EMPLOYEE ATTRITION

MMA7-GROUP 5

HANNAH WANG
RICHARD EL CHAAR
YURI XU
CLAUDIA NI

https://github.com/McGill-MMA-EnterpriseAnalytics/Employee-Attrition

TABLE OF CONTENTS

PROBLEM STATEMENT

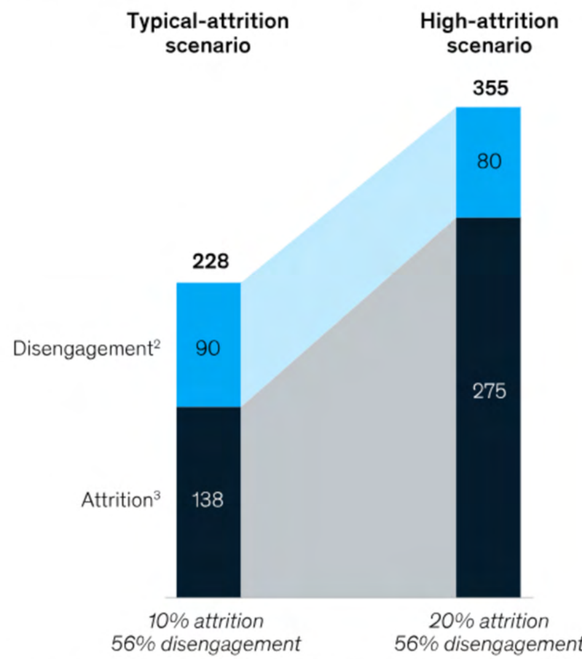
PROBLEM STATEMENT

Why Attrition Matters?

- Lost productivity
- Employee burnout
- Lost tribal knowledge
- Wasted time and money associated with hiring a replacement
- Cost of training a new person and risk of mis-hire

For a median-size S&P 500 company, the estimated cost of employee disengagement and attrition is $228 million a year—and can be much higher.

Annual cost split by disengagement and attrition,[1] $ million



https://www.forbes.com/sites/forbeseq/2023/03/21/five-hidden-costs-of-employee-attrition/

https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/some-employees-are-destroying-value-others-are-building-it-do-you-know-the-difference

PROBLEM STATEMENT

Hypotheses and Objectives

Attrition Risk Classification

Hypothesis:
We can predict individual employee attrition risk using historical HR data.

Action:
Proactively flag high-risk employees and prioritize them for targeted retention interventions.

Causal Inference & Treatment Effects

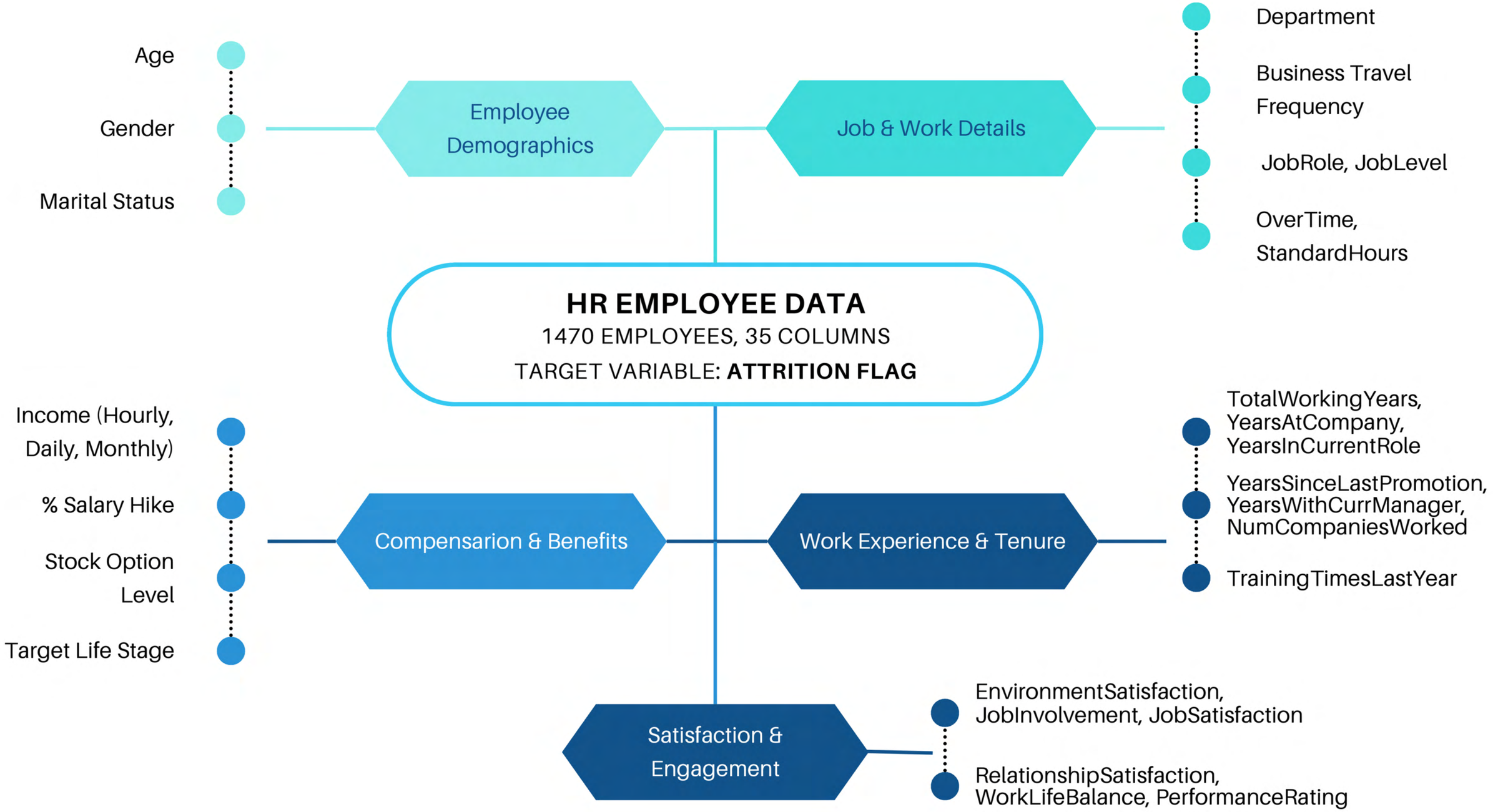
Hypothesis:
By applying causal inference methods to our employee data, we can identify treatment effects from specific interventions that would lead to a measurable reduction in attrition.

Action:
Allocate resources to interventions that demonstrate a meaningful impact on lowering churn.

Employee Segmentation

Hypothesis:
Clustering employees into distinct groups will reveal unique attrition drivers, enabling tailored retention strategies.

