- **WorkLifeBalance: significant effect**

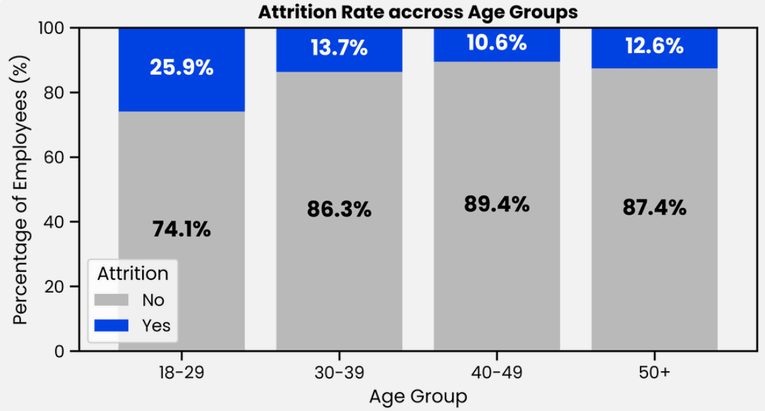


The Hidden Cost of Low Satisfaction

Our data shows a clear, powerful link between low satisfaction scores and an employee's decision to leave the company.

Statistical Findings

- **JobSatisfaction: significant effect**
- **EnvironmentSatisfaction: significant effect**



Are Younger Employees a Higher Flight Risk?

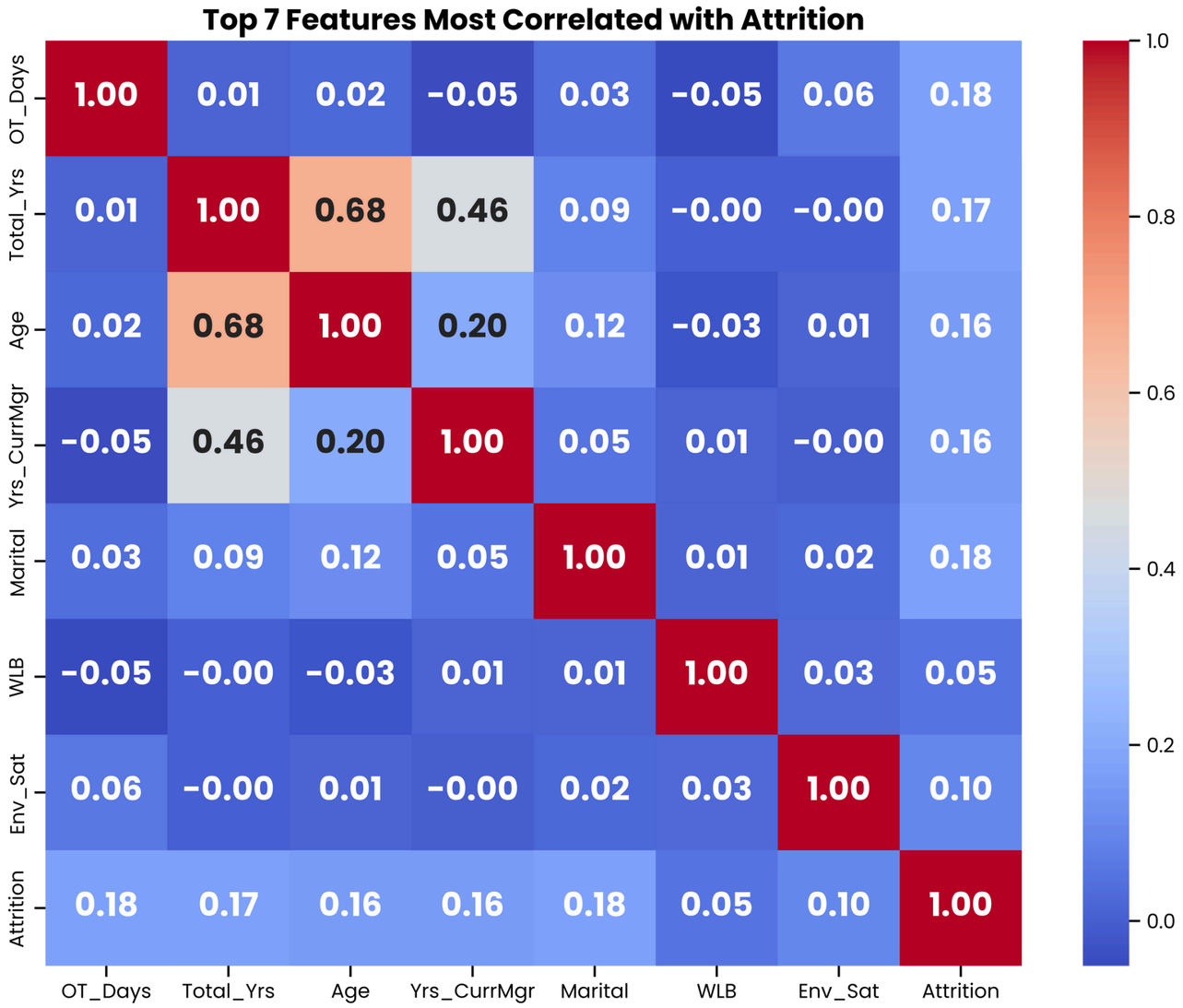
Age and total experience are stronger attrition predictors than pay. Younger, less experienced employees have the highest flight risk.

