PROJECT 1
Employee Attrition

Spencer Thompson (thompsonsm@smu.edu)

21 October 2025

DataScience@SMU

Agenda

01	**Overview**	Slide 03
02	**EDA**	Slide 04
03	**Statistical Tests**	Slide 10
04	**Modeling Approaches**	Slide 15
05	**Model Performance Comparison**	Slide 21
06	**Final Recommendations**	Slide 23

Overview

Goals

- To identify the potential causes of employee attrition
- Develop a model to predict and identify possible attrition

Methods

- Using T Tests, ANOVA, and Kruskal-Wallis tests to analyze significance of variables
- Predicting attrition using Logistic Regression, kNN, and Naive Bayes models

Results

- Three best predictors: overtime, job satisfaction, and job involvement
- 5 very significant predictors
- 16 significant predictors in total

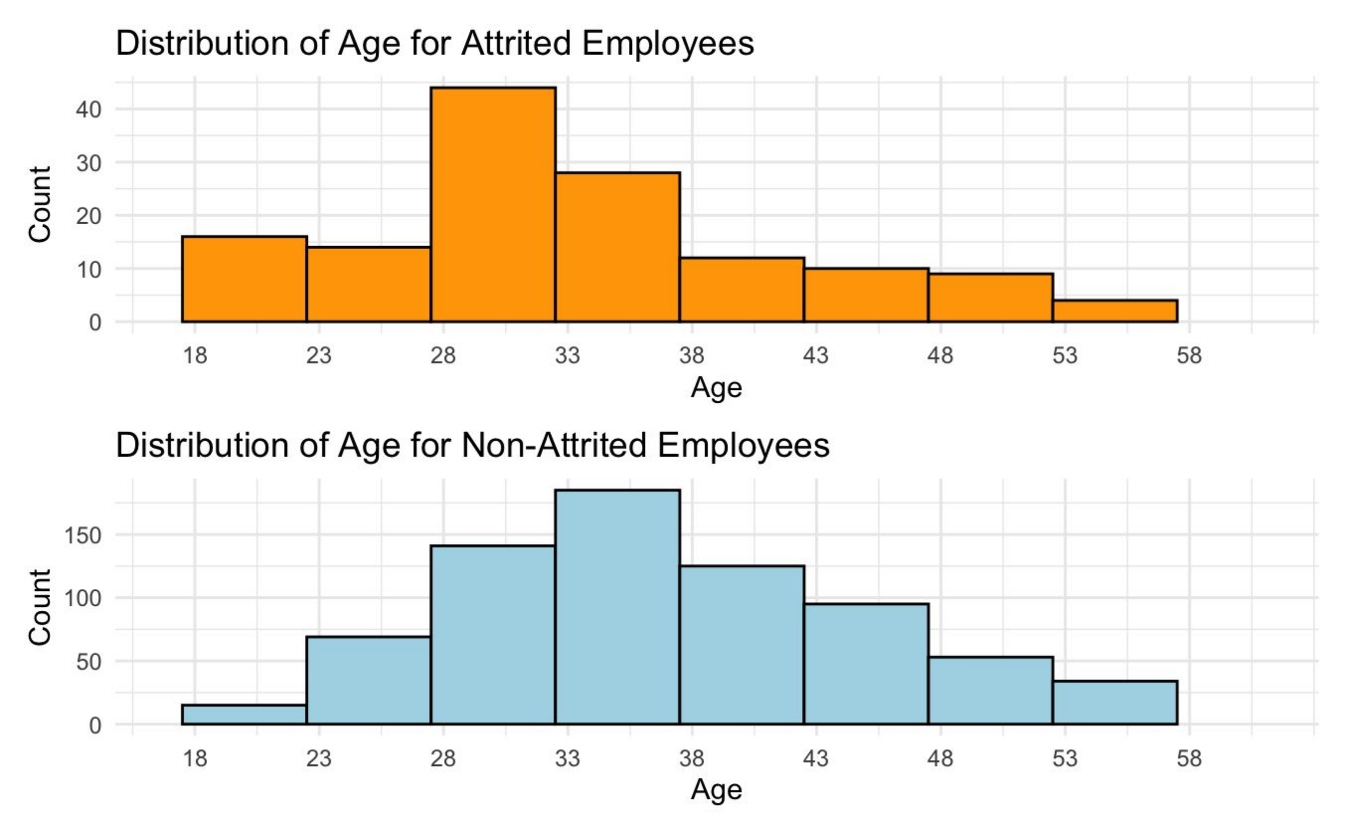
Raw Numbers

- 140 attrited employees (16.1%)