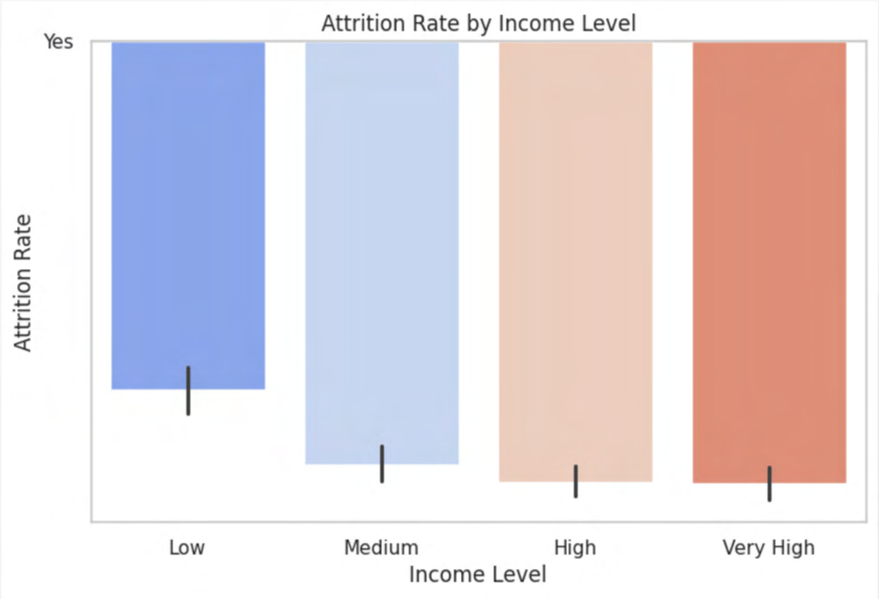
Action:
Customize retention programs based on the specific needs and risk profiles of each employee segment.

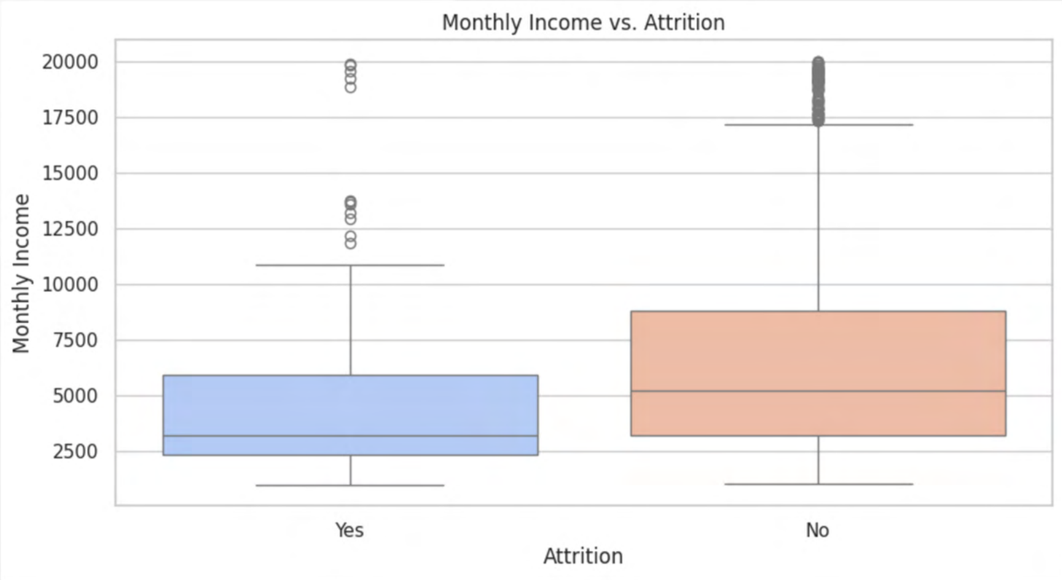

Age
Gender
Marital Status
Employee Demographics
Job & Work Details
Department
Business Travel Frequency
JobRole, JobLevel
OverTime, StandardHours
HR EMPLOYEE DATA
1470 EMPLOYEES, 35 COLUMNS
TARGET VARIABLE: ATTRITION FLAG
Income (Hourly, Daily, Monthly)
% Salary Hike
Stock Option Level
Target Life Stage
Compensarion & Benefits
Work Experience & Tenure
TotalWorkingYears, YearsAtCompany, YearsInCurrentRole
YearsSinceLastPromotion, YearsWithCurrManager, NumCompaniesWorked
TrainingTimesLastYear
Satisfaction & Engagement
EnvironmentSatisfaction, JobInvolvement, JobSatisfaction
RelationshipSatisfaction, WorkLifeBalance, PerformanceRating

EDA

EDA

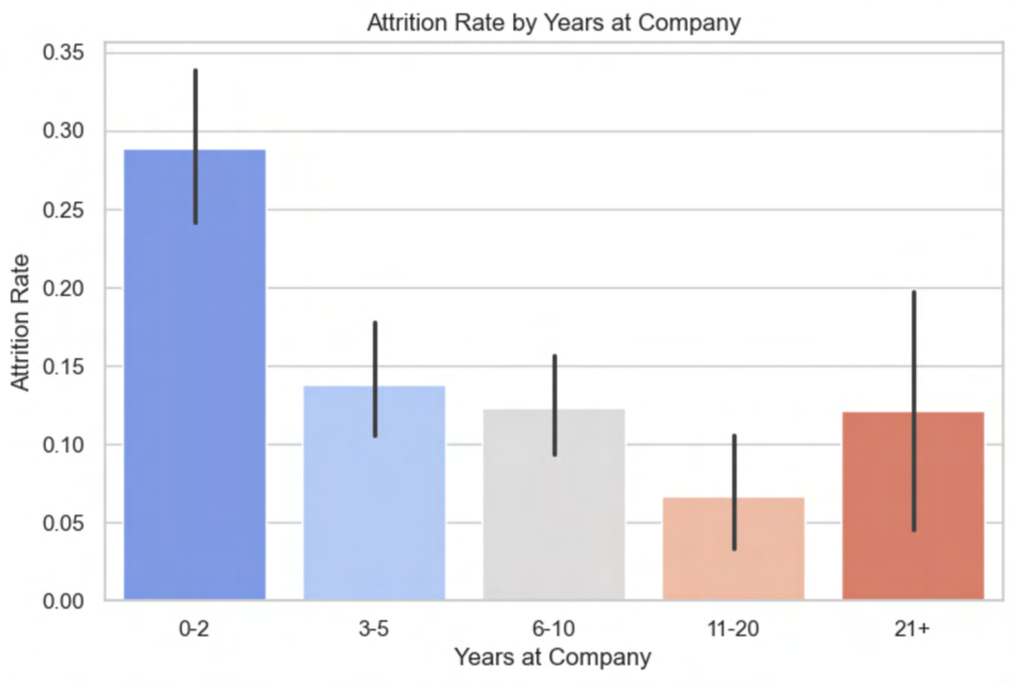
Attrition Tends to be Higher in Employees Earning Less, possibly seeking better opportunities