Statistical Findings

- **Age: significant effect**
- **TotalWorkingYears: significant effect**

MULTIVARIATE ANALYSIS

3SIGMA SQUAD



No "Silver Bullet" for Attrition

Attrition is **highly complex** and isn't driven by a single factor. Research shows **no single variable strongly correlates** with employees' decision to leave (all correlations below 0.2).

"The decision to leave a company is multivariate: There is no single cause."

— Mike West, People Analytics For Dummies (Wiley)

3SIGMA SQUAD

Rakamin Academy

DATA PREPROCESSING

Handling Missing Values • Handling Outliers • Handling Imbalanced Class • Feature Extraction • Feature Encoding • Feature Selection

HANDLING MISSING VALUES & OUTLIERS

3SIGMA SQUAD

The dataset initially had **2.5% missing values**, **imputed** using the **mode** for categorical and the **median** for numerical variables. A **Yeo-Johnson** transformation was then applied to address **~60% extreme values**, reducing outliers to **16.78%**, which now remain only in TotalWorkingYears, TrainingTimeLastYear, and YearsAtCompany.

Data Transformation

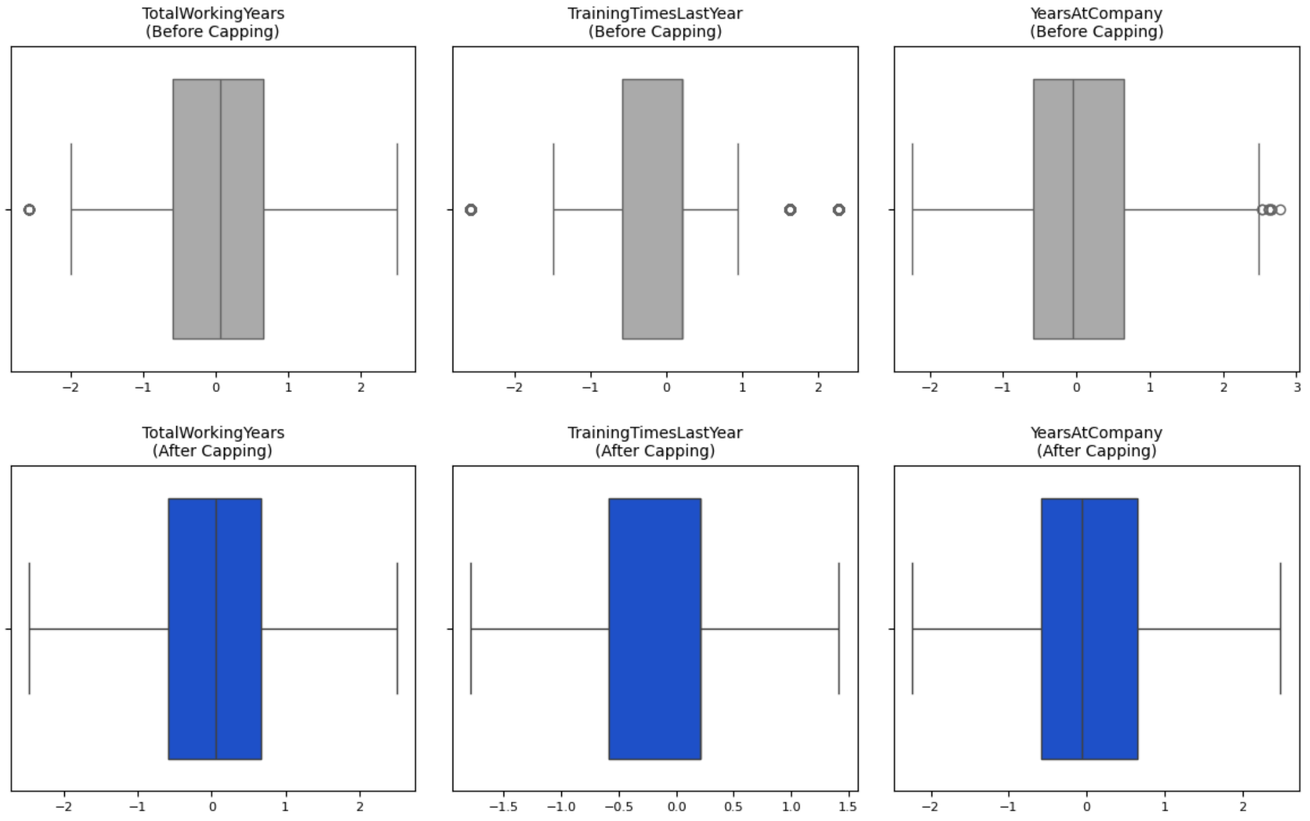
Applied **Yeo-Johnson Transformation** and **standardization** to normalize distributions and ensure uniform scaling

Outlier Detection & Handling

Utilized the **IQR (Interquartile Range)** method for precise identification of remaining outliers, which were handled using **capping**

Validating the Impact

Effectiveness is confirmed using **Skewness Values** and **Histograms** signifying successful transformation



