ADSP 31017 IP03 Machine Learning I

ANALYZING EMPLOYEE ATTRITION

Anusha Bhat, Devi Mahajan, Mahima Masetty, Nidhi Pareddy

03/11/25

CONTENT

BUSINESS PROBLEM

IDENTIFYING EMPLOYEES AT RISK OF ATTRITION

OUR DATA

This IBM dataset is comprised of employee attrition data, where we are given information related to the employee's demographic, work and academic backgrounds, and performance and satisfaction at work.

PROBLEM STATEMENT

IBM is experiencing employee attrition. To present the Chief HR Officer with retention strategies to ultimately reduce attrition, we must predict which employee profiles are most likely to contribute to higher attrition rates.

OUR APPROACH

We aim to implement a combination of supervised and unsupervised models to cluster and identify at-risk employees, and thereby determine a strategy to improve employee retention.

DATA SUMMARY AND EDA

DATASET OVERVIEW AND PRE-PROCESSING

01 **Data Overview:**

There are 1470 rows in our data, with 34 predictors. In addition, we have our target “Attrition” column we plan to predict. Out of this, we do not have any null values in our dataset.

02 **Erroneous Data:**

We dropped Employee Count, Over 18, and Standard Hours as they contain only 1 value. We also dropped Employee Number as this column contains unique employee ID’s.

03 **Correlated Data:**

We noticed high correlation between certain predictors, such as Monthly Income and Job level with the correlation level of 95%. We therefore decided to drop 3 columns to reduce correlation.