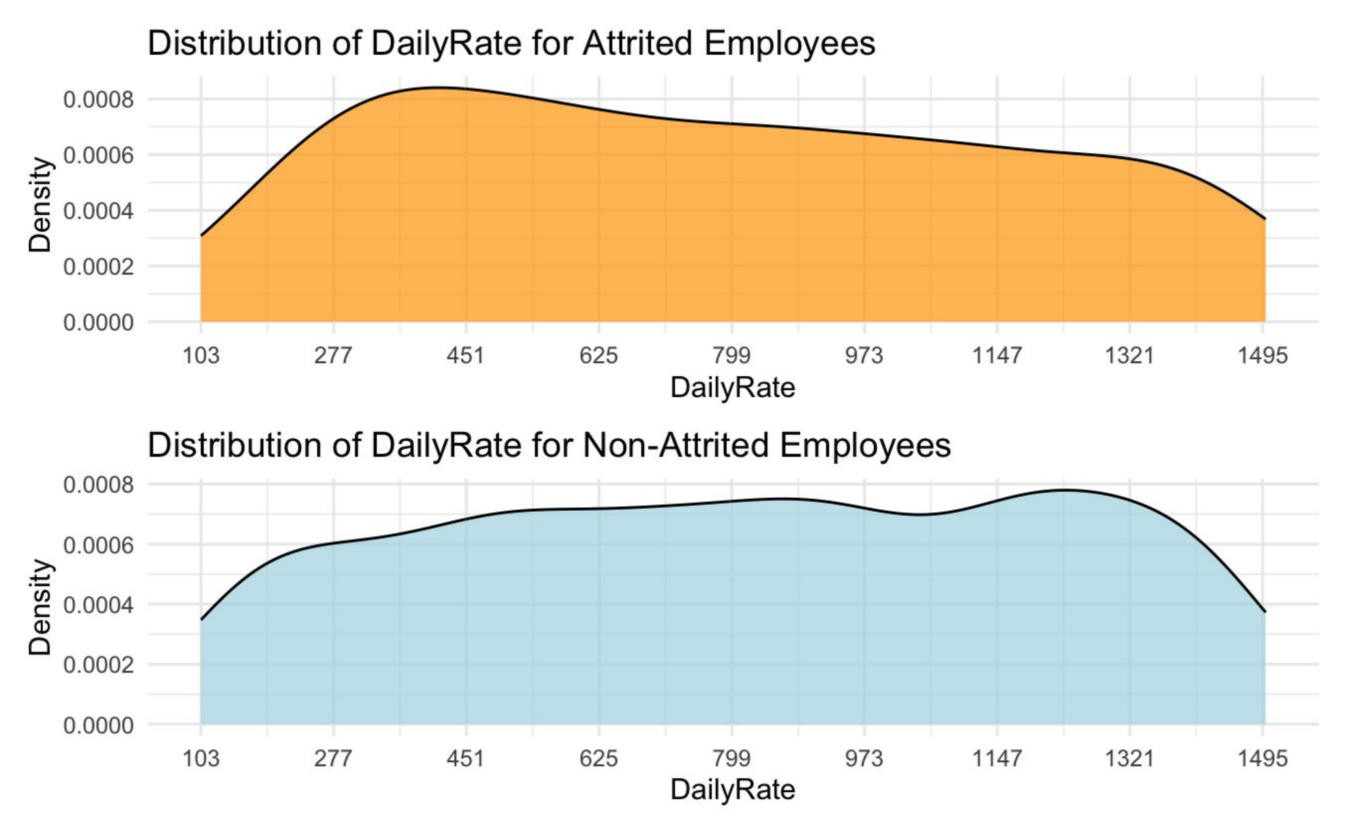
EDA

Exploratory Data Analysis