HANDLING IMBALANCED CLASS

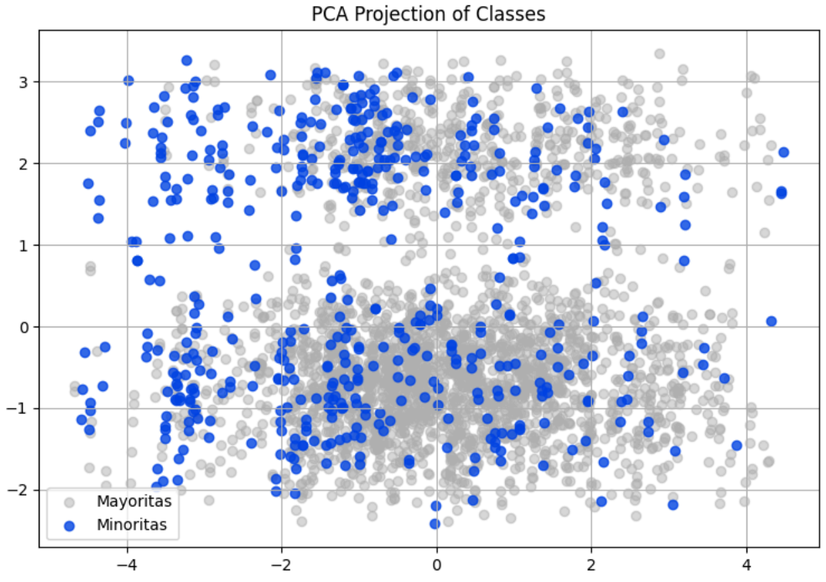
3SIGMA SQUAD

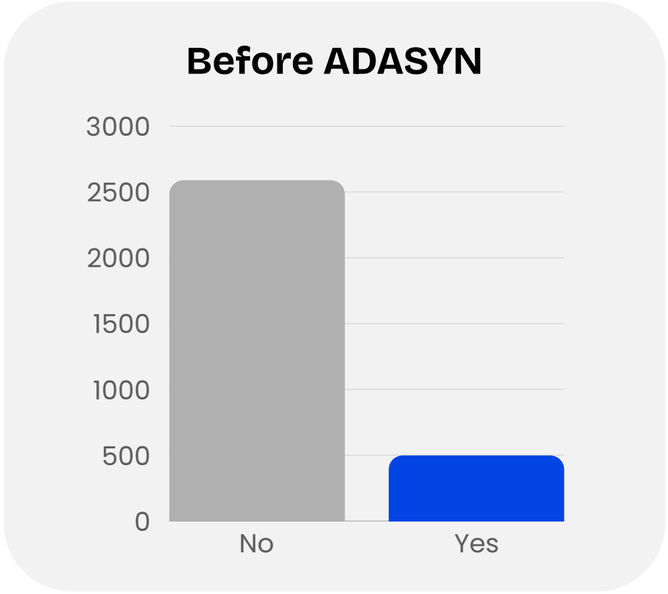
Addressing **class imbalance** in the target variable (y_train) using the **ADASYN** (Adaptive Synthetic Sampling) technique. This is applied only to the **training data** to prevent data leakage into the test set.

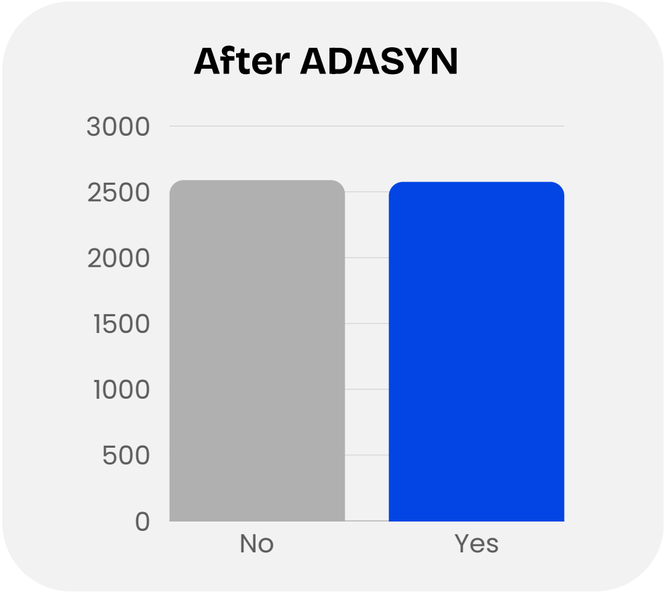
The **minority class** is **scattered** among the majority (source)

The decision boundary is **complex** and **non-linear** (source)

ADASYN **outperformed** other resampling methods in our tests







FEATURE EXTRACTION & ENCODING

3SIGMA SQUAD

AvgWorkHours

Average number of working hours per day (Al-suraihi, 2021)

Date-based

AbsentDays

Total number of working days the employee was absent (Morrow et al., 1999)

Date-based

OvertimeDays

Number of days the employee worked more than standard hours (Al-suraihi, 2021)

Date-based

Feature Extraction ↑

Feature Encoding ↓

ML models require numerical input. Encoding converts categorical data into a machine-readable format.

