Employee Attrition

Data Mining Project

Jaime Bunay, Siddarth Bhagirath, Jonathan Caussin, Nisargvan Goswami, Cindy Li, & Edosa Odia

Motivation

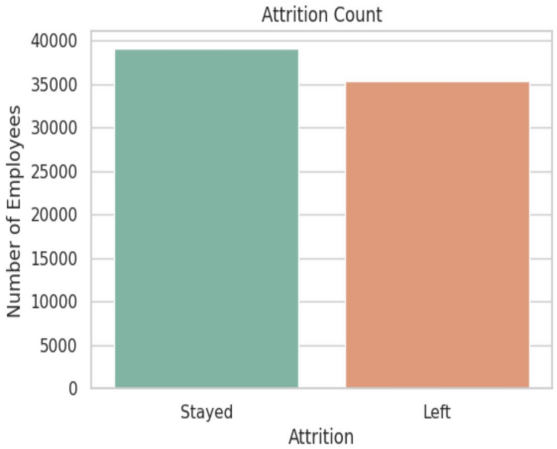
Dataset Research Question:

- What factors predict the most significant levels of employee attrition?

Implications:

- Job Satisfaction
- Factors of Employee Turnover
- Workplace Culture

Data Source



Employee Attrition Dataset

Source: Kaggle

Summary: Collection of employee records

Attrition: If employee has left the company