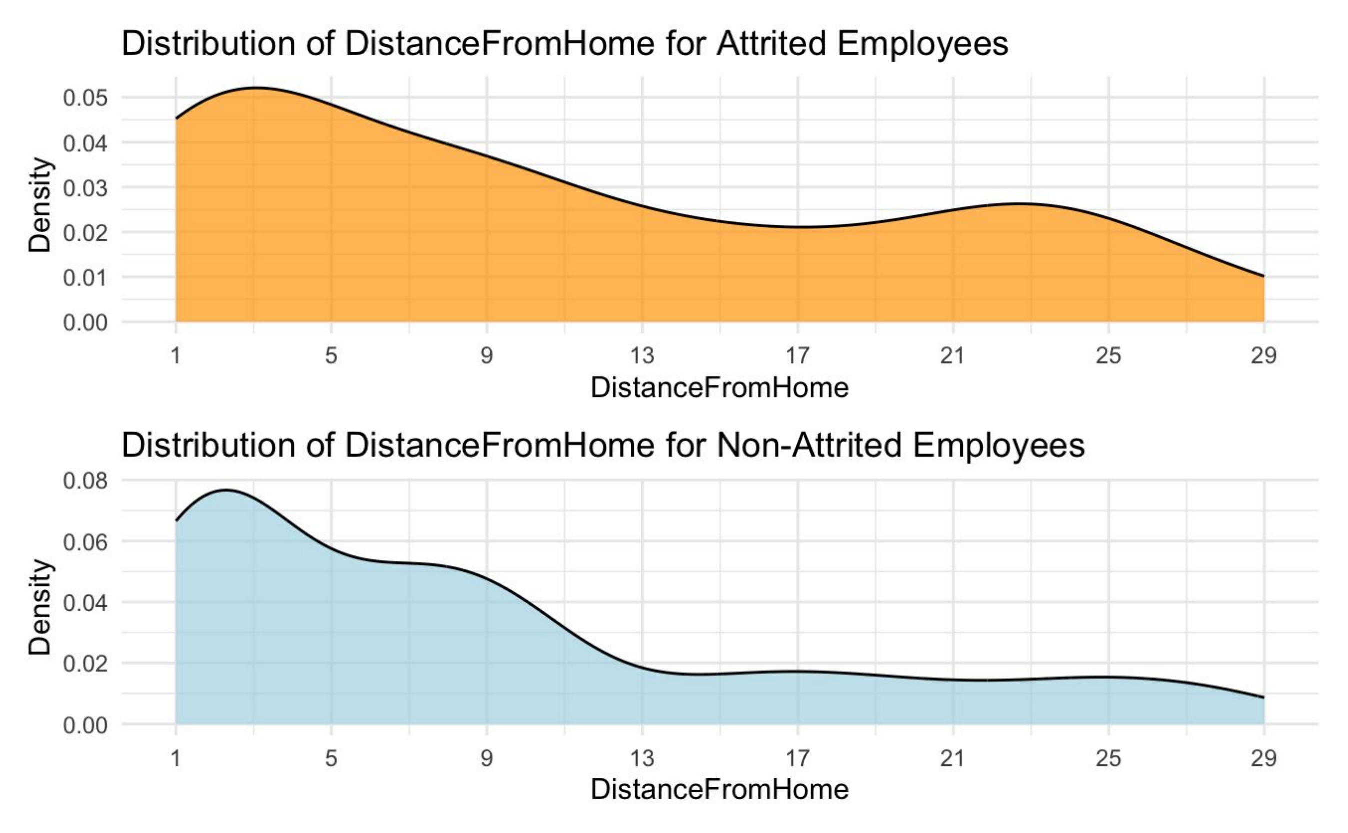
Distribution of Age