Nominal Data

For categories like marital status, departments, or job role where there's **no inherent rank or order**, we used:

One-Hot Encoding

Ordinal Data

For categories that do have a **clear order or hierarchy** like job satistifaction, job level, or work-life-balance, we used:

Ordinal Encoding

FEATURE SELECTION

3SIGMA SQUAD

Feature Removal

- **EmployeeID**
  Unique identifier, no predictive value
- **EmployeeCount**
  Constant value (1) across all records
- **StandardHours**
  Fixed at 8 hours for every employee
- **Over18**
  All employees are over 18 (all values Yes)

Multicollinearity Note

- **High VIF values** observed in some **features**
- **Removal** of high VIF features tested, resulted in **lower model performance**
- **Engineered features** from highly correlated variables **underperformed** compared to original feature configuration

The cleaned dataset with 42 columns split into train and test sets available for download here

3SIGMA SQUAD

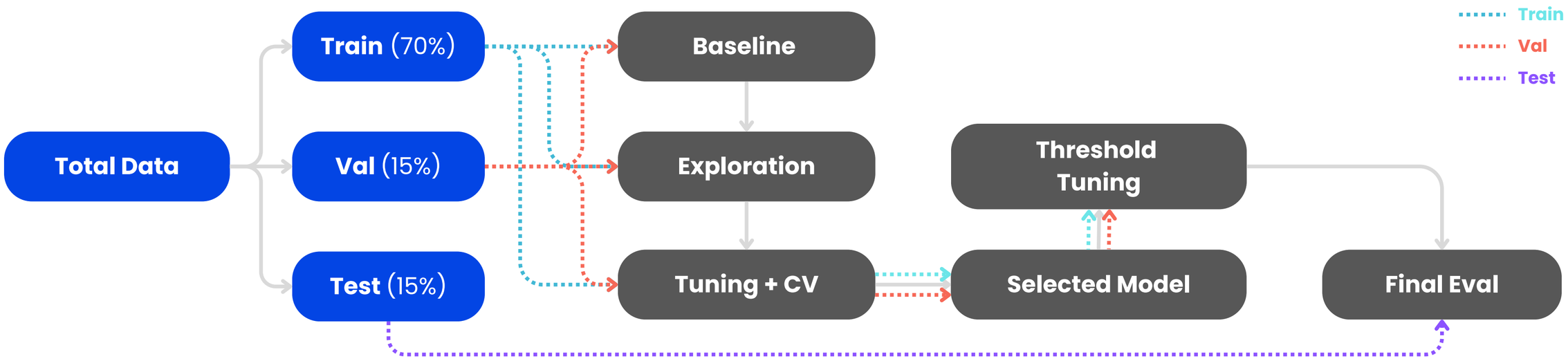
Rakamin Academy

MODELING & INTERPRETABILITY

Baseline Modeling • Model Exploration • Hyperparameter Tuning • Fairness • Feature Importance • Error Analysis • Impact Assessment

MODEL EXPLORATION

3SIGMA SQUAD



Initial Model Exploration

- **Logistic Regression** baseline
- Tested **various models**: linear, kernel, instance, tree, boosting, neural nets
- **Top 6** models chosen for tuning

Refinement & Tuning

- Performed **grid search** for **hyperparameter tuning**
- Used **stratified 5-fold CV** for balanced class distribution
- Optimized for **F2-Score** (prioritizing recall)
- Selected the **best model** as the final choice

PRIMARY METRIC

3SIGMA SQUAD

Our primary metric is the **F2 Score**. This calculation is designed to treat the error of **missing an at-risk employee** as **twice** as significant as the error of **incorrectly flagging a secure one**.

Primary Objective

- Detect all at-risk employees
- Missing risks means losing key talent
- Maximize identification efforts
- Act early to prevent turnover

Efficiency Focus

- Unnecessary actions waste resources
- May impact employee experience
- Focus on targeted, efficient response
- Balance accuracy with efficiency