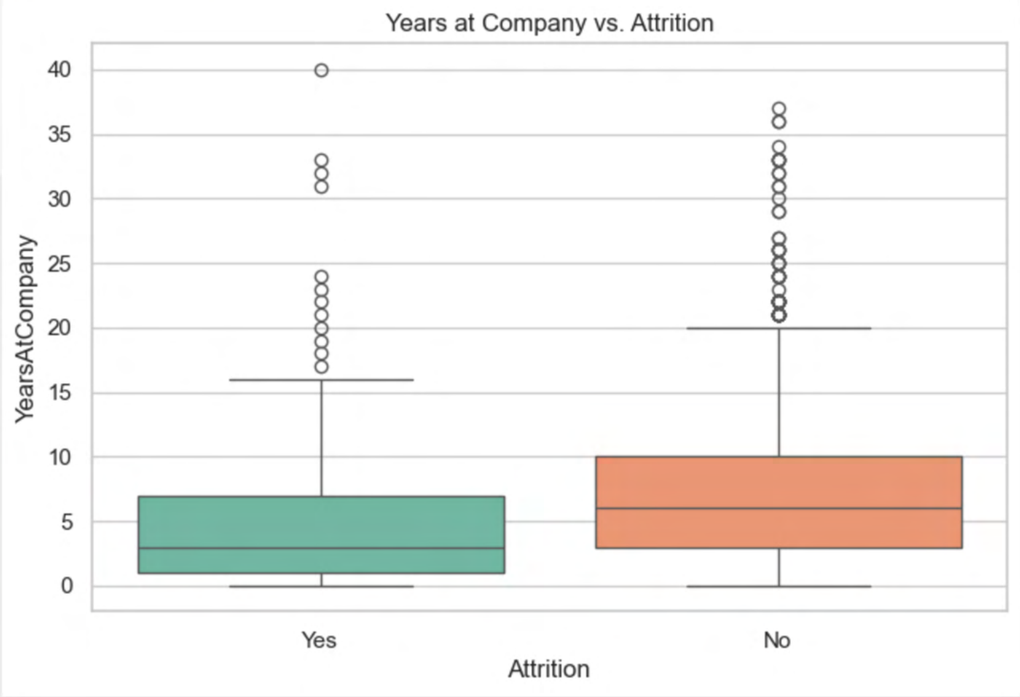




EDA

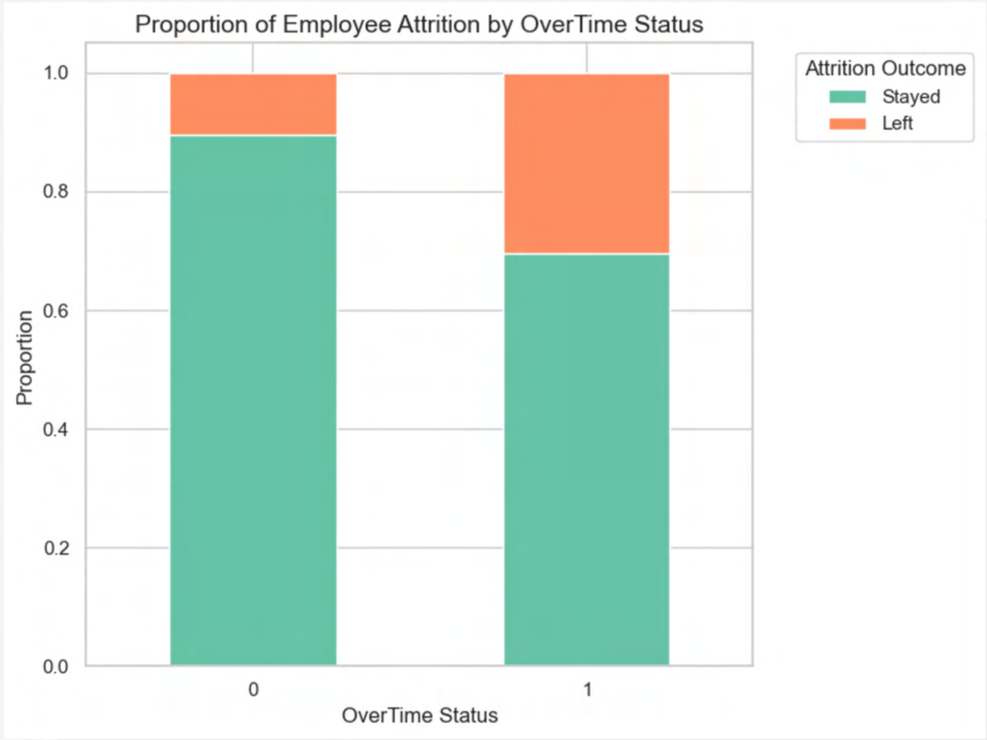
Company tenure matters:
Early-stage employee turnover is high due to poor job or culture fit.
Attrition steadily declines as tenure increases, except for a slight uptick at 21+ years, likely due to retirement or career stagnation.