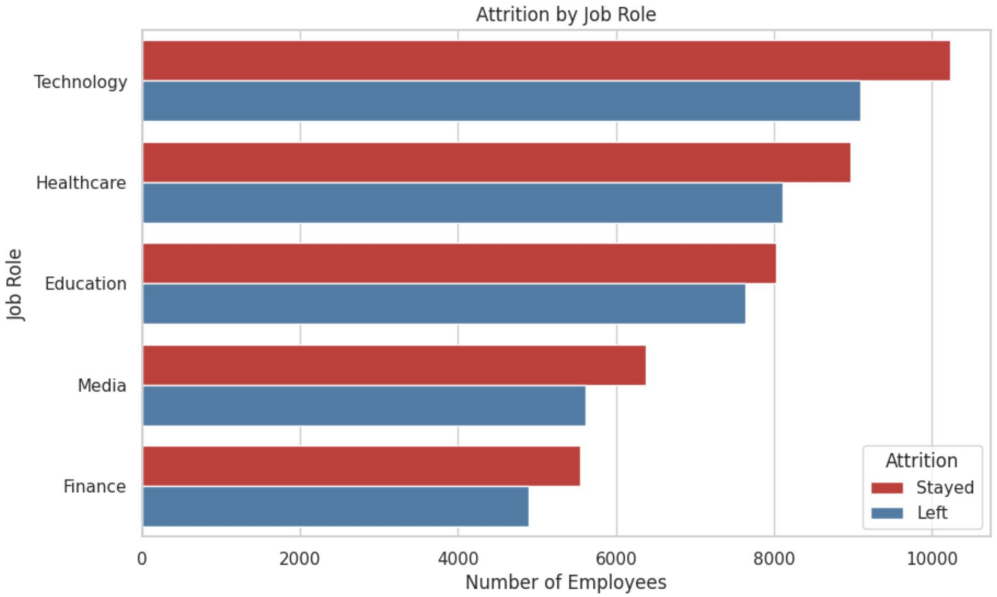
0 = stayed, 1 = left

Factors: Employee demographics & performance

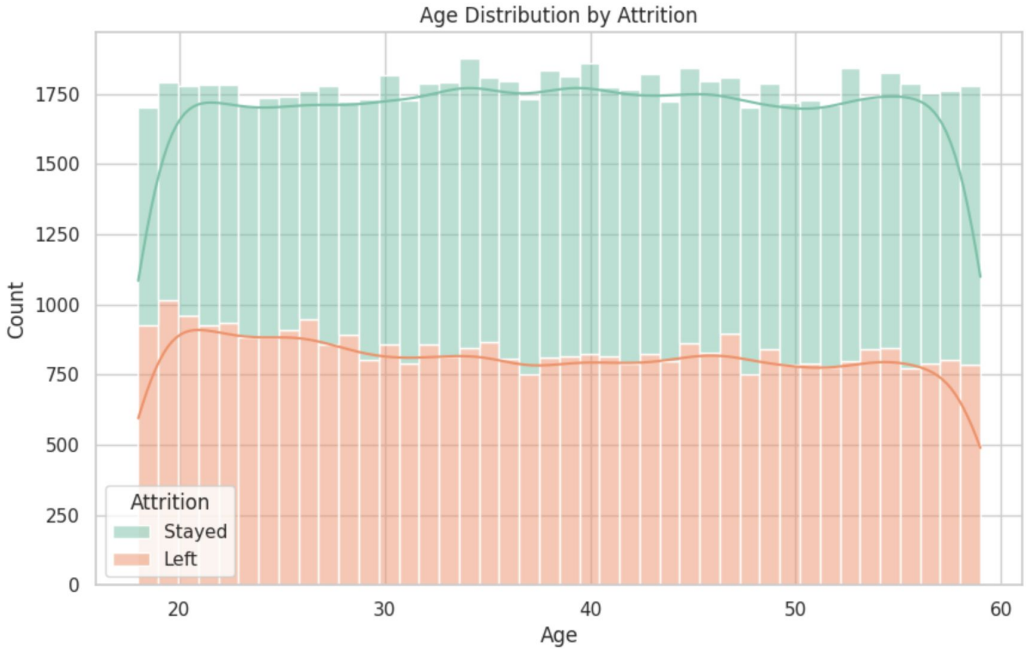
Job & company details

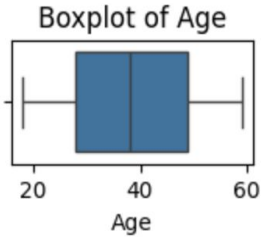
Samples: 74,498

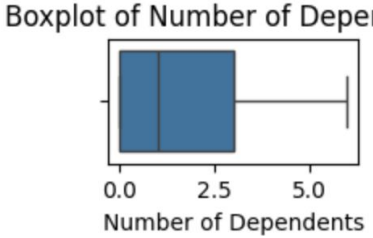
Summary Statistics

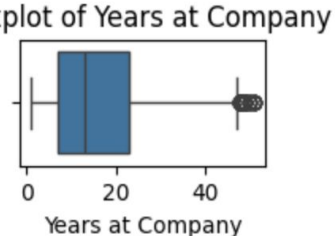


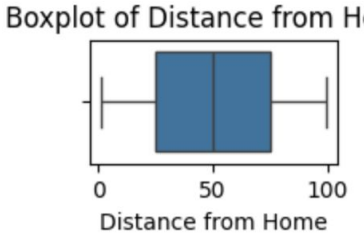
Summary Statistics











Model Selection

Predict Label: Attrition (Binary 0 - stayed and 1 - left)