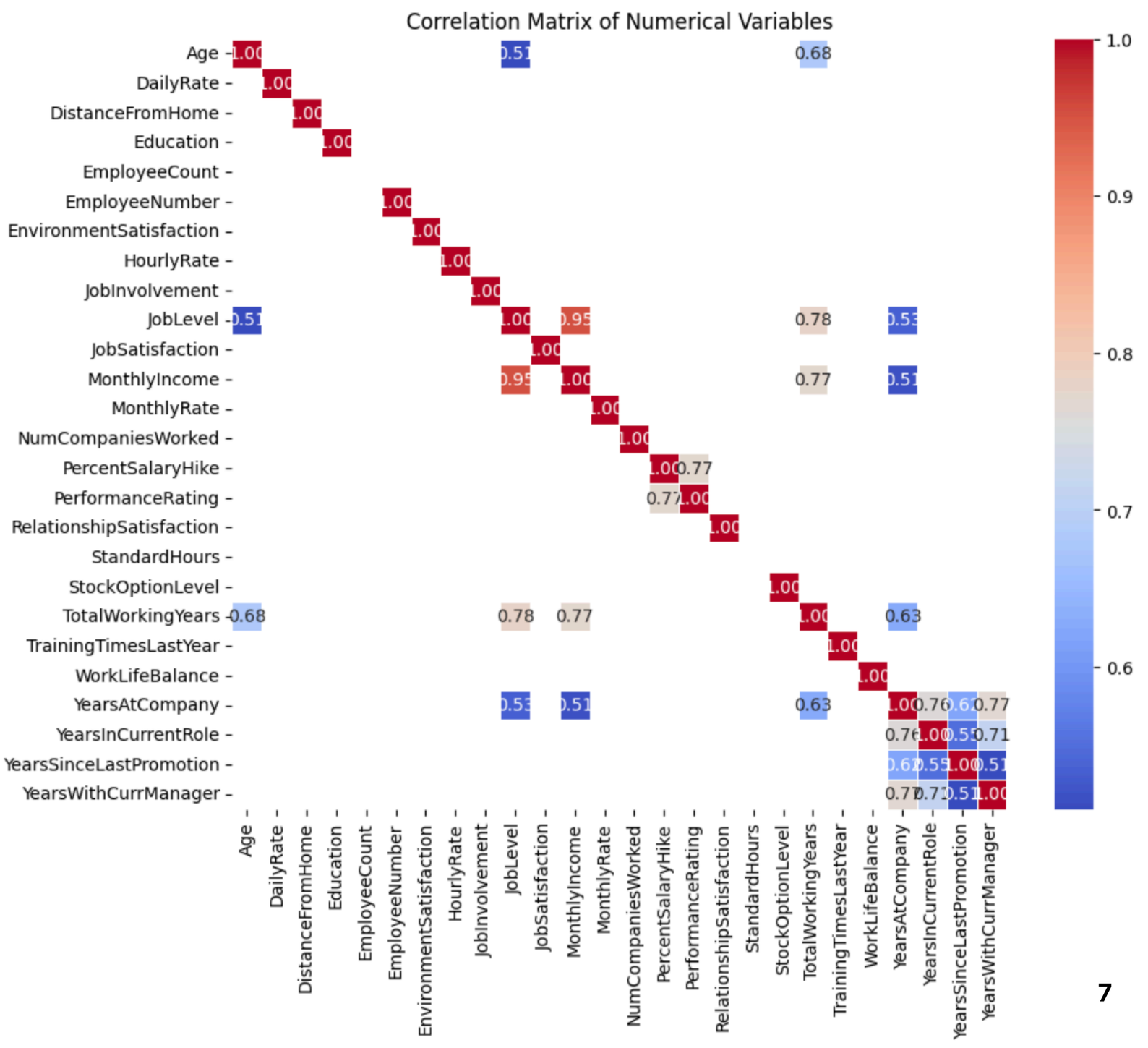
04 **Feature Engineering :**

Binary categorical columns originally stored as int64 data types were converted to object for one-hot encoding. The data was then scaled using standard or