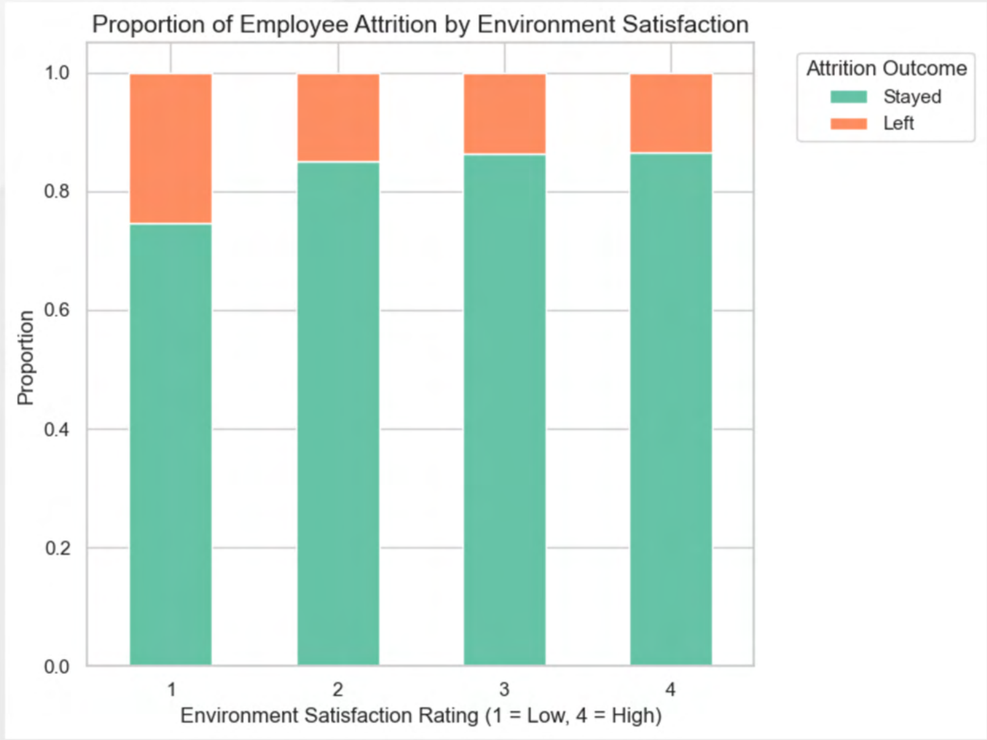




EDA

Burnout matters:
Employees working overtime have much higher chances of leaving



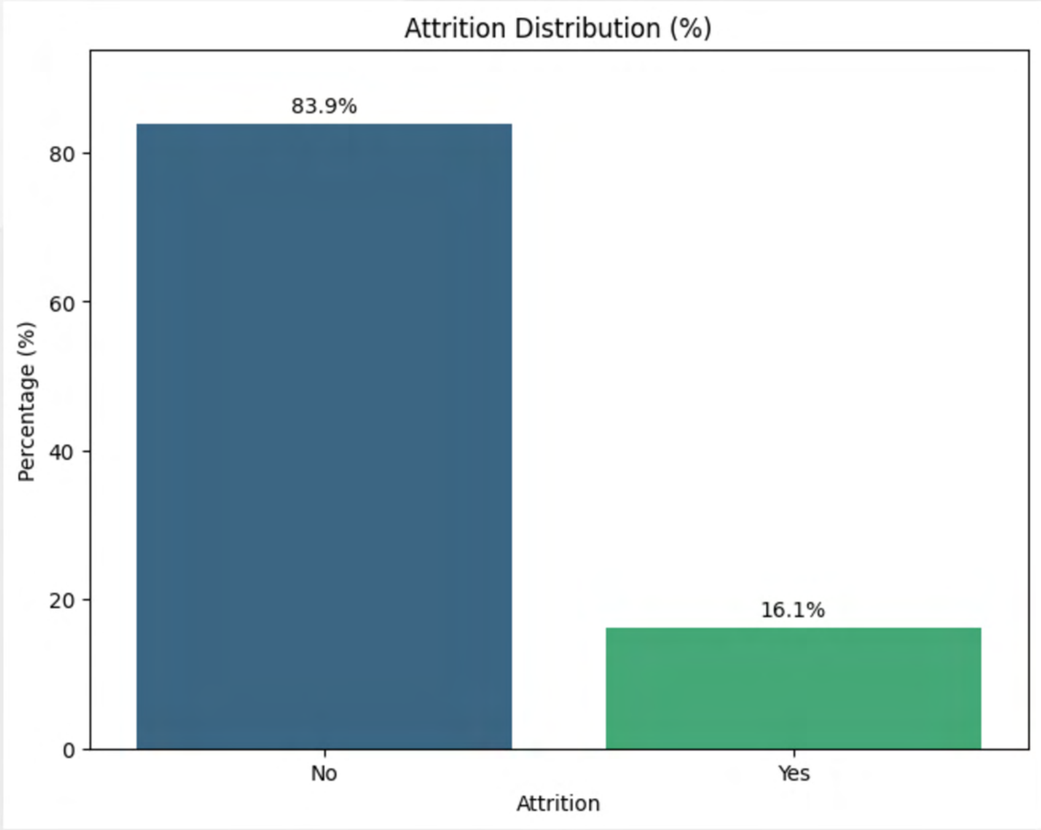


PRE-PROCESSING

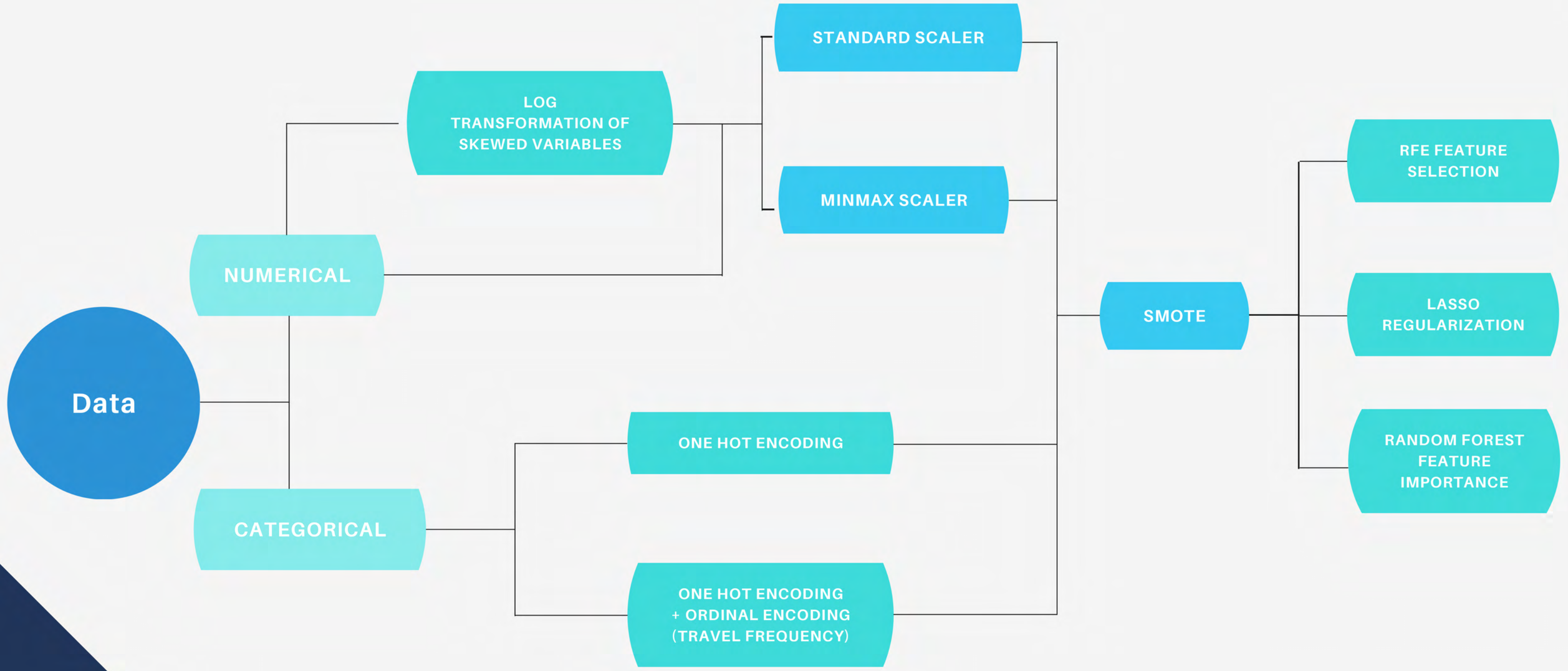
PRE-PROCESSING

Data Cleaning & Splitting

- Removed duplicates and non-informative columns (EmployeeCount, StandardHours, EmployeeNumber)
- Label encoded "Attrition"
- Performed a stratified train/test split



PREPROCESSING & FEATURE ENGINEERING PIPELINES



MODELLING

SEQUENTIAL FINE TUNING APPROACH

Classification Models

- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boost
- MLP

Hyperparameter Tuning

- Selected the best pipeline from each base classifier.
- Tuned model hyperparameters using **RandomizedSearch**.

Preprocessing & Feature Selection

Used **GridSearch** to identify the best combination of preprocessing steps for each base model.

Stacking Ensemble

- Picked the best-tuned version of each classifier.
- Built stacking ensembles to combine strengths of top models.