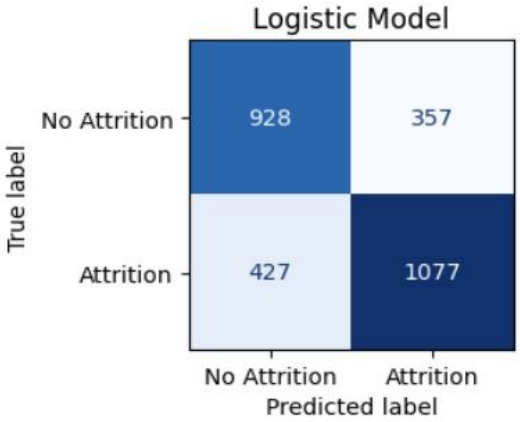
Classification Models

- Logistic Regression
- Decision Tree
- Random Forest
- SVM

Model Performance

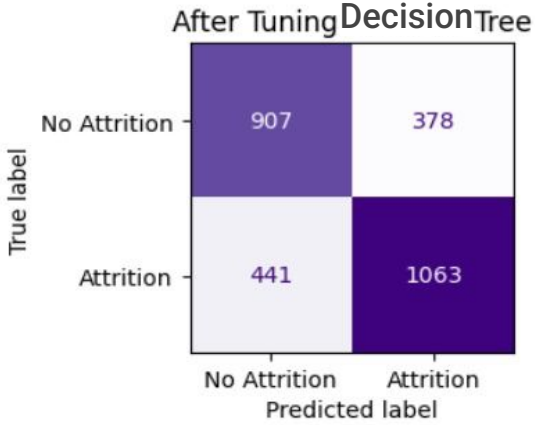
Logistic Regression

Accuracy | **72%**
Precision | 75%
Recall | **72%**
F1 Score | **73%**



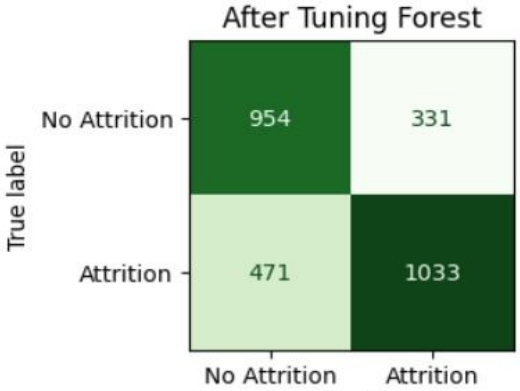
Decision Tree

Accuracy | 71%
Precision | 74%
Recall | 71%
F1 Score | 72%



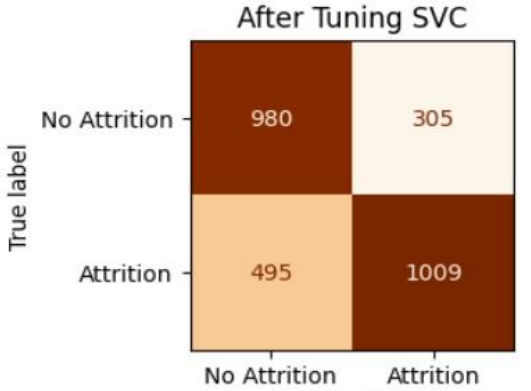
Random Forest

Accuracy | 71%
Precision | 76%
Recall | 69%
F1 Score | 72%



SVM

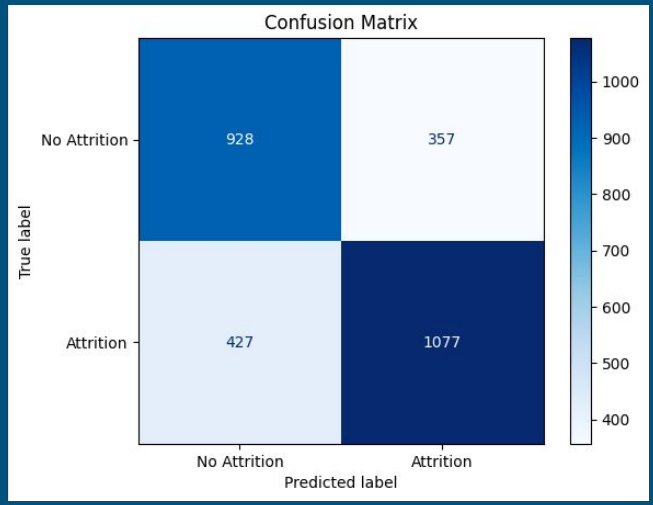
Accuracy | 71%
Precision | **77%**
Recall | 67%
F1 Score | 72%



Logistic Regression

- **Overall Accuracy: 72%**
- **Class 0**
  - Precision: 0.68
  - Recall: 0.72
  - F1 Score: 0.70
- **Class 1**
  - Precision: 0.75
  - Recall: 0.72
  - F1 Score: 0.73
- Out of this, 75% identified from Class 1 (left) versus 68% from Class 0 (stayed)
- Logistic regression was more effective in identifying Class 1 than Class 0