Missing At-Risk Employee

$35,000

www.peoplekeep.com

Unnecessary Intervention

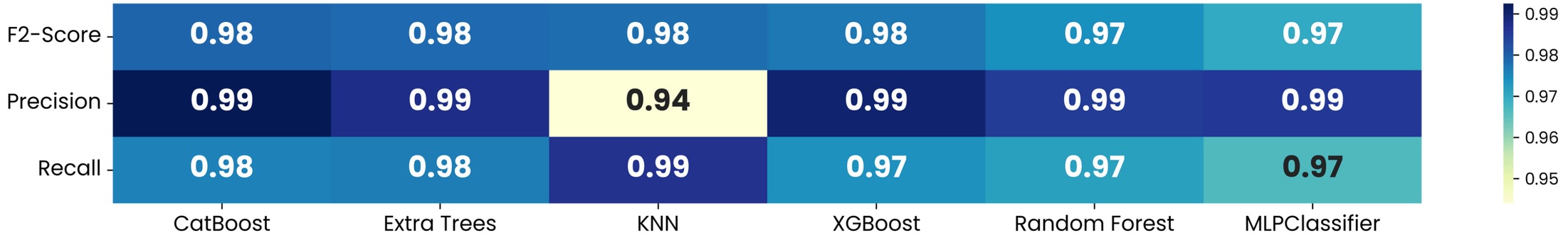
$5,000

www.aihr.com

MODEL COMPARISON

3SIGMA SQUAD

Initial Exploration Results



Baseline

Logistic Regression

F2 Score on train set

0.8076

F2 Score on val set

0.4826

After Hyperparameter Tuning

Model	Best Param	F2 Mean
CatBoost	depth=8, iter=200, lr=0.2, l2=1	0.9806
KNN	n=11, p=1, weights=distance, metric=minkowski	0.9757
MLPClassifier	hl=(100,), act=tanh, alpha=0.001, lr=const, solver=adam	0.9685
Extra Trees	crit=entropy, max_feat=log2, n_est=300	0.9797
XGBoost	max_depth=6, lr=0.2, n_est=300, subsample=0.8	0.9805
Random Forest	crit=entropy, max_feat=sqrt, n_est=300	0.9750

FINAL MODEL

3SIGMA SQUAD

CatBoost

0.0057s

Fast prediction time makes the model ideal for practical implementation

0.9586

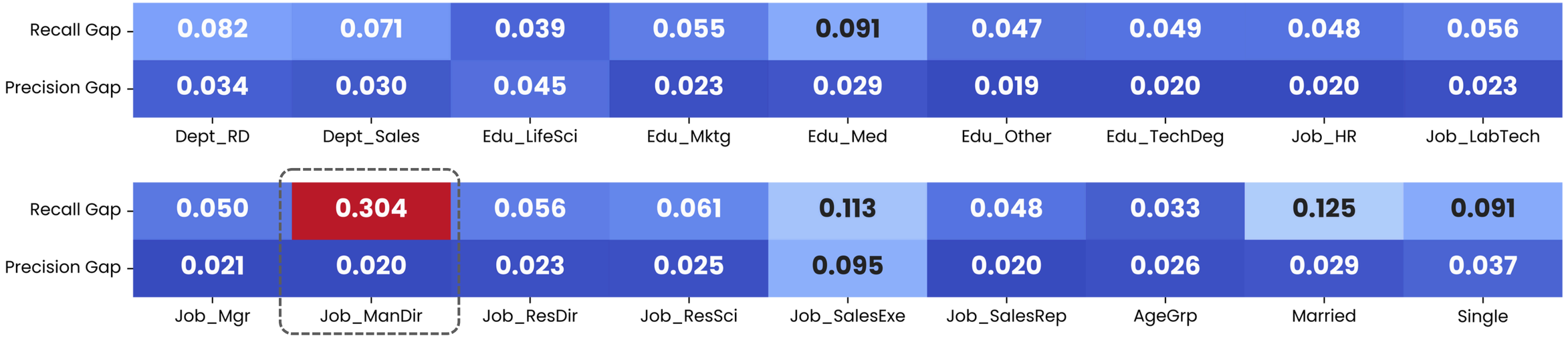
Achieved excellent test performance, indicating strong generalization

```
best_params = {
    'depth':8,
    'iterations':200,
    'l2_leaf_reg':1,
    'learning_rate':0.2,
    'verbose':False
}
```

- More Reliable Generalization
- Superior Prediction Speed
- Automated Feature Engineering
- Individual Prediction Explanations
- Overfitting Resistance

FAIRNESS ANALYSIS RESULTS

3SIGMA SQUAD



Bias Detection

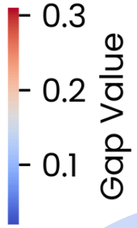
All segments perform **well** — except **Manufacturing Director** with **recall gap > 0.20**

Why It Matters

Manufacturing Directors are **high-impact roles**. Missing true attrition (false negatives) is costly, while false alarms are acceptable.

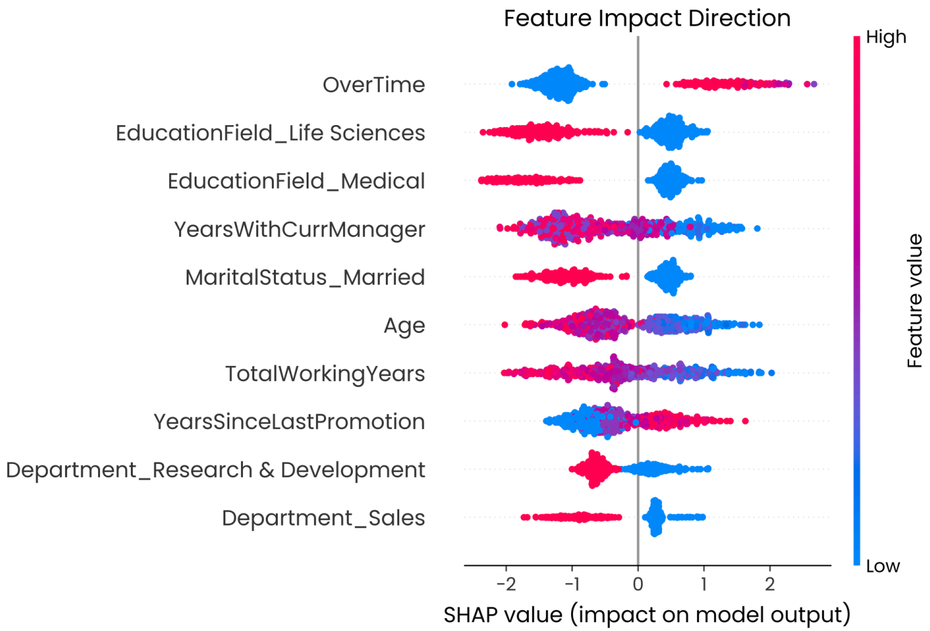
Solution

Request **additional data** for Manufacturing Director segment to improve model fairness and reduce bias gap.