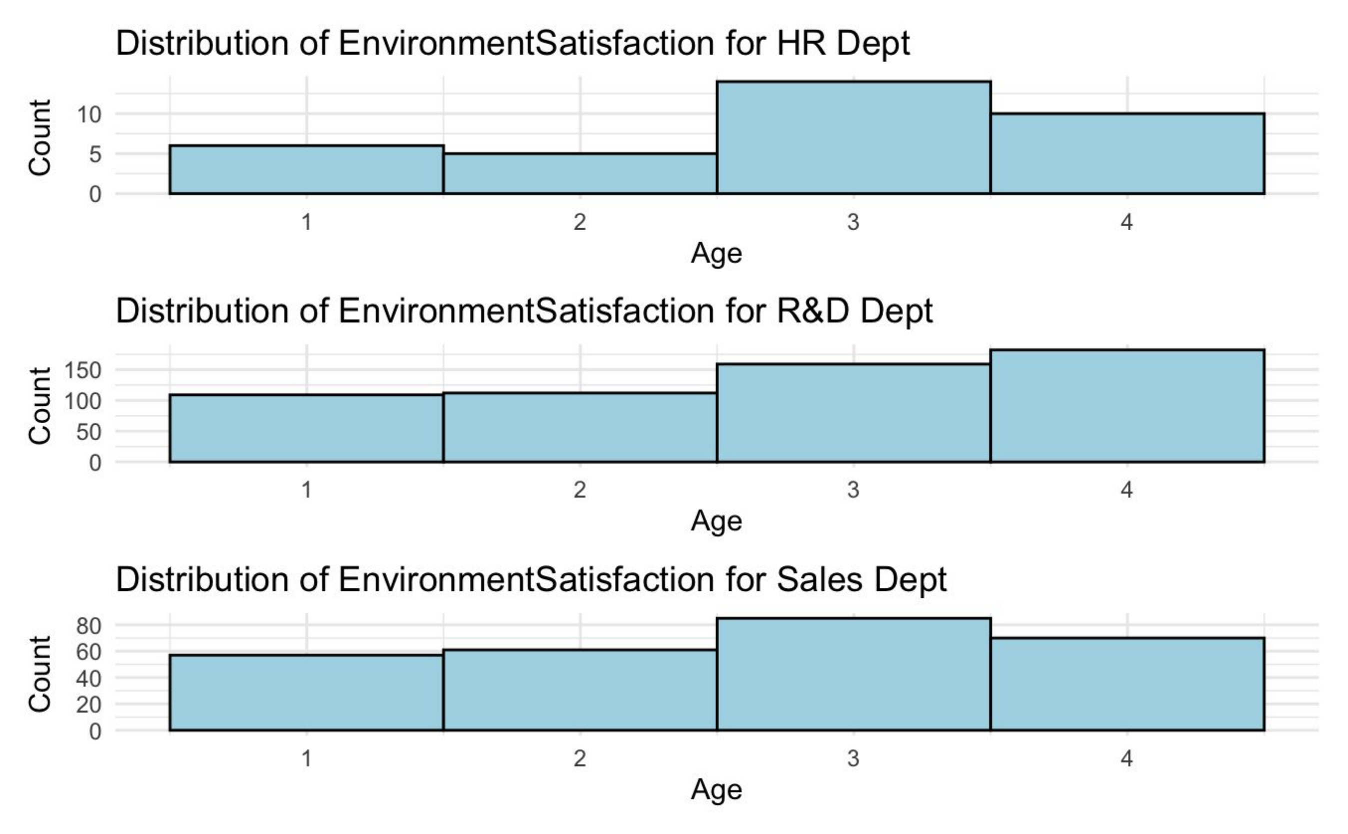
Distribution of Daily Rate (Salary)



