Predicting Employee Attrition

Jyoti Kumari, Sarah Lee, Hrishi Salitri,
Vannie Sung, Milan Vaghani

Agenda

Problem Introduction

Exploratory Analysis

Modeling

Attrition insight

Companies need to understand employee attrition

Employee Attrition Dataset

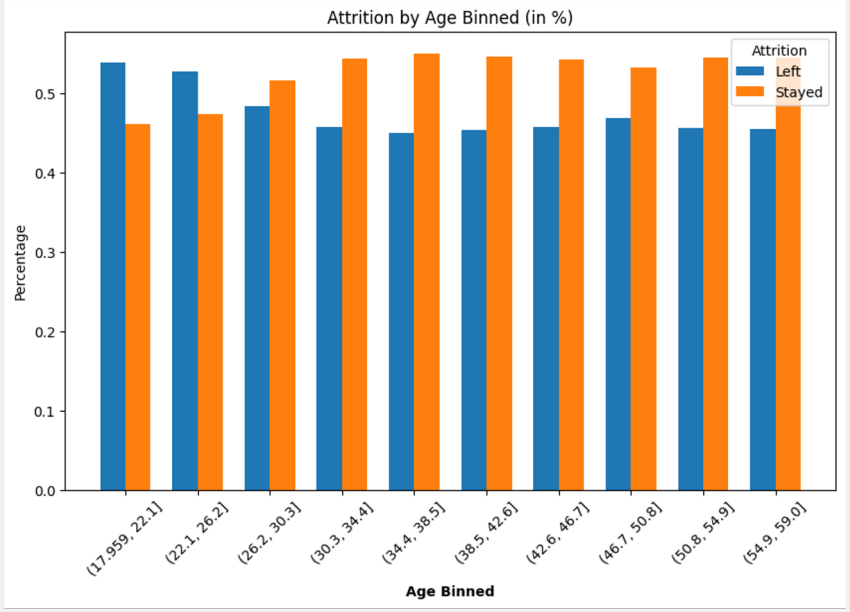
- ~ 70,000 rows of data
- 23 predictor variables
- Target variable: Attrition

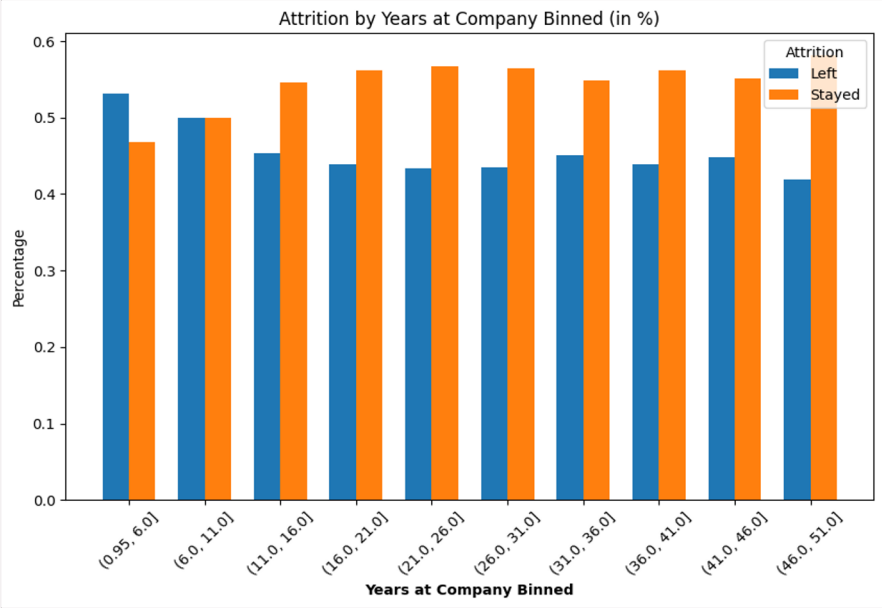
52% Stayed

Exploratory Data Analysis(EDA)