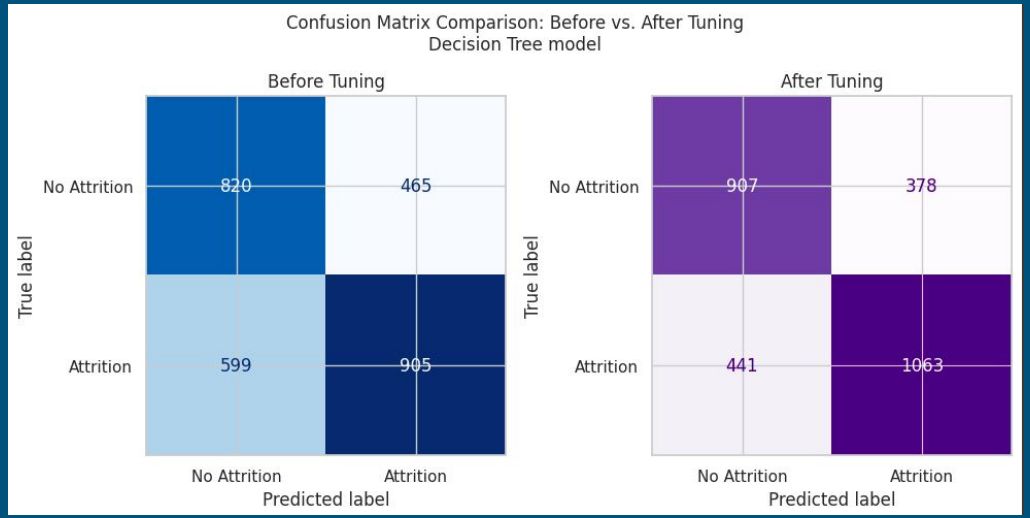
```
Accuracy: 0.7188956615274292
Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.72      0.70      1285
           1       0.75      0.72      0.73      1504

    accuracy                           0.72      2789
   macro avg       0.72      0.72      0.72      2789
weighted avg       0.72      0.72      0.72      2789
```



Decision Tree Modelling

```
Accuracy: 61.85%
Classification Report:
              precision    recall  f1-score   support

           0       0.58      0.64      0.61      1285
           1       0.66      0.60      0.63      1504

    accuracy                           0.62      2789
   macro avg       0.62      0.62      0.62      2789
weighted avg       0.62      0.62      0.62      2789

30
```

- **Overall Accuracy: 62%**
- **Class 0**
  - Precision: 0.58
  - Recall: 0.64
  - F1 Score: 0.61
- **Class 1**
  - Precision: 0.66
  - Recall: 0.60
  - F1 Score: 0.63
- Out of this, 66% identified from Class 1 (left) versus 58% from Class 0 (stayed)
- Decision tree modelling was more effective in identifying Class 1 than Class 0