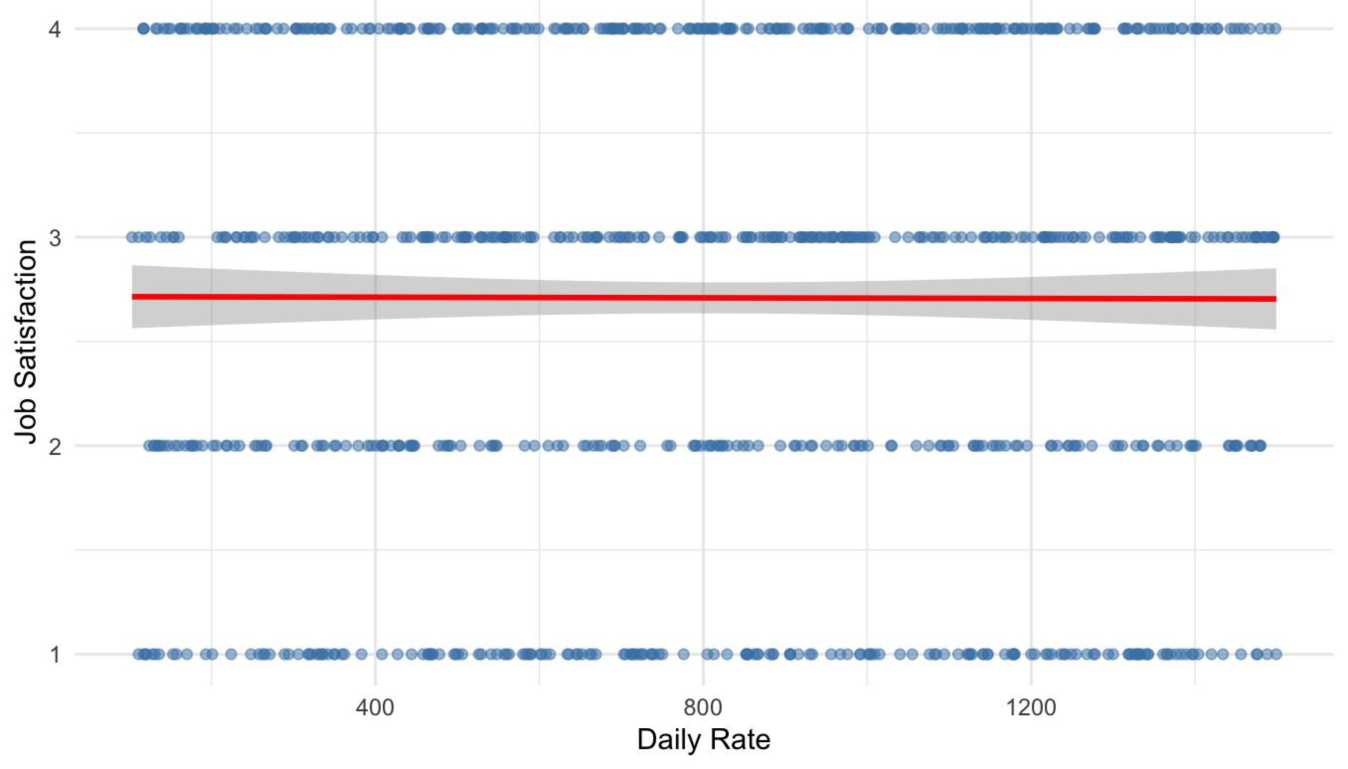
Distribution of Distance From Home



Distribution of Environment Satisfaction



Job Satisfaction vs Daily Rate



Tests

- 2 Sample T-Tests
- Kruskal-Wallis Test
- 3 Way ANOVA

Mean DistanceFromHome T-Test

- Claim: Mean miles traveled to work (ie DistanceFromHome) differs between employees that left and employees that stayed
- Conclusion: There was a significant difference in mean miles traveled to work between attrited employees and remaining employees
- Interval: We are 95% confident that the plausible difference in mean miles traveled to work between the groups is between -0.35 and -3.5. Where attrited employees on average drove at least 0.35 miles further to work.