Age vs Attrition

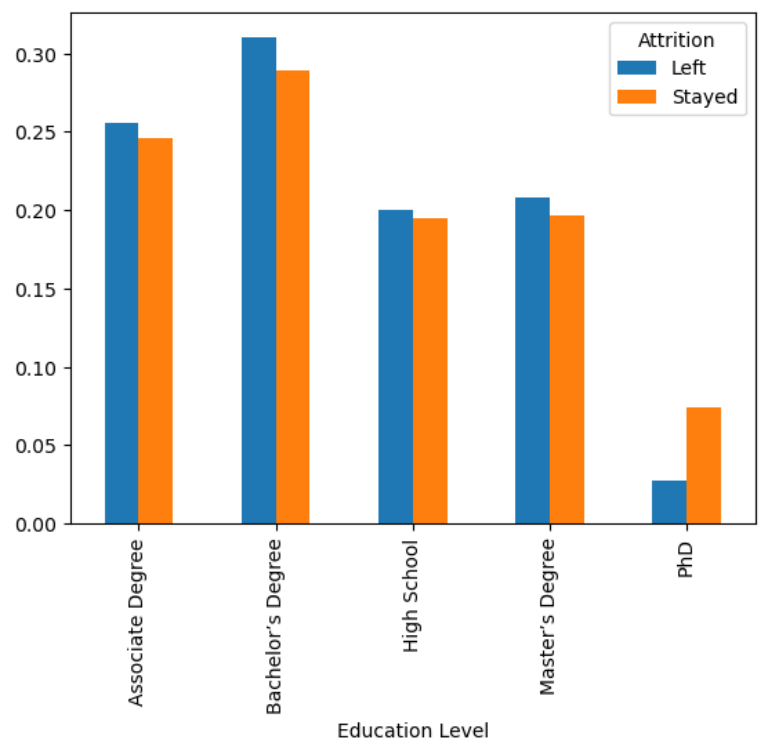
- Older employees stay longer compared to younger employees who leave.
- Older workers seek stability, resulting in longer company tenure.