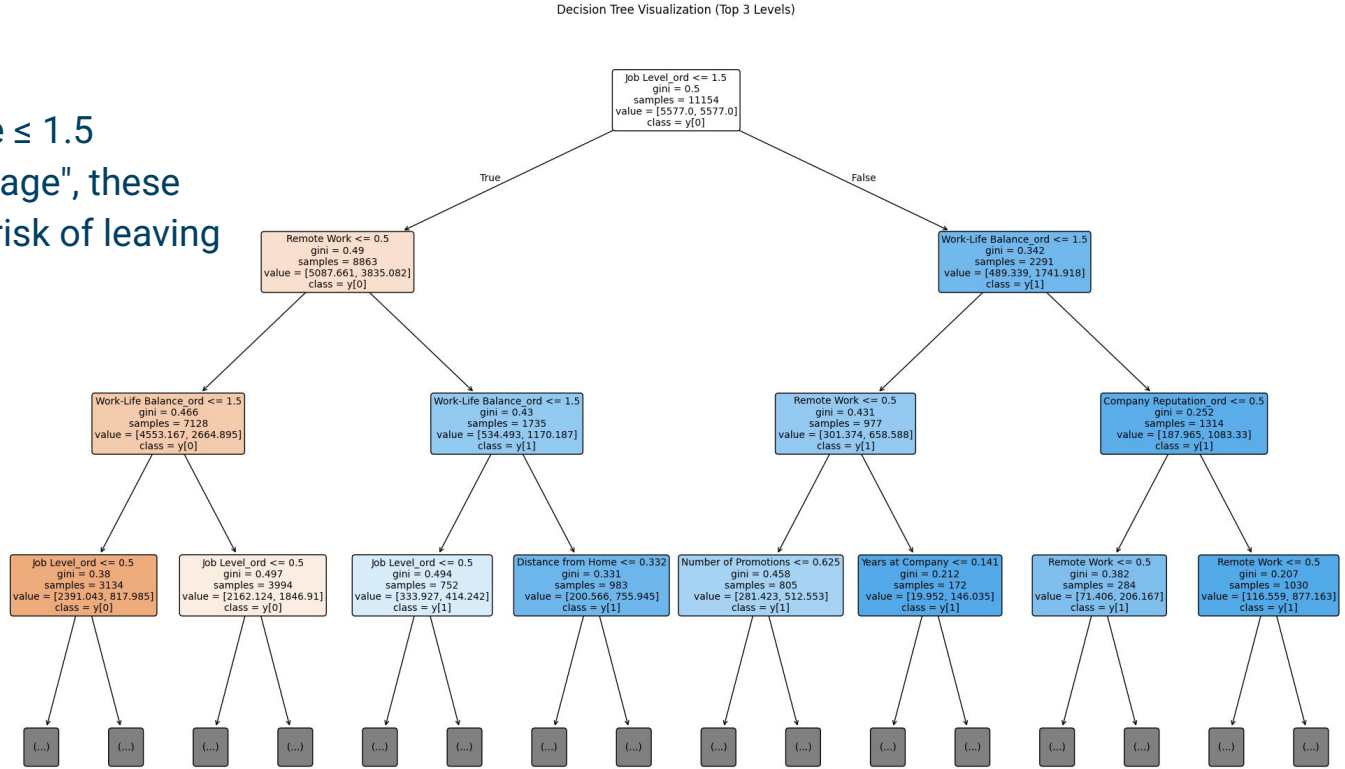
Decision Tree Modelling

Top Node: Job Level ≤ 1.5

- Employees at lower job levels (Job Level_ord) are more likely to leave.

Right Node: Work Life balance ≤ 1.5

- Rating of "Poor" or "Average", these employees are more at risk of leaving



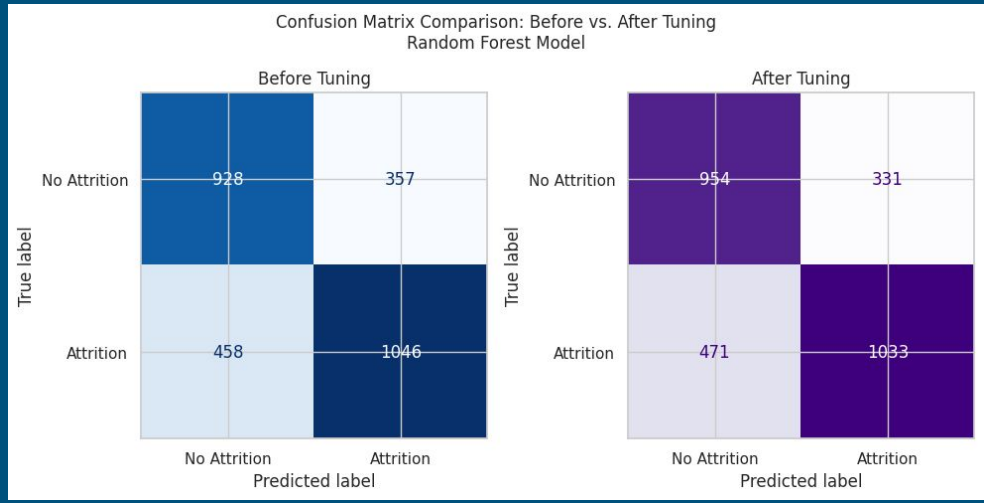
Random Forest Classification

- After Tuning Metrics (class 1)
  - Precision: 0.75
  - Recall: 0.68
  - Accuracy: 71.24%
- Precision-Recall tradeoff
  - Decrease in FP(331)
  - Increase in FN(471)
- Improved precision helps identify employees likely to leave
- Slight drop in recall may miss some attrition cases