Mean JobSatisfaction T-Test

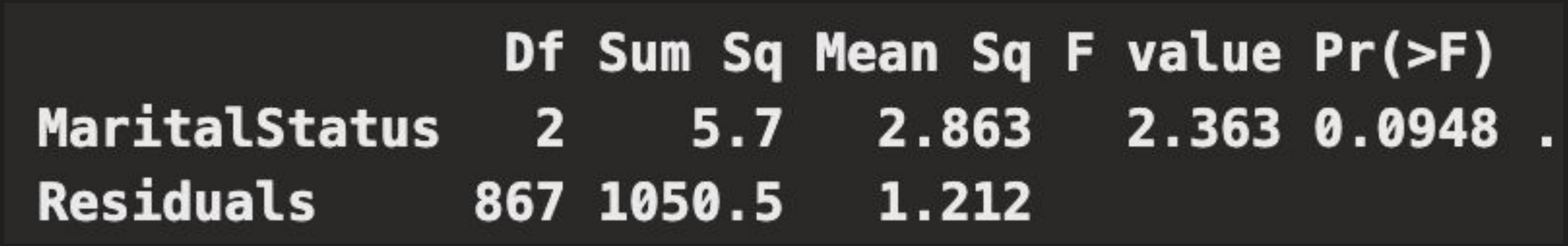
- Claim: Mean job satisfaction differs between employees that work overtime and employees that don’t
- Conclusion: There was no significant difference in mean job satisfaction between overtime employees and non-overtime employees
- Interval: We are 95% confident that the plausible difference in mean job satisfaction between the groups is between -0.24 and 0.09.

Work-Life Balance vs Travel Frequency

- Claim: Mean work-life balance differs between the 3 levels of travel frequency
- Method: Kruskal-Wallis test (since the data is ordinal)
- Conclusion: There was no significant difference in mean work-life balance score between the 3 types of travel frequency

RelationshipSatisfaction vs Marital Status

- Claim: Mean RelationshipSatisfaction score differs between the 3 levels of marital status (divorced, single, married)
- Method: ANOVA test on 3 groups
- Conclusion: There is no significant difference in mean Relationship Satisfaction across marital status groups. Although the P-value (0.09) suggests it’s worth examining closer.



```
               Df Sum Sq Mean Sq F value Pr(>F)
MaritalStatus   2    5.7   2.863   2.363 0.0948 .
Residuals     867 1050.5   1.212
```