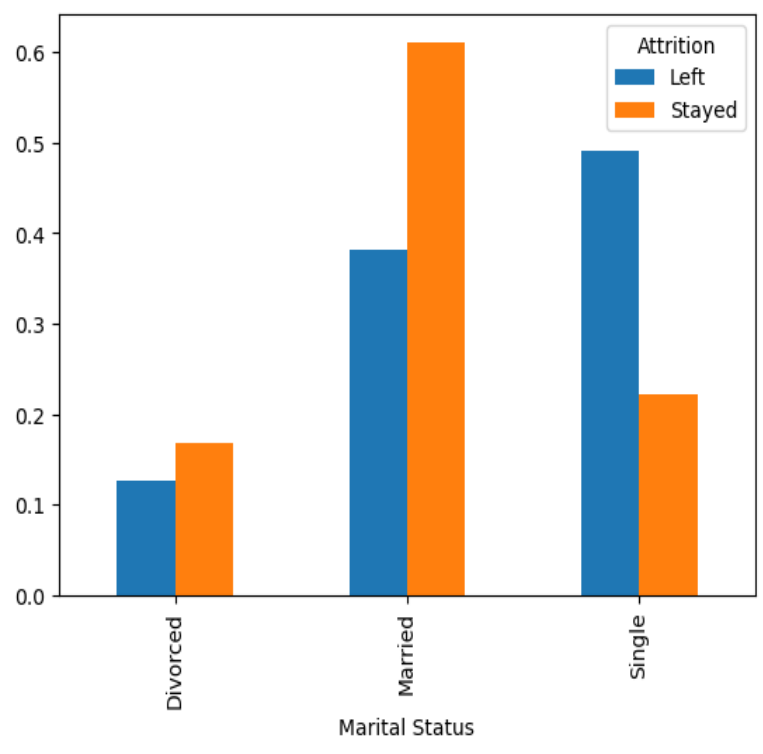


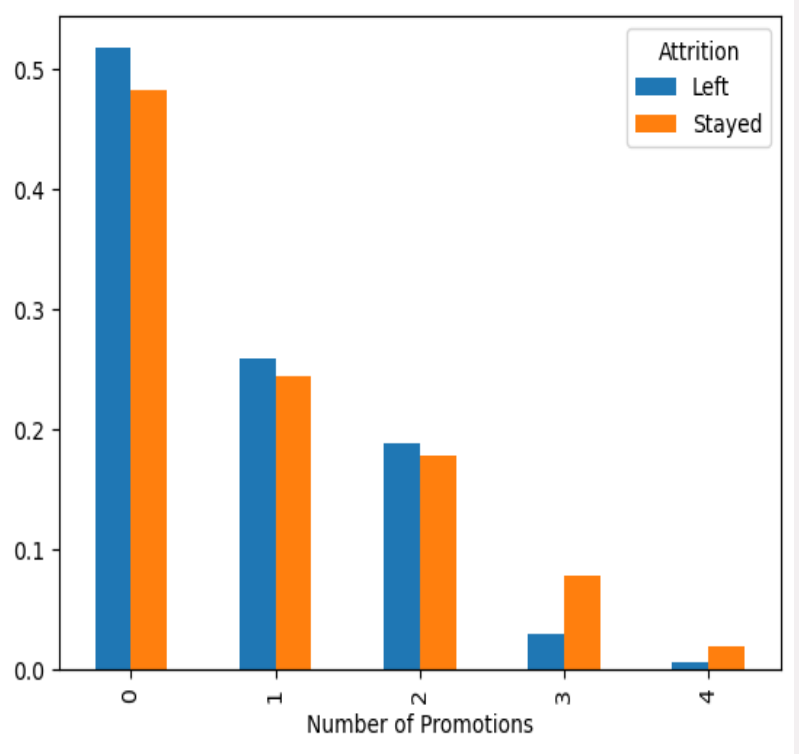


Years at Company vs Attrition

- Employees with higher tenure tend to stay
- As these employees are invested in the company's growth and have ESOPs they tend to stay







Education Level vs. Attrition

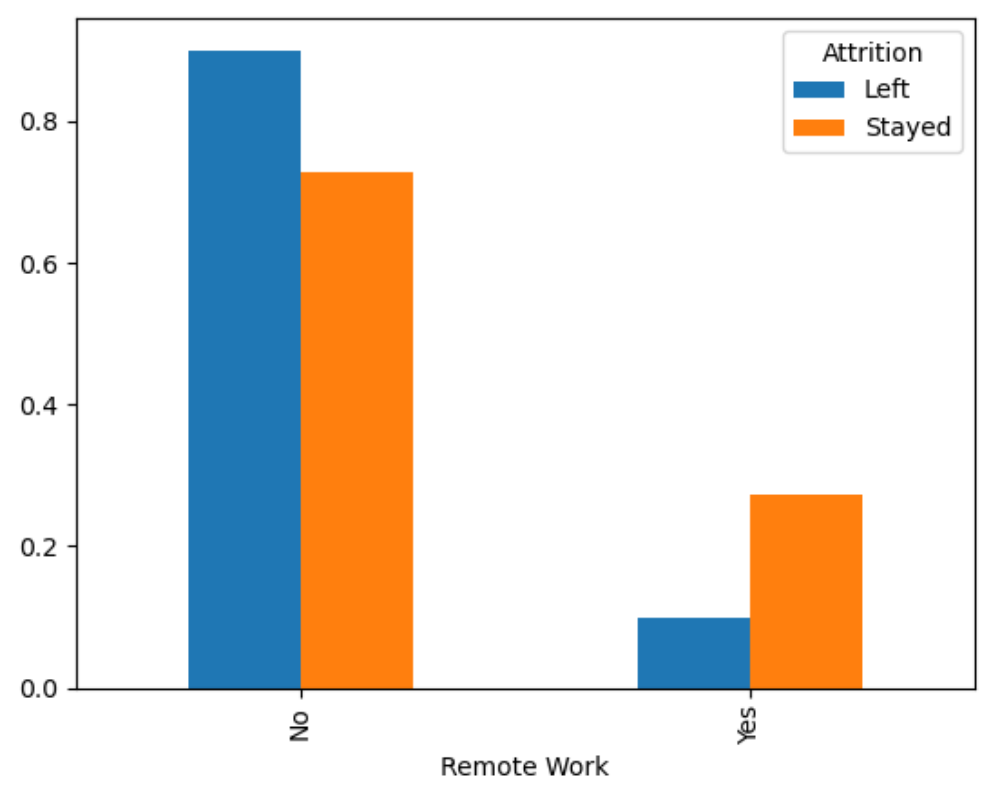
- More education led to more loyalty
- Employees with a bachelor's degree exhibit the highest attrition rate