min-max scaling, based on the model.

EXPLORING CORRELATIONS BETWEEN PREDICTORS

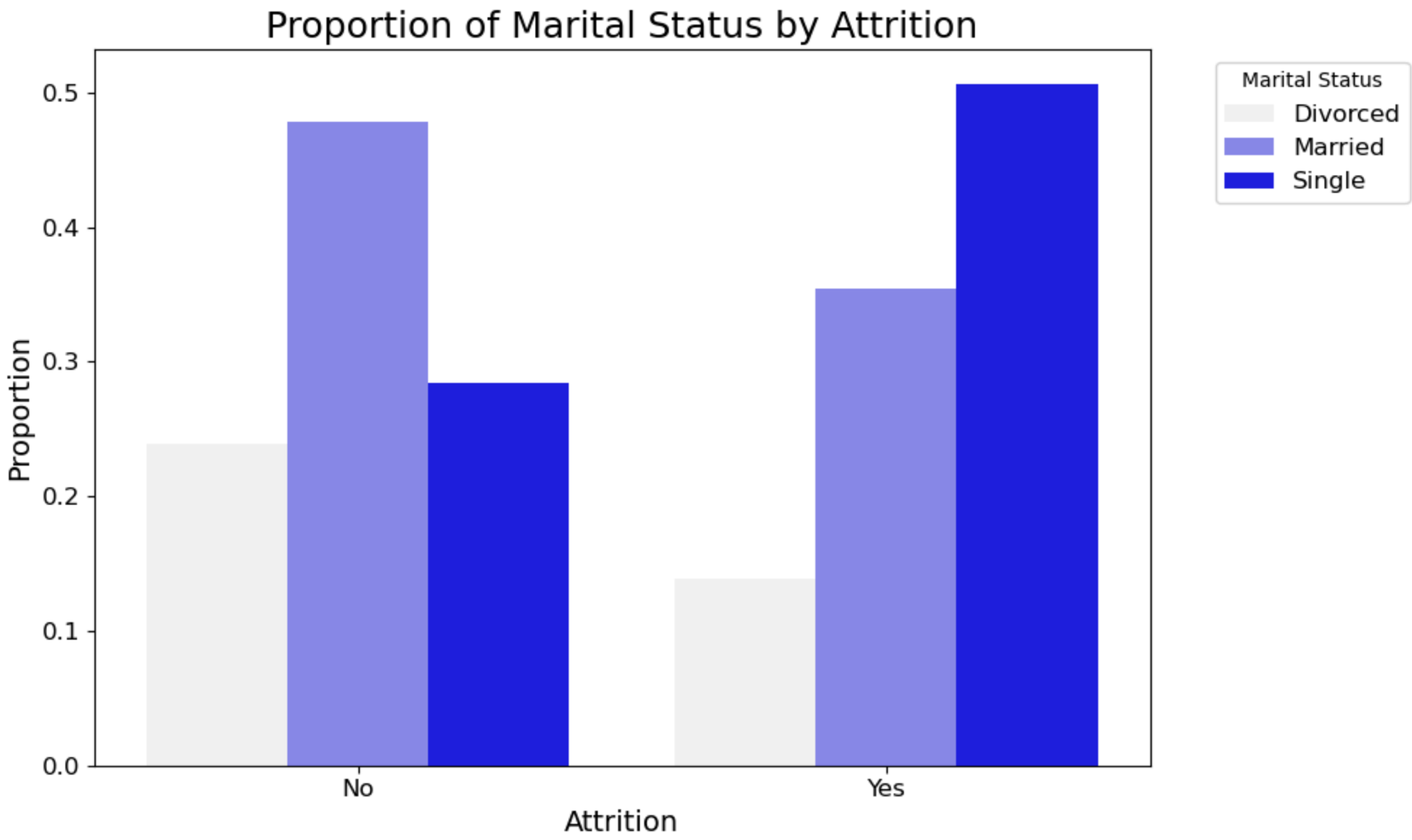
The Correlation matrix highlights strong relationships between key employment factors. These include:

1. Job Level
2. Years At Company
3. Years with Current Manager

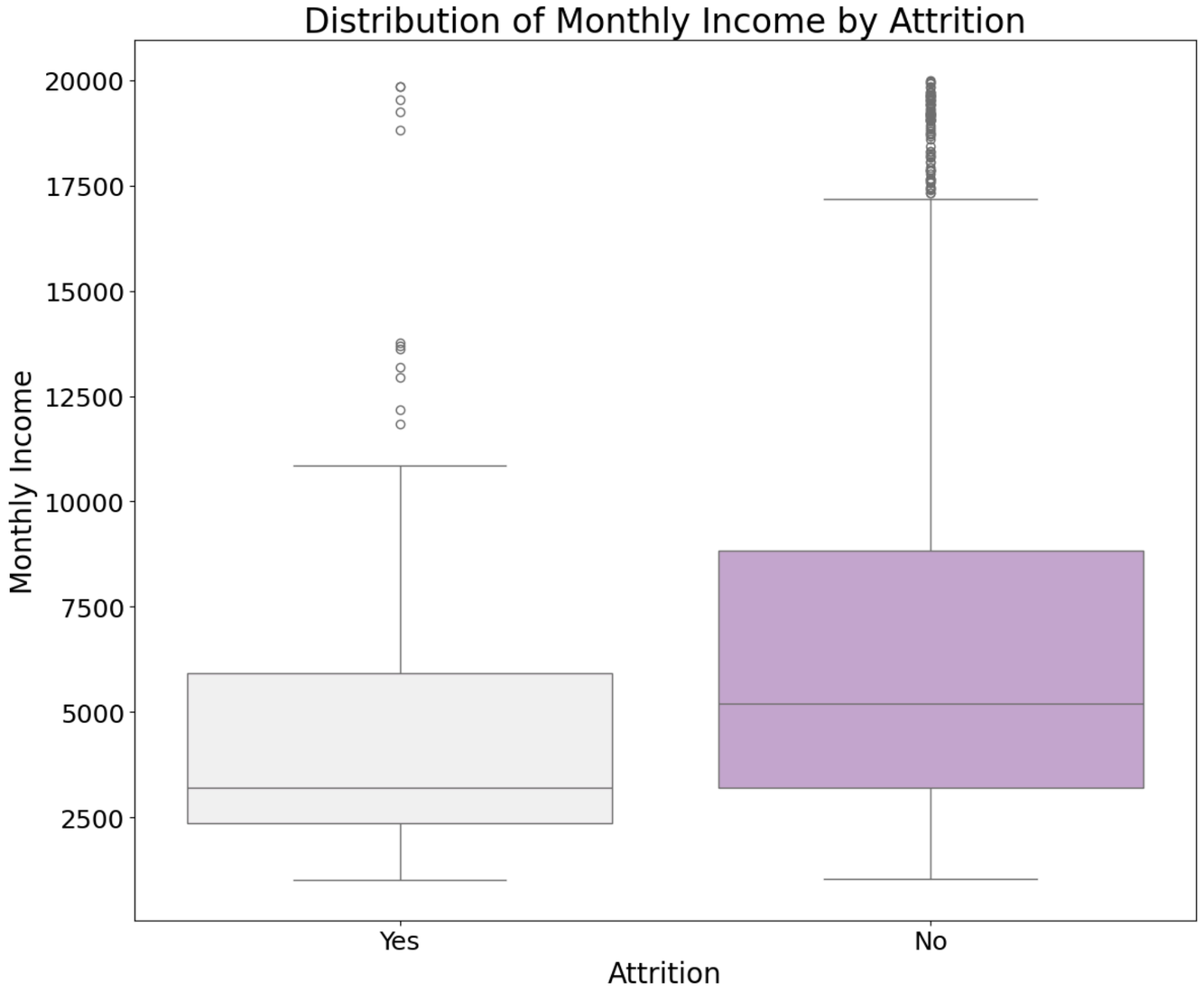


EMPLOYEES IN TECHNICAL RESEARCH AND SALES ROLES HAVE A HIGHER PROPORTION OF ATTRITION

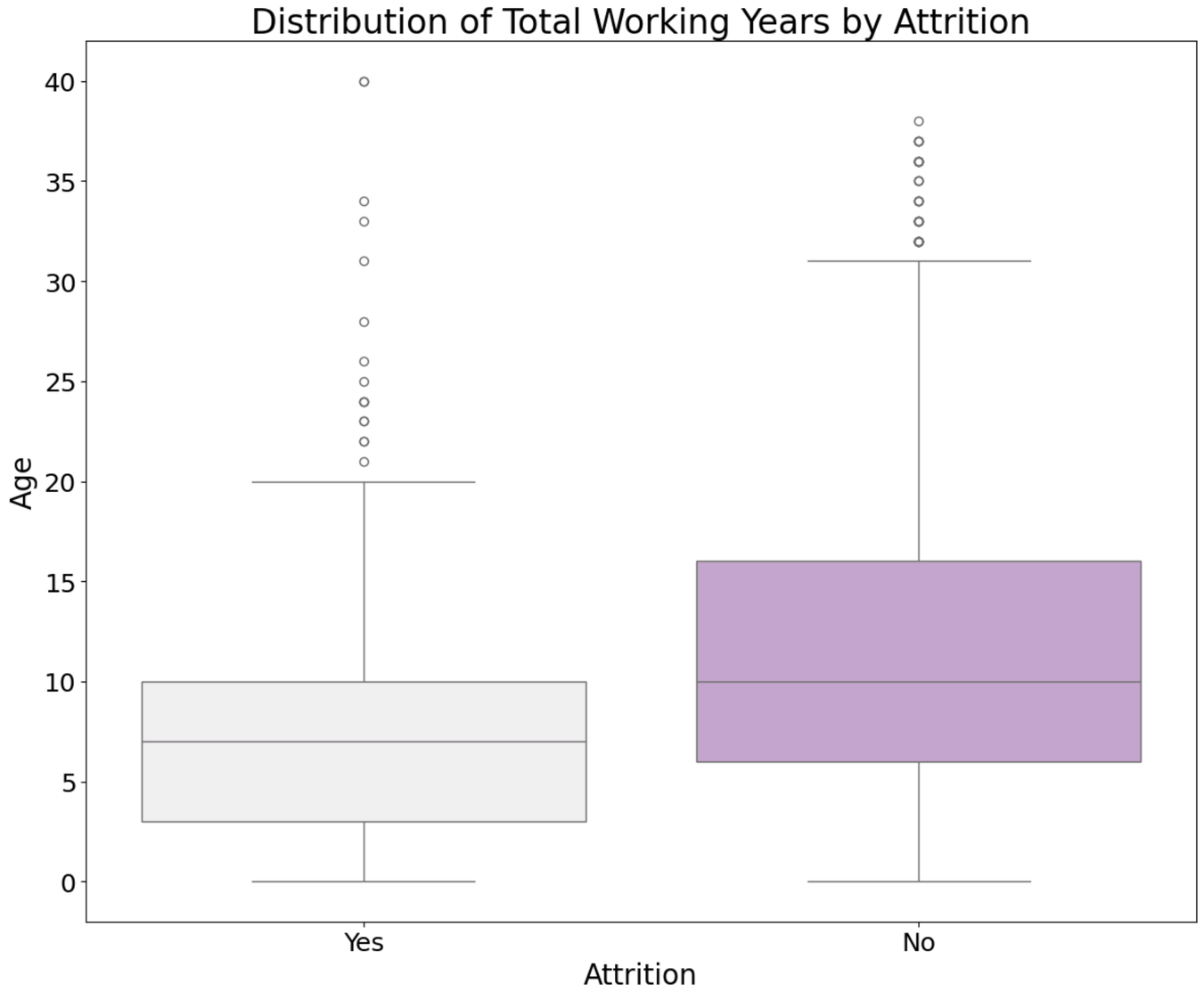
SINGLE EMPLOYEES HAVE A HIGHER PROPORTION OF ATTRITION