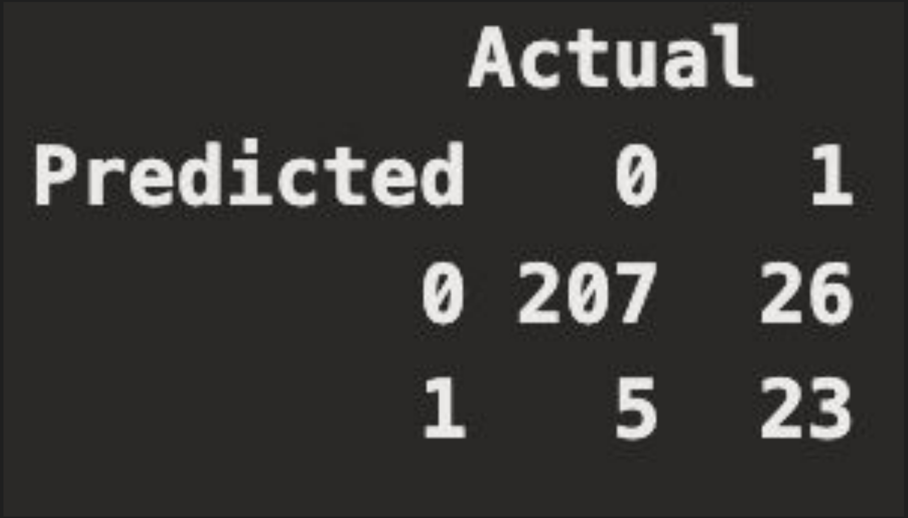
Modeling

- Logistic Regression
- K Nearest Neighbors
- Naive Bayes

Logistic Regression Model

Performance Metrics

- Accuracy: 88.2%
- Sensitivity: 46.9%
- Specificity: 97.6%

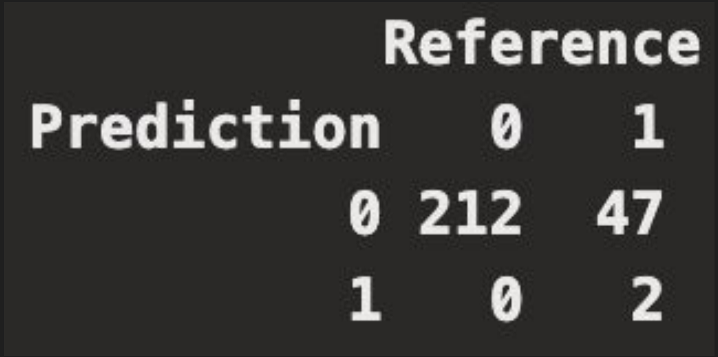


```
        Actual
Predicted   0   1
        0 207  26
        1   5  23
```

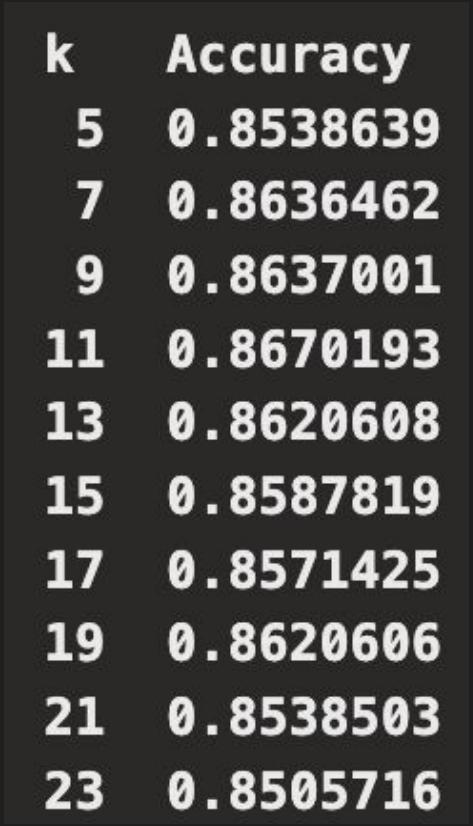
K Nearest Neighbors

Performance Metrics

- Accuracy: 81.9%
- Sensitivity: 4%
- Specificity: 100%



```
          Reference
Prediction   0   1
         0 212  47
         1   0   2
```

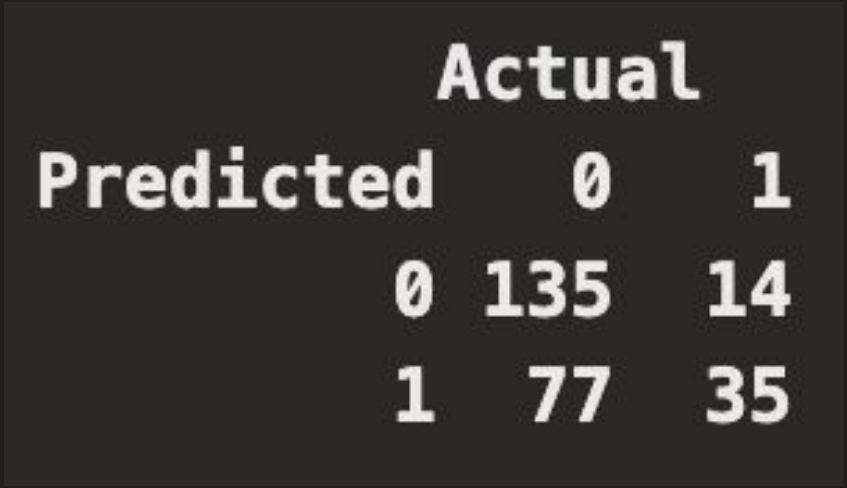


```
k   Accuracy
 5  0.8538639
 7  0.8636462
 9  0.8637001
11  0.8670193
13  0.8620608
15  0.8587819
17  0.8571425
19  0.8620606
21  0.8538503
23  0.8505716
```