Evaluated our models using 5-fold cross-validation
With the main objective of maximizing the F2-score: we want to prioritize recall to ensure we identify as many high-risk employees as possible

PHASE 1 OPTIMAL PIPELINES

Model	Numeric Transformation	Numeric Scaler	Categorical Encoding	Feature Selection Method
LogisticRegression	Apply log1p transformation	MinMax Scaling	OneHot + Ordinal Encoding	Logistic Regression with L1 regularization (Lasso)
DecisionTree	Apply log1p transformation	MinMax Scaling	OneHot + Ordinal Encoding	RandomForest Feature Importance
RandomForest	No transformation (passthrough)	MinMax Scaling	OneHot + Ordinal Encoding	Recursive Feature Elimination (RFE)
GradientBoosting	Apply log1p transformation	MinMax Scaling	OneHot encoding	Logistic Regression with L1 regularization (Lasso)
MLP	Apply log1p transformation	Standard Scaling	OneHot + Ordinal Encoding	No feature selection (passthrough)

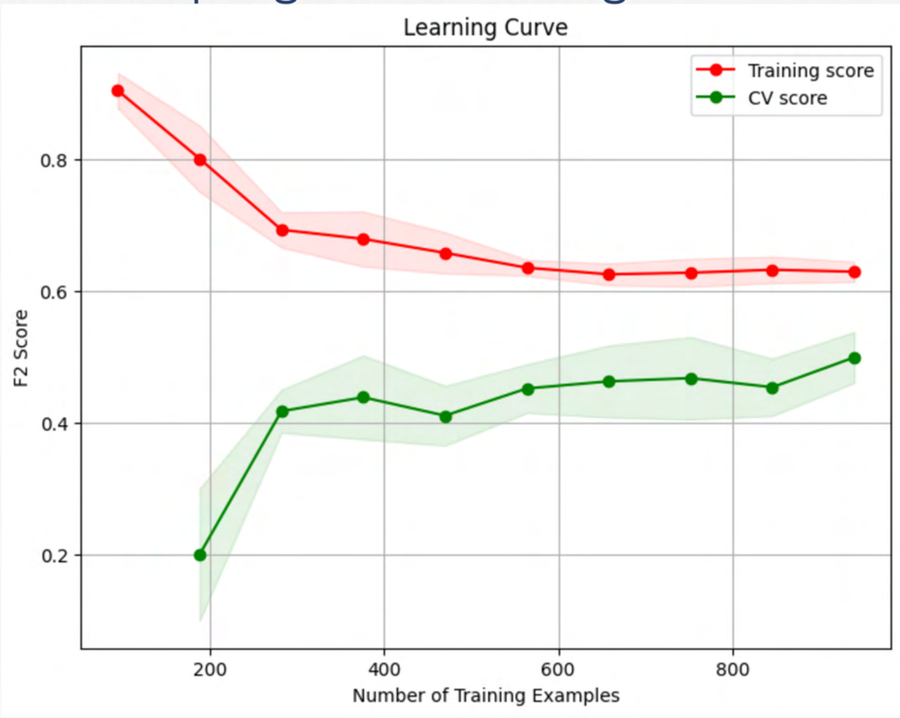
PERFORMANCE EVALUATION

	Base Model		Fine Tuned Model		Train Set
Model	**ROC-AUC**	**F2 Score**	**ROC-AUC**	**F2 Score**	**F2 Score**
LogisticRegression	0.8369	0.6317	0.8368	0.6319	0.6667
DecisionTree	0.6245	0.3998	0.6632	0.4024	0.7432
RandomForest	0.7281	0.3839	0.7519	0.4807	0.5504
GradientBoosting	0.7994	0.462	0.7947	0.4874	0.6112
MLP	0.8031	0.5472	0.8222	0.5577	1
Stacked Model			0.8337	0.5202	0.7542

Clear signs of overfitting!

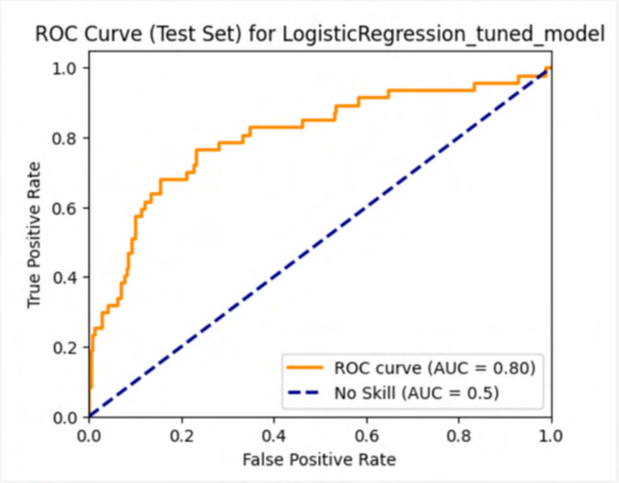
Gradient Boosting Learning Curve

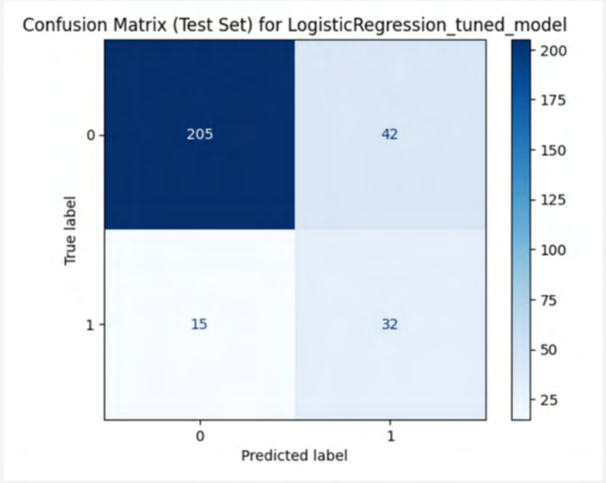
More data is needed for complex models, oversampling was not enough