MODEL EXPLAINABILITY ANALYSIS

3SIGMA SQUAD



Key Risk Drivers

A **high-risk** profile is identified in employees who frequently work **overtime**, are **younger**, **unmarried**, and have a **short tenure** with their **current manager**

Key Mitigating Factors

A **lower-risk** profile is associated with employees who have **longer tenures** and an educational background in **Medical** or **Life Sciences**

SHAP-Based Feature Pruning

Based on SHAP analysis, we removed low-impact features — excluding one-hot encoded ones — to simplify the model without compromising structure or interpretability. Removed features: **JobInvolvement**, **JobLevel**, **AbsentDays**, **StockOptionLevel**, **Education**, **PerformanceRating**.

Final Model Evaluation

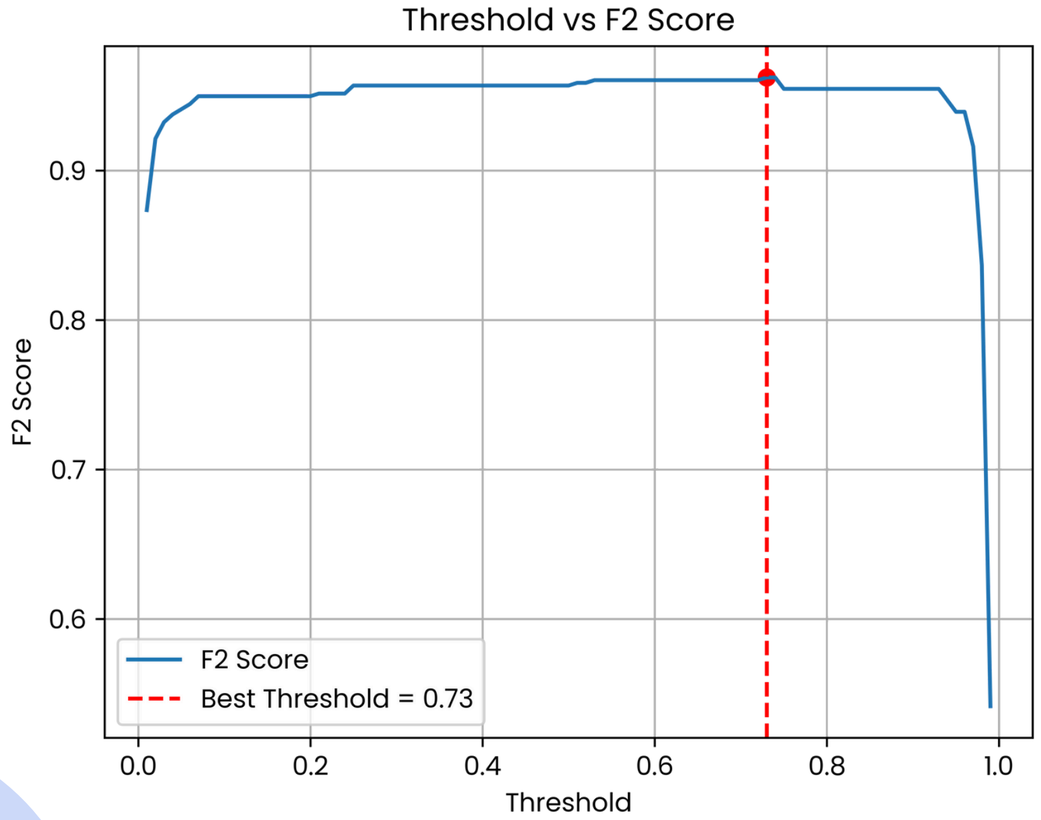
Model performance remains unchanged across all evaluation metrics:

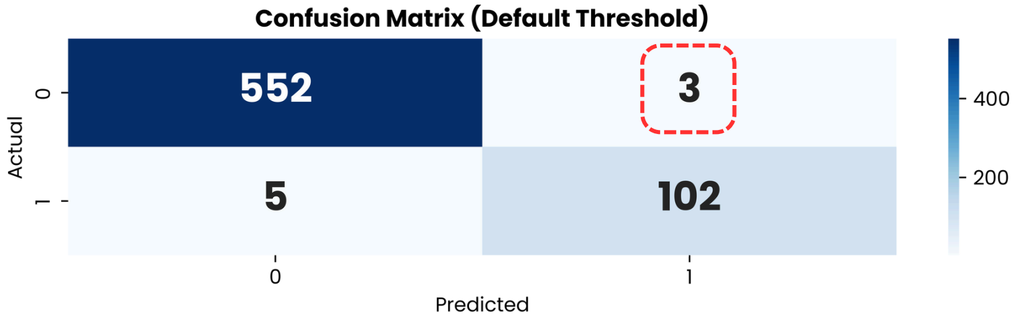
- **F2-Score: 0.9568**
- **Precision: 0.9714**
- **Recall: 0.9533**