RETAINED EMPLOYEES HAVE HIGHER MEDIAN MONTHLY INCOMES



RETAINED EMPLOYEES HAVE HIGHER HIGHER MEDIAN WORKING YEARS



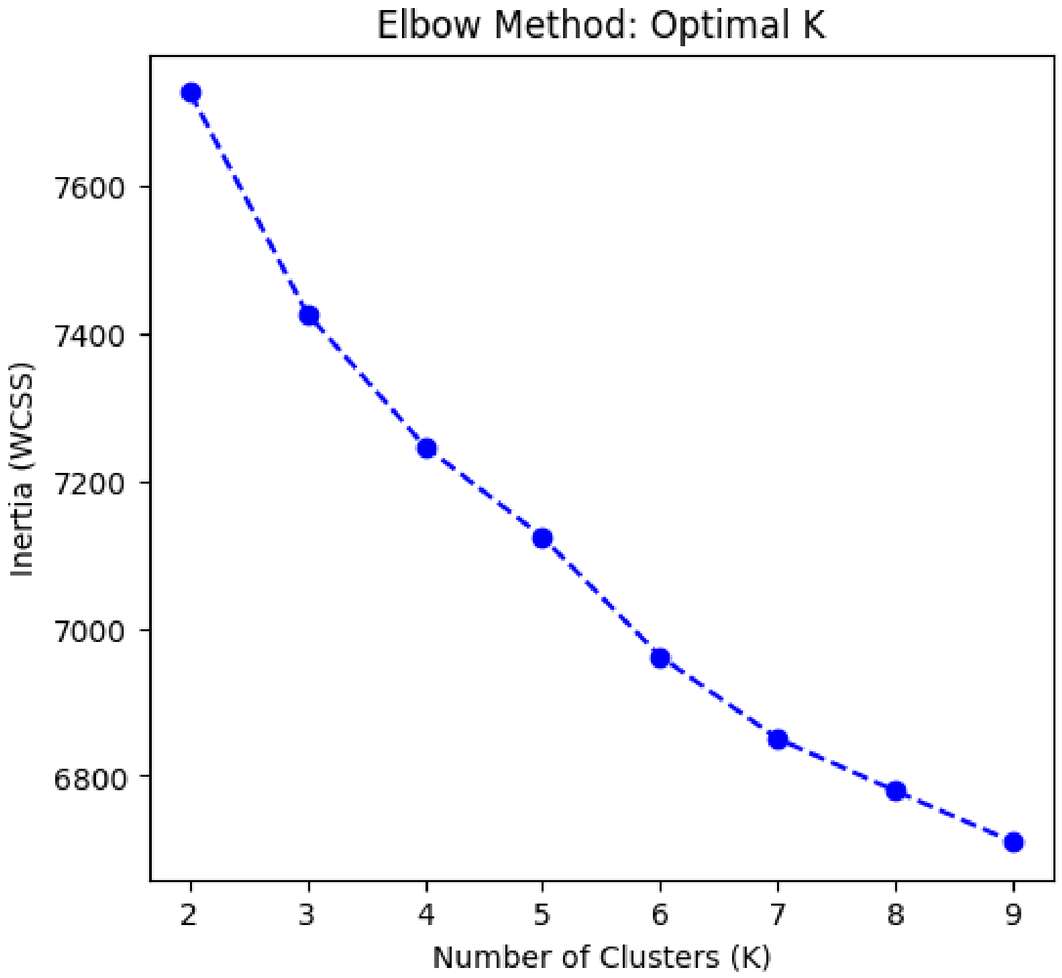
CLUSTERING

EVALUATING K-MEANS

We opted to use the best k indicated by the elbow plot for k-means and obtained the following results.