Accuracy: 70.78%

Classification Report:

	precision	recall	f1-score	support
0	0.67	0.72	0.69	1285
1	0.75	0.70	0.72	1504
accuracy			0.71	2789
macro avg	0.71	0.71	0.71	2789
weighted avg	0.71	0.71	0.71	2789



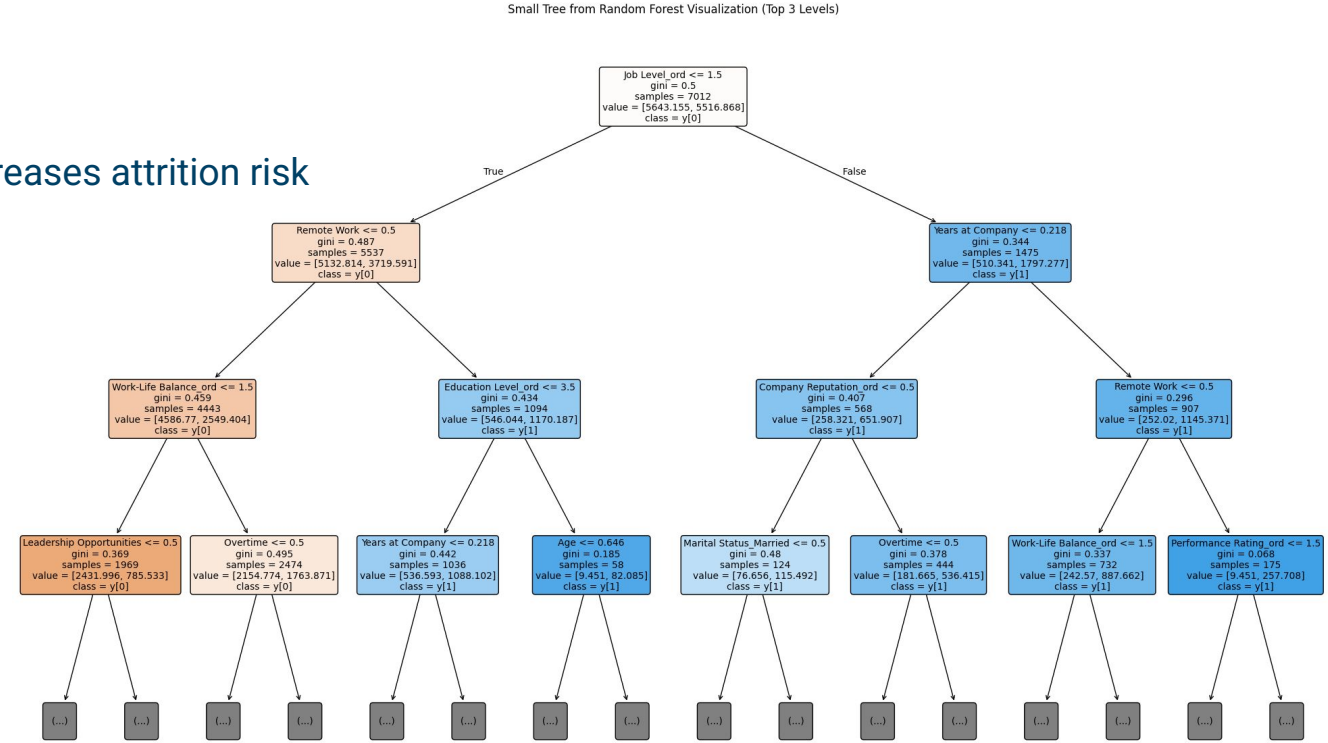
Random Forest Classification

Top Node: Job Level ≤ 1.5

- Employees at lower job levels (Job Level_ord) are more likely to leave.