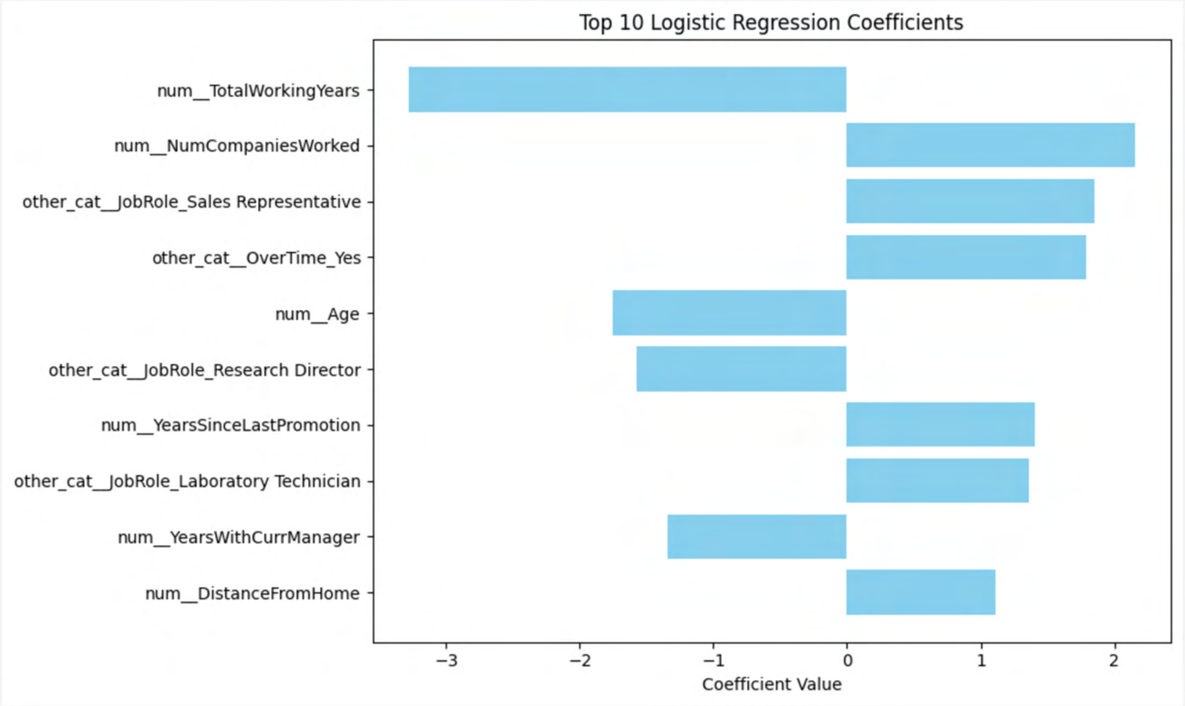


BEST MODEL EVALUATION ON TEST SET : LOGISTIC REGRESSION

Precision (Class 1)	Recall (Class 1)	F2 Score	Accuracy	ROC-AUC
0.43	0.68	0.61	0.8	0.8







CAUSAL INFERENCE

Treatment:

1. Overtime
2. Job Involvement
3. Many Companies Worked
4. Work-Life Balance
5. Long Time Since Promotion
6. Above Median Income
7. Relationship Satisfaction

Methodology:

- DoWhy
  - Directed Acyclic Graph
  - Treatment Variable Assumption
  - Cofounders Rational
- Heterogenity Analysis
- Refutation Tests

Treatment	Average Treatment Effect	Significant	Confidence Interval
Overtime	+ 21.63%	Yes	[0.172, 0.262]
Job Involvement	-7.82%	Yes	[-0.122, -0.035]
Company Worked >=3	+ 6.18%	Yes	[0.020, 0.104]
Work-Life Balance	-4.9%	Yes	[-0.088, -0.006]
Long Time Since Promotion	+ 3.66%	No	[-0.013, 0.086]
Above Median Income	-3.52%	No	[-0.079, 0.004]
Relationship Satisfaction	-2.25%	No	[-0.059, 0.014]

IMPLICATIONS

- **Overtime**

  - Sales Representatives (+37.85 pp) and Laboratory Technicians (+34.26 pp)
  - Youngest employees (18-30): +32.62 pp

- **Job Involvement**

  - Strongest effect in Human Resources (-16.05 pp)

- **Targeted Interventions**

  - Younger employees (18-30) are most sensitive to overtime, income differences, and previous job mobility
  - Relationship satisfaction matters most for single employees and those in the 31-40 age group

CLUSTERING

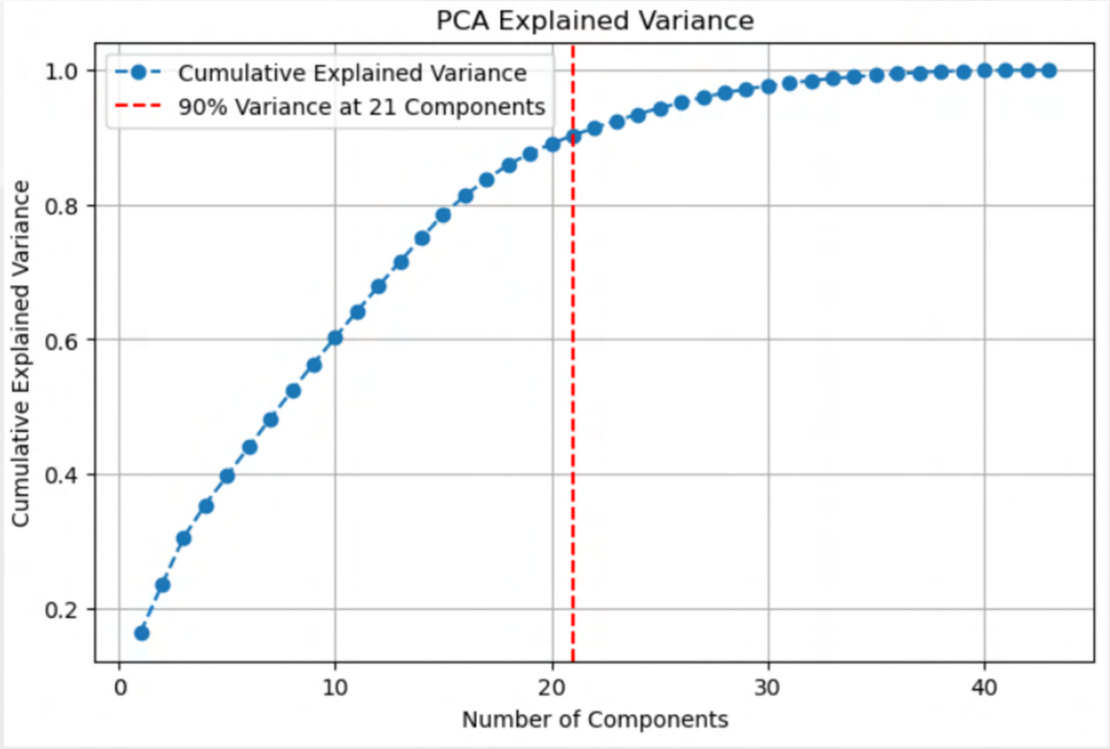
Objective

Applying Clustering Techniques to segment employees and explore how attrition patterns emerge across clusters, revealing insights into employee retention risks and uncovering opportunities for targeted HR strategies to improve retention.

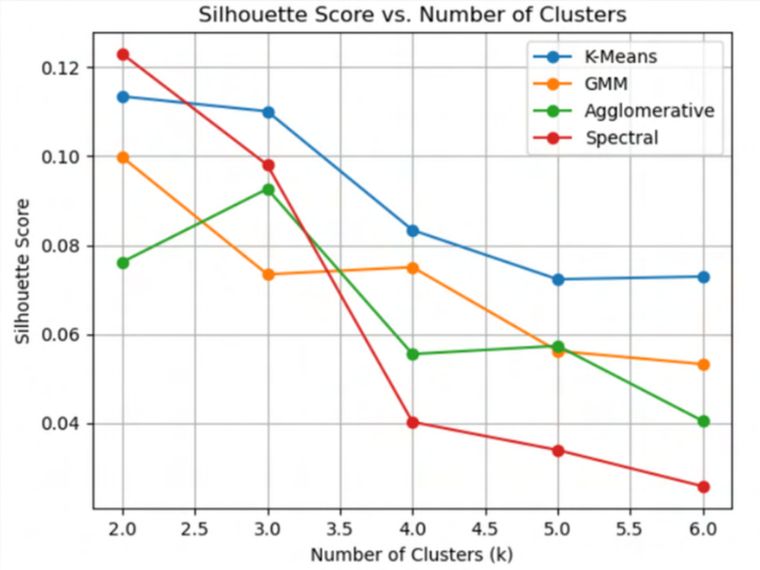
CLUSTERING

Pre-Process:

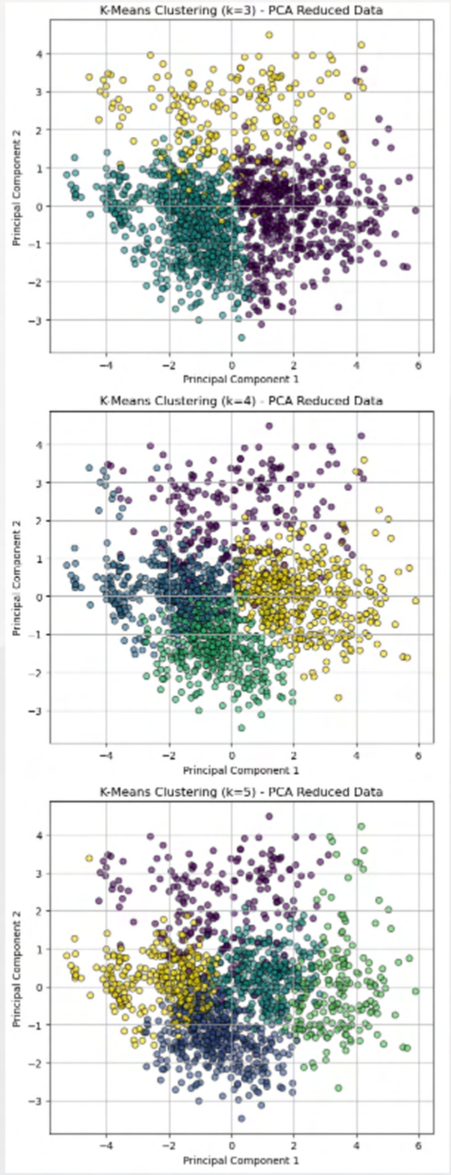
- Drop Attrition
- Log Transformation
- Standardization
- One-Hot Encoding
- PCA



CLUSTERING



- The highest **silhouette score** is observed at k=2, but possibly due to continuous dataset, the method of bisection might always work well



- **Cluster Separation:**
  - **k=3:** relatively **clear boundaries** and **stable structure**
  - **k=4 or 5:** clusters more **dispersed**, significant **overlap** occurs between multiple clusters.
- **Cluster Compactness:**
  - **k=3:** data points distributed relatively **tight**
  - **k=4 or 5: loosely distributed** without a clear dense center

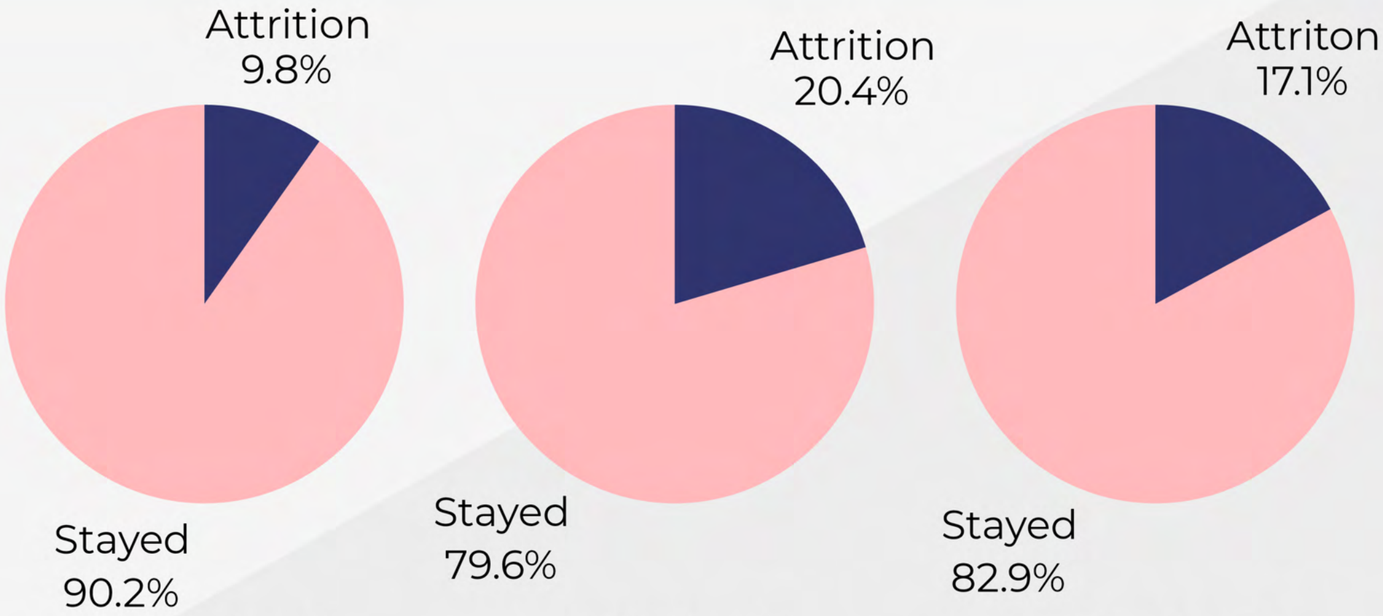
Final Decision:

KMeans with k=3

CLUSTERING

	Cluster	counts
0	0	521
1	1	732
2	2	217

- Each cluster has balanced size, indicating the segmentation is well-distributed across the employee population.