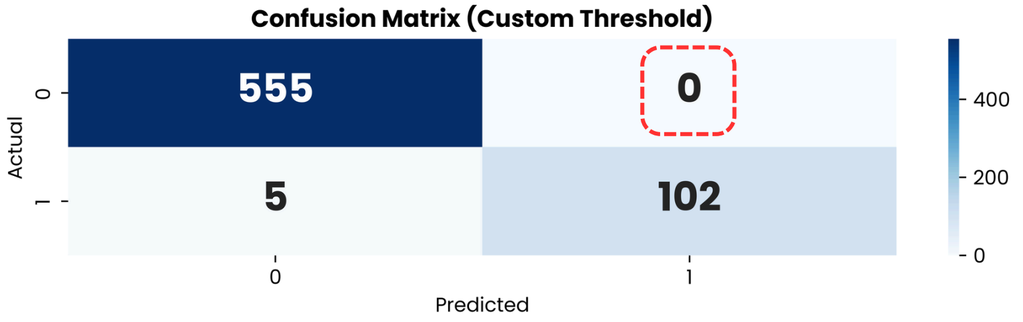
THRESHOLD OPTIMIZATION

3SIGMA SQUAD

We optimized the threshold to **maximize F2-score**, prioritizing recall. The **optimal threshold of 0.73** effectively balanced precision and recall, **eliminating all 3 false positives**.







INDUSTRY & FINANCIAL IMPACT

3SIGMA SQUAD

Industry Impact

- **Integration** with existing **HR tech stack**
- **Data-driven retention** strategies adoption
- Predictive HR becomes **standard practice**

Key Talent Attrition Impact

- **50%** HR success rate
- **147** expected resignations
- **65** resignations prevented

44.2%

key talent attrition drop

Financial Impact

$39.3M

Without Model

- 710 employees resigned = **$39.3M**

$24.9M

With Model

- Intervention costs: **$3.4M**
- Model monitoring & maintenance: **$50K**
- Team operational costs: **$810K**
- 373 employees resigned: **$20.7M**

$14.4M

annual savings

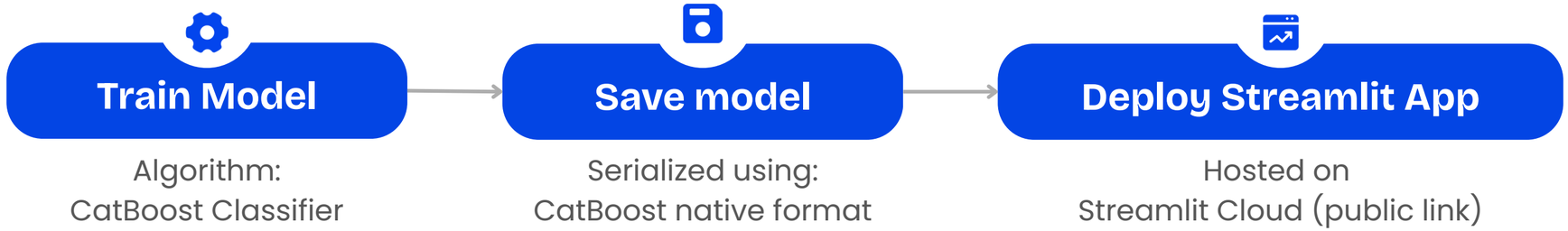
Achieved through $14.4M strategic investment in predictive modeling and targeted retention programs — delivering 340.85% ROI across 4,410 employees

Full cost analysis available at: [click here]

DEPLOYMENT & RECOMMENDATION

3SIGMA SQUAD

ML Deployment Architecture



Strategic Recommendation

Company-Level Interventions

A macro-level, data-driven strategy to address attrition risks across the workforce proactively

Individual-Level Interventions

ML flags high-risk employees, while SHAP reveals root causes—guiding HR in focused, confidential interventions

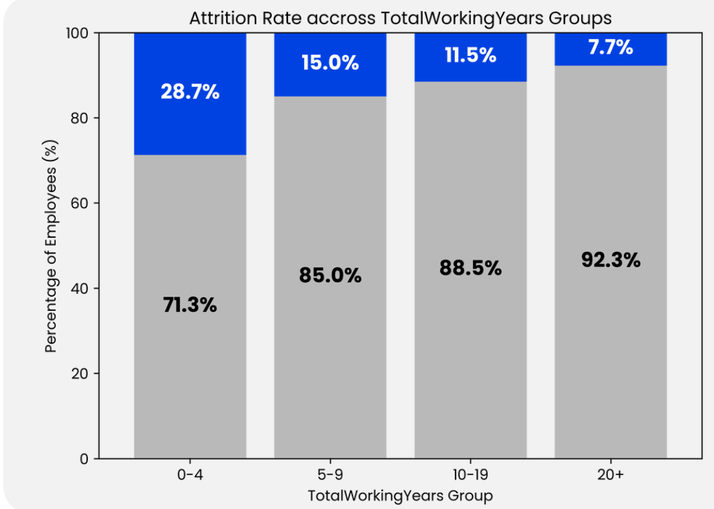
COMPANY-LEVEL INTERVENTIONS

3SIGMA SQUAD

"According to SRT, receiving socialization resources facilitates employees' adjustment to the new environment, with subsequent direct and indirect outcomes for employees' overall attitudes, an example being turnover intention."