Elbow Plot

Best k: 5
Train Silhouette Score: 0.052
Test Silhouette Score: 0.051

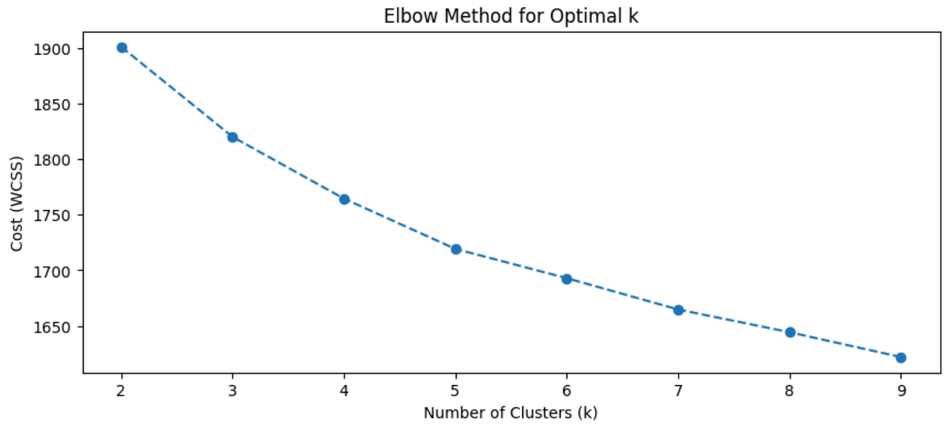


EVALUATING K-PROTOTYPES

We also opted to perform k-prototypes clustering, as it effectively handles mixed data types and is well-suited for our dataset. As k-prototypes has a higher silhouette score than k-means, we determined that it is the better performing clustering model for our data.

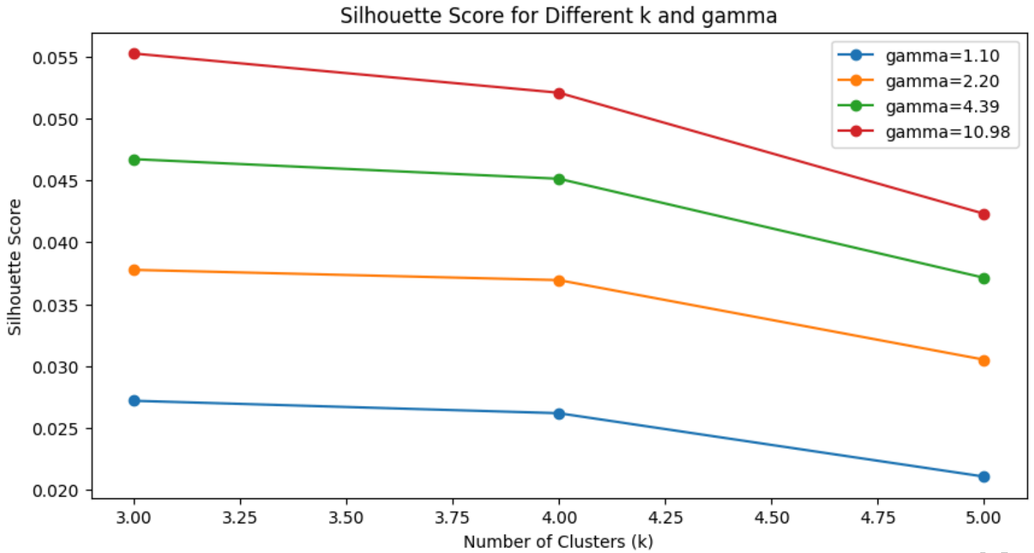
Elbow Plot

The elbow plot shows that the optimal k could between 3-5. To better understand the best k, we look at silhouette scores.



Silhouette Plot

Best k: 3
Train Silhouette Score: 0.055
Test Silhouette Score: 0.061



CLUSTERS RESULTING FROM K-PROTOTYPES

Cluster Name and Size	Cluster 0: Early-Career Researchers (49.1%)	Cluster 1: Mid-Career Technicians (26.6%)	Cluster 2: Senior Sales Executives (24.7%)
Demographics	75.3% Male, 44.2% Single	56.1% Female, 59.9% Married	53.6% Female, 58.1% Married
Education	51.5% Bachelor's	Mix of College, Bachelor's, Master's	46.4% Master's
Job Roles	Research Scientists (28%), Lab Technicians (20.7%)	Lab Technicians (29.8%), Manufacturing Directors (14.4%)	Sales Executives (66.8%), Managers (11.8%)
Daily Compensation	$773.89	$865.63	$797.35
Years Since Promotion	1.83 years	2.10 years	2.9 years
Experience	3.75 years in current role	4.25 years in current role	5.17 years in current role
Satisfaction	High environment (41.2%), Very high job satisfaction (41.4%)	Very high environment (46.2%), High job satisfaction (45.5%)	Low environment (32.5%), Low job satisfaction (29.4%)
Overtime	19.1% work overtime	55.8% work overtime	19.4% work overtime
Attrition	16.0%	16.7%	15.9%
Recommendations	**Invest in training and development programs to retain early-career talent**	**Implement flexible work policies to reduce overtime and prevent burnout**	**Conduct satisfaction surveys and career path discussions to re-engage senior employees**