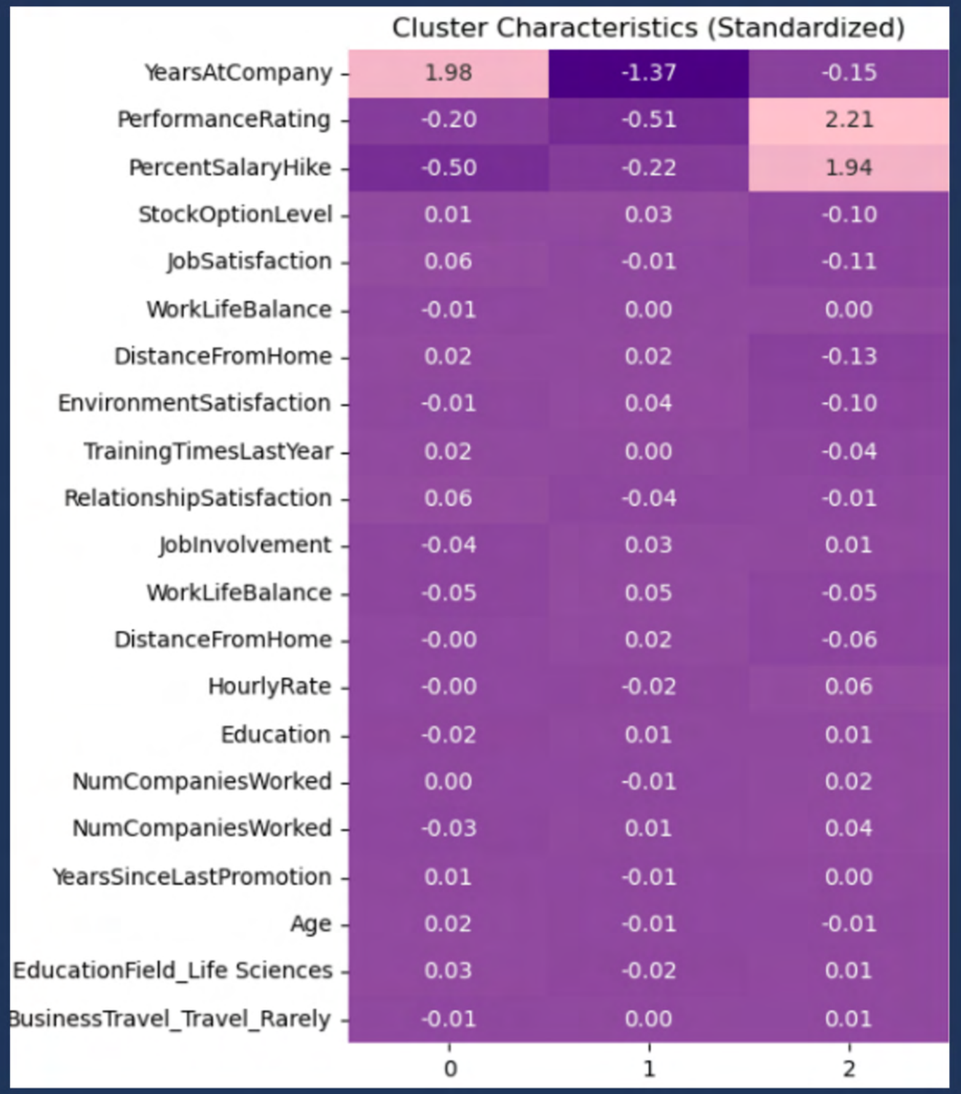
Cluster 1 **Cluster 2** **Cluster 3**

- Attrition rates differ significantly among different clusters
- Highlights its importance as a distinguishing factor

	Feature	P-Value
0	Attrition	0.000003

- P-value below 0.05 indicates a significant relationship with Attrition (Kruskal-Wallis Test).

CHARACTERISTICS & STRATEGY



- **Cluster 0:** Lowest attrition rate (9.78%)
  - **YearsAtCompany (1.98):** Long-tenured employees
  - **PercentSalaryHike (-0.50):** Below average salary increase
  - **PerformanceRating (-0.20):** Performance ratings lower than other clusters.

  Business Strategy: Focus on **career development programs**, offer skill-building opportunities

- **Cluster 1:** Highest attrition rate (20.35%)
  - **YearsAtCompany (-1.37):** Shortest tenure in the company.
  - **PerformanceRating (-0.51):**Lowest performance ratings.
  - **PercentSalaryHike (-0.22):** Salary increases below average.

  Business Strategy: Offer **mentorship programs**, **career coaching**, and **job rotation opportunities**

- **Cluster 2:** Medium attrition rate (17.05%)
  - **YearsAtCompany (-0.15):** Shorter tenure than average
  - **PerformanceRating (2.21):** Highest performance ratings
  - **PercentSalaryHike (1.94):** Largest salary increases.

  Business Strategy: Go beyond financial incentives, offer **leadership development programs**, **challenging assignments, structured career progression plans .**

CONCLUSION

- Predictive modeling
  - Identified key drivers
- Causal inference
  - Quantified their impact
- Segmentation:
  - Tailor retention strategies

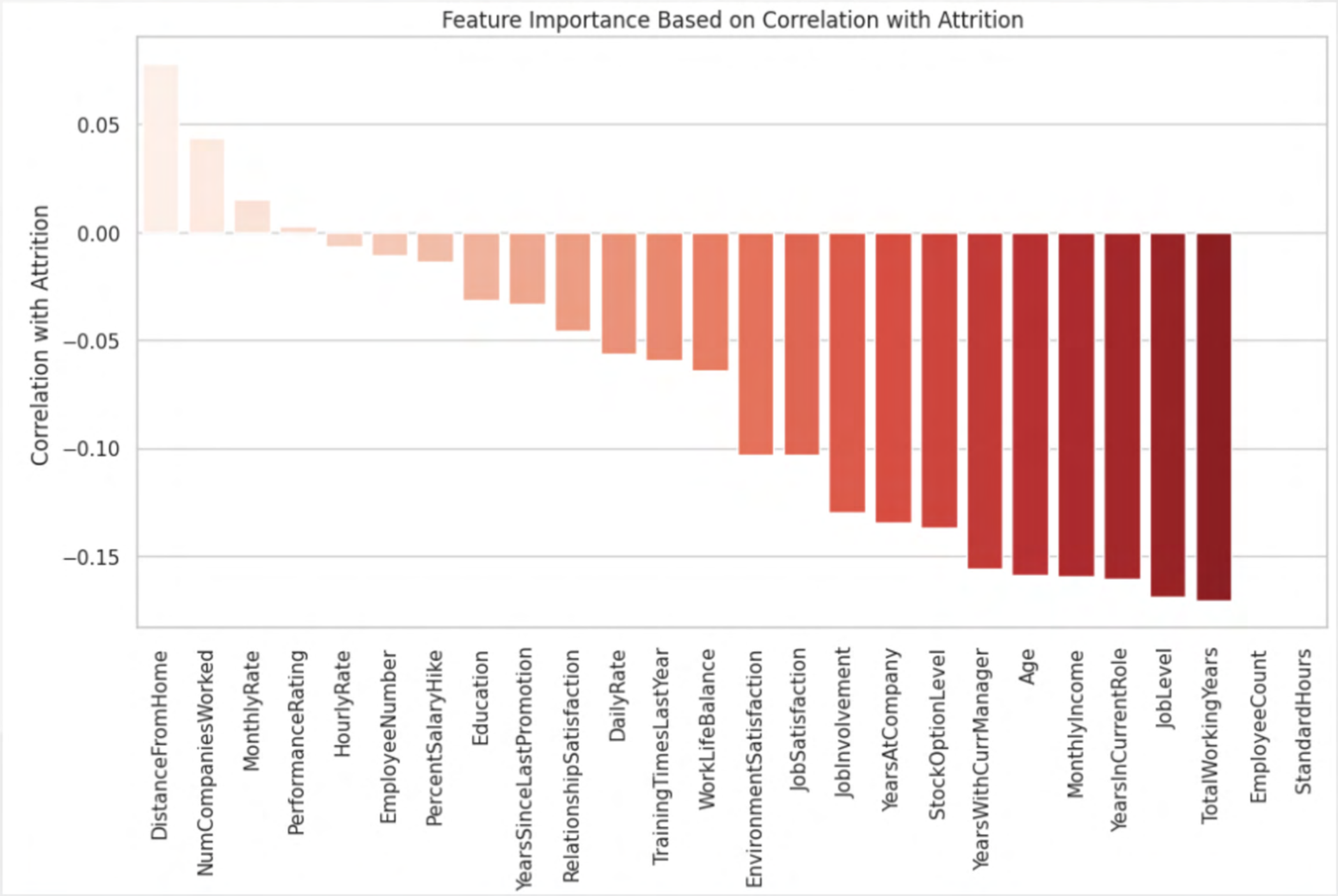
NEXT STEPS

- Implement and test HR interventions
- Monitor their effectiveness over time
- Refine models with real-time data for improvement

THANK YOU

APPENDIX I
EDA

- Work-Life Balance & Job Satisfaction Matter
- Performance Rating & Satisfaction Metrics are Weak Predictors



APPENDIX II. CLUSTERING