Strategic Onboarding & Integration

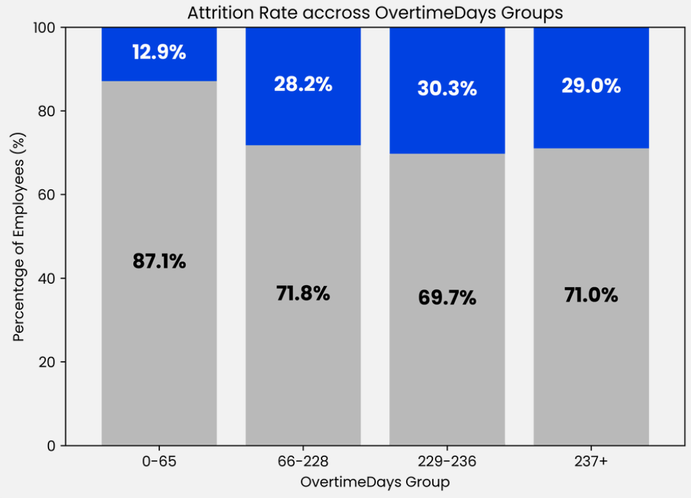
pmc.ncbi.nlm.nih.gov



Mentorship & Coaching

#5 Attrition Driver
(per SHAP Analysis, Slide 43)

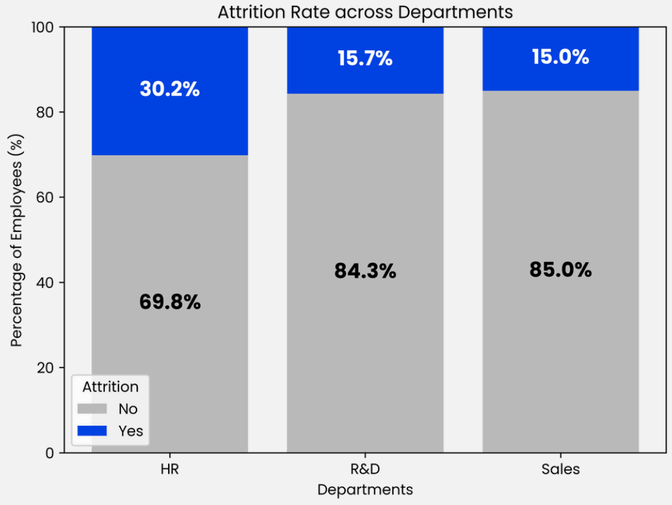
CNBC/SurveyMonkey



Overtime & Workload Management

#1 Attrition Driver
(per SHAP Analysis, Slide 43)

pmc.ncbi.nlm.nih.gov



HR Departmental Review

Statistically Significant
(Chi-Square Test)

INDIVIDUAL-LEVEL INTERVENTIONS

3SIGMA SQUAD

Our **SHAP analysis** identifies the **top three drivers** of **individual employee attrition**, providing crucial insights for HR. These findings act as a **catalyst for focused 1-on-1 discussions**, guiding HR to pinpoint potential issues. While SHAP highlights key areas, the specific interventions will be co-created and customized based on the outcomes of these individual conversations, ensuring truly effective and personalized support for retention.

Work-Life Balance

Implement flexible work policies

EY 2022 Survey

Career Development

Development and training programs

LinkedIn for Learning

Financial & Compensation

Offer competitive salaries

gallup.com

Work Environment

Conduct stay interviews

hr.nih.gov

Roles & Positions

Strategic job rotation

Study by SHRM

Access the Live Dashboard & Prediction App: [click here]

CHALLENGES & LIMITATIONS

3SIGMA SQUAD

These are the **main problems** that **blocked project progress**, **reduced model effectiveness**, and **created workflow inefficiencies**. Each issue required a specific solution to keep the project on track:

Imbalanced Dataset Challenge

Apply ADASYN oversampling to balance the class distribution and ensure better model performance overall

Accuracy Metrics Misalignment

Use the F2-Score to optimize the model so that results stay relevant to core business objectives and needs

Black Box Nature of CatBoost

Leverage SHAP to interpret CatBoost results, increase transparency, and build stakeholder confidence fully

Project Team Stagnation Issue

Restructure and realign team roles and workflows to resolve dysfunction and accelerate project delivery

CONCLUSION

3SIGMA SQUAD

Key Drivers

- High workload
- Low satisfaction
- Younger age, limited experience

Area of Concern

The HR department's higher-than-average attrition rate should be investigated further to understand its root causes and develop appropriate solutions

Predictive Model

- Accurately flags ~95% of potential leavers
- Minimizes unnecessary false positives

Action Point

- Use the model to monitor risk proactively
- Pilot targeted interventions to boost retention

Recommendation

Add data for missing roles • Add relevant predictive features • Use survey insights • Monitor drift & retrain

Details on retraining & maintenance here

3SIGMA SQUAD

THANK YOU

- GitHub Repository: [click here]
- Streamlit App: [click here]
- Dataset: [click here]