Left Node: Remote Work ≤ 0.5

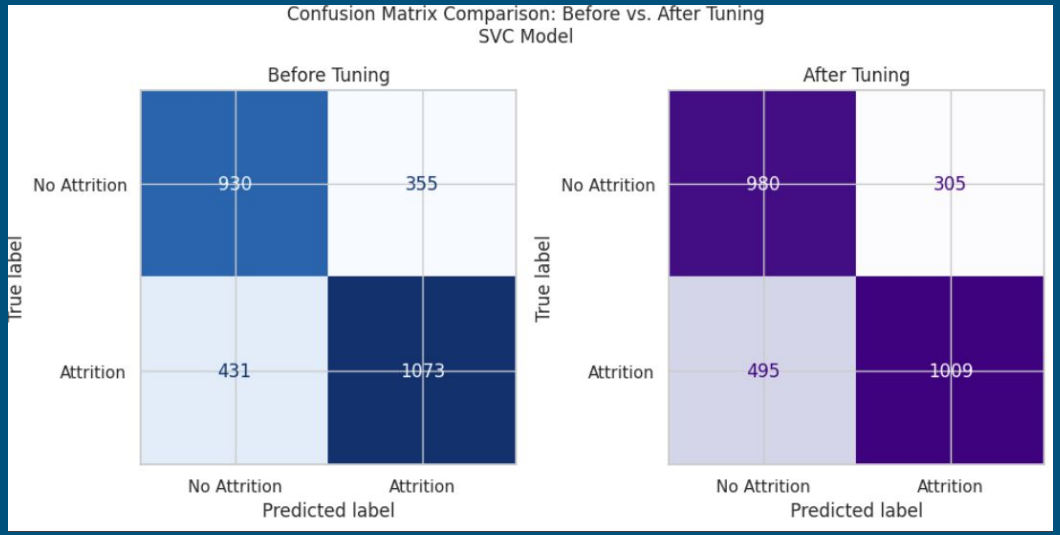
- Lack of remote work increases attrition risk