Naive Bayes

Performance Metrics

- Accuracy: 65.1%
- Sensitivity: 71.4%
- Specificity: 63.7%



```
         Actual
Predicted   0   1
        0 135  14
        1  77  35
```

Best Predictors

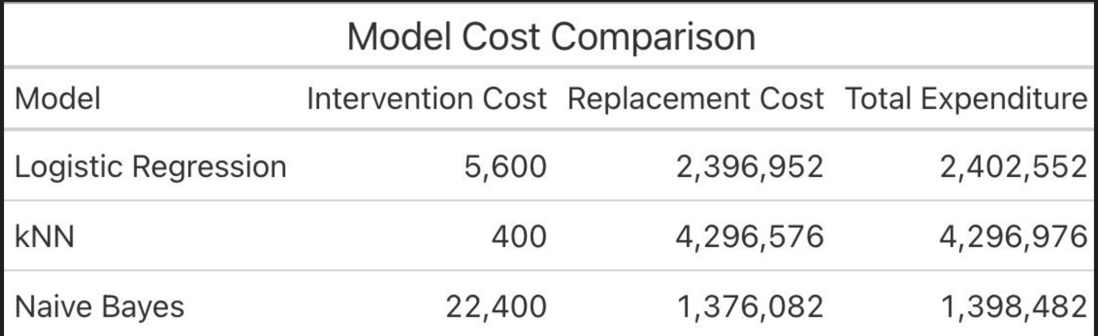
- OverTime_num (z = 7.499)
- JobInvolvement (z = -5.046)
- JobSatisfaction (z = -4.554)
- NumCompaniesWorked (z = 4.234)
- YearsSinceLastPromotion (z = 4.037)

- BusinessTravel_num (z = 3.571)
- DistanceFromHome (z = 3.361)
- isDirector (z = -2.997)
- WorkLifeBalance (z = -2.973)
- TrainingTimesLastYear (z = -2.726)
- EnvironmentSatisfaction (z = -2.714)

- MaritalStatus_num (z = 3.416)
- TotalWorkingYears (z = -2.554)
- YearsWithCurrManager (z = -2.301)
- RelationshipSatisfaction (z = -2.214)
- YearsInCurrentRole (z = -2.102)

Model Comparison

Model Performance Comparison



Model Cost Comparison

Model	Intervention Cost	Replacement Cost	Total Expenditure
Logistic Regression	5,600	2,396,952	2,402,552
kNN	400	4,296,576	4,296,976
Naive Bayes	22,400	1,376,082	1,398,482

Naive Bayes

- Very conservative (lots of false positives)
- Best at catching actual attrition (71% sensitivity)
- Overall, best in terms of total expenditure with given replacement & intervention costs

Final Recommendations

Final Recommendations

- Improve job satisfaction and job involvement
  - Prioritize equal involvement
  - Solicit and listen to employee feedback
- Reduce business travel and commuting distance
  - Test out work-from-home options for select employees/teams
- Prioritize manager consistency and training
  - Keep employees working with the same manager
  - Incentivize going to trainings
- Emphasize work environment, relationships, and work-life balance
  - Let employees pick their teammates
  - Modify the PTO policy, encourage time off
  - Encourage cross-functional work and collaborating with new people