CLASSIFICATION MODELS ON CLUSTERS

SMOTE - SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE

SMOTE was used to generate new synthetic samples of data for the minority class to balance the dataset. This increased class balance from approximately 16.16% for the minority to 50%

Benefits:

- Improves class balance and the model's ability to detect patterns and model performance in the minority class
- Reduces bias towards majority class
- Interpolates between existing data points to create diversity rather than duplicating like oversampling does
- Can be used by any classifier

Drawbacks:

- Risk overfitting if new samples are too similar to original data
- Could potentially add noise to the data since it uses interpolation
- Increased computational complexity

Class Balance

Before SMOTE:

No - 83.84&

Yes - 16.16%

After SMOTE:

No - 50%

Yes - 50%

LOGISTIC REGRESSION PERFORMANCE OVERVIEW

Logistic Regression

Benefits:

- Simplistic model, offering easy interpretation
- Predicts binary outcomes

Drawbacks:

- Sensitivity to outliers
- Risk overfitting with high-dimensional data

Steps

1. Balanced classes with SMOTE
2. Used forward selection with AIC to train models for each cluster
3. Performed grid search cross validation on each model to optimize regularization type, regularization strength, and class weights
4. Performed testing for accuracy and recall evaluation

Cluster Breakdown

Cluster 0:
Test Accuracy:
83.8%
Test Recall:
10%

Cluster 1:
Test Accuracy:
86.9%
Test Recall:
33.3%

Cluster 2:
Test Accuracy:
82.6%
Test Recall:
66.7%

TREE MODELS PERFORMANCE OVERVIEW

Random Forest Classifier

Benefits:

- Handles high dimensional data well
- More robust model (handles overfitting)

Drawbacks:

- Slow for real time predictions
- Lower accuracy compared to GBM for complex patterns

Gradient Boosting Machine

Benefits:

- Hypothesized higher accuracy for complex customer behavior
- Better for imbalanced datasets

Drawbacks:

- More prone to overfitting
- Slower training time

Cluster Breakdown

Cluster 0:

Test Accuracy:

1. Random Forest: 86.2%
2. Gradient Boosting: 83.6%

Test Recall:

1. Random Forest: 58.7%
2. Gradient Boosting: 50.0%

Cluster 1:

Test Accuracy:

1. Random Forest: 90.1%
2. Gradient Boosting: 90.8%

Test Recall:

1. Random Forest: 63.3%
2. Gradient Boosting: 73.3%

Cluster 2:

Test Accuracy:

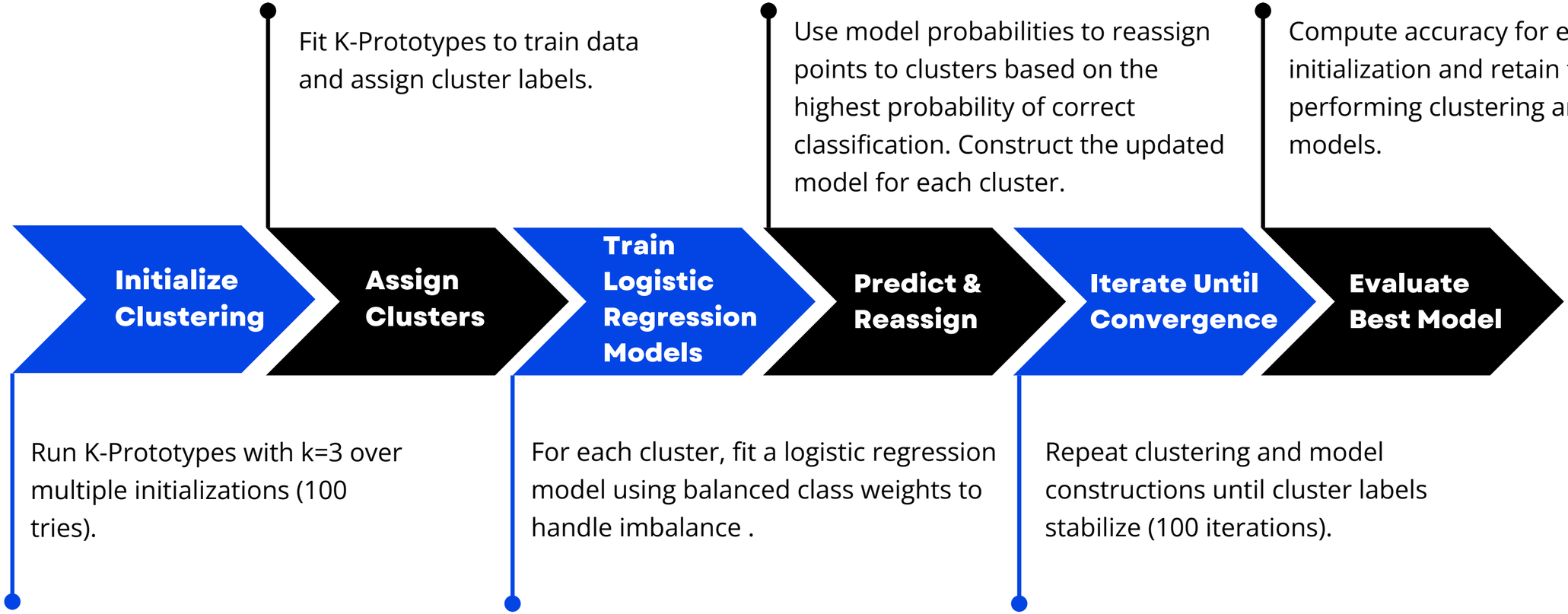
1. Random Forest: 93.4%
2. Gradient Boosting: 91.7%