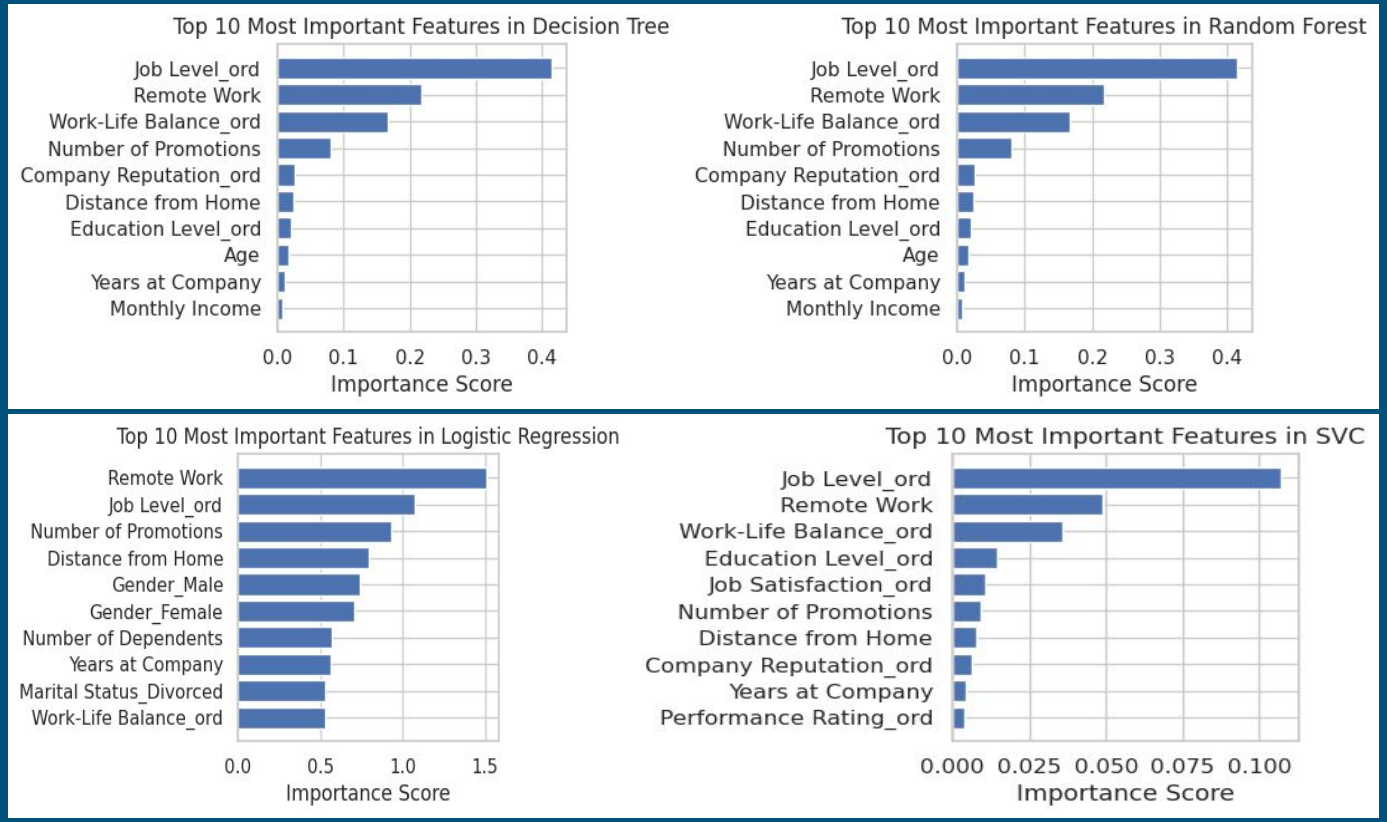
Support Vector Machine

- After Tuning Metrics (class 1)
  - Precision: 0.76
  - Recall: 0.67
  - Accuracy: 0.7132
- Precision-Recall tradeoff
  - Increase in FN(495)
  - Decrease in FP(305)
- Increase in FN may have negative effects a company

```
Accuracy: 71.82%

Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.72      0.70      1285
           1       0.75      0.71      0.73      1504

    accuracy                           0.72      2789
   macro avg       0.72      0.72      0.72      2789
weighted avg       0.72      0.72      0.72      2789
```



Feature Importances



Example Employee

Feature	Value	Feature	Value
Age	0.902	Leadership Opportunities	0.000
Years at Company	0.513	Innovation Opportunities	0.000
Monthly Income	0.655	Job Satisfaction (ord)	3.000
Number of Promotions	0.000	Performance Rating (ord)	2.000
Overtime	1.000	Work-Life Balance (ord)	1.000
Distance from Home	0.724	Education Level (ord)	0.000
Number of Dependents	0.333	Job Level (ord)	0.000
Gender: Female	1.000	Company Size (ord)	1.000
Job Role: Healthcare	1.000	Company Reputation (ord)	3.000
Marital Status: Married	1.000	Employee Recognition (ord)	0.000
Remote Work	0.000		

Actual Label: Left (0)

Logistic Regression	Decision Tree	Random Forest	SVM
Predicted: Left	Predicted: Left	Predicted: Left	Predicted: Left

Key Takeaways

1. **Support Early-Career Employees:**
   Lower job levels are the strongest predictor of attrition
   a. Invest in career development, mentorship, and promotion paths.
   b. The higher the seniority the less likely one may leave.

2. **Expand Remote Work Options:**
   Employees without remote flexibility are more likely to leave
   a. Offer hybrid/remote policies
   b. Improve work/life balance if remote work isn't feasible

3. **Address Commute Burden:**
   Long commute times increase quit risk
   a. Offer location flexibility or commuter benefits.


The End