Test Recall:

1. Random Forest: 68.4%
2. Gradient Boosting: 63.1%

CLUSTER-WISE REGRESSION

CLUSTER-WISE REGRESSION MODEL OVERVIEW



Model Results

Test Accuracy: 75.2%
Test Recall: 70.2%

MODEL COMPARISON

	Test Accuracy	Test Recall
Logistic Regression (average)	84.4%	36.7%
Random Forest (average)	89.9%	63.5%
GBTM (average)	88.7%	62.1%
Cluster-wise Regression	75.2%	70.2%

01 **K-Prototypes vs Cluster-wise Regression:**

- K-prototype clusters focus on job function, compensation, and satisfaction.
- Cluster-wise regression emphasizes career stage transitions, and promotions.

02 **Regression vs Tree Models:**

- Tree models and Logistic Regression have higher test accuracy compared to Cluster-wise regression.
  - Tree models can handle imbalanced datasets better than Cluser-wise Regression
- Test Recall for Cluster-wise Regression is better than other models
  - Better at handling minority classes within clusters and has lower bias.

CLUSTER PROFILES

Cluster Name and Size	Cluster 0: "Experienced Contributors" (48.5%)	Cluster 1: "Mid-Career Sales Employees" (29.1%)	Cluster 2: "Young Innovators" (21.3%)
Demographics	63.2% Male, 46.8% Married	62.2% Female, 48.4% Married	71.6% Male, 44% Married
Education	42% Master’s	48.3% Bachelor’s	40.1% Bachelor's
Job Roles	30% Managerial, 20% Research Directors	56.8% Sales Executive	32.8% Laboratory Technician, 28.5% Research Scientist
Daily Compensation	$787.95	$870.40	$771.12
Years in Current Role	6.52 years	4.71 years	2.96 years
Years Since Promotion	4.56 years	2.16 years	1.17years
Satisfaction	High environment (36.4%), Very high job satisfaction (39.2%)	Very high environment (38.0%), High job satisfaction (31.4%)	Low environment (34.9%), High job satisfaction (32.3%)
Attrition	8.4%	15.2%	20%
Recommendations	**Implement tailored recognition and mentoring programs to leverage their expertise and provide leadership development**	**Foster a culture of continuous learning and provide clear pathways for career advancement to keep them motivated.**	**Encourage skill development and innovation initiatives to engage and retain younger employees while addressing their satisfaction concerns.**

MAIN TAKEAWAYS

01 **Distinct Clusters**

After performing clustering on our data we were able to find three distinct employee clusters and thereby identify which features of each cluster were most correlated with attrition.

02 **Attrition Prediction is Accurate**

Our model allowed us to predict employee attrition with satisfactory accuracy and recall, where our ‘Young Innovators’ cluster is most likely to leave.

03 **Steps to Improve Model**

To enhance our model, future work should focus on refining the class sampling method, gathering more data on employees who have attrited, and implementing strategies to retain employees while evaluating their effectiveness.

THANK YOU!