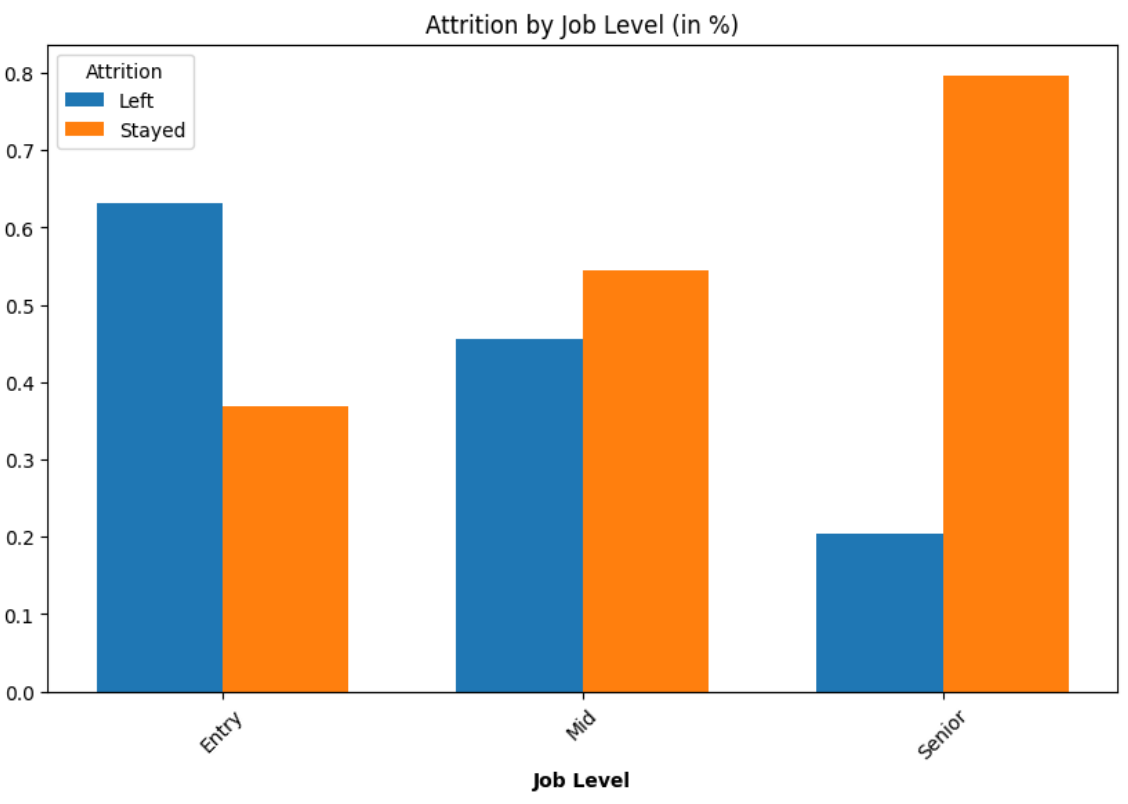
Marital Status vs. Attrition

- 'Single' tends to leave company frequently
- 'Married' tends to look for more stability, hence more likely to stay

Promotion vs. Attrition

- No big impact until a certain threshold
- 75% of employees that stayed had at least 3 promotions.





Work Environment vs. Attrition

- Companies offering more flexible work-from-home arrangements have lower attrition rates

Job Level vs. Attrition

- Higher up an employee was at the company, the less likely he or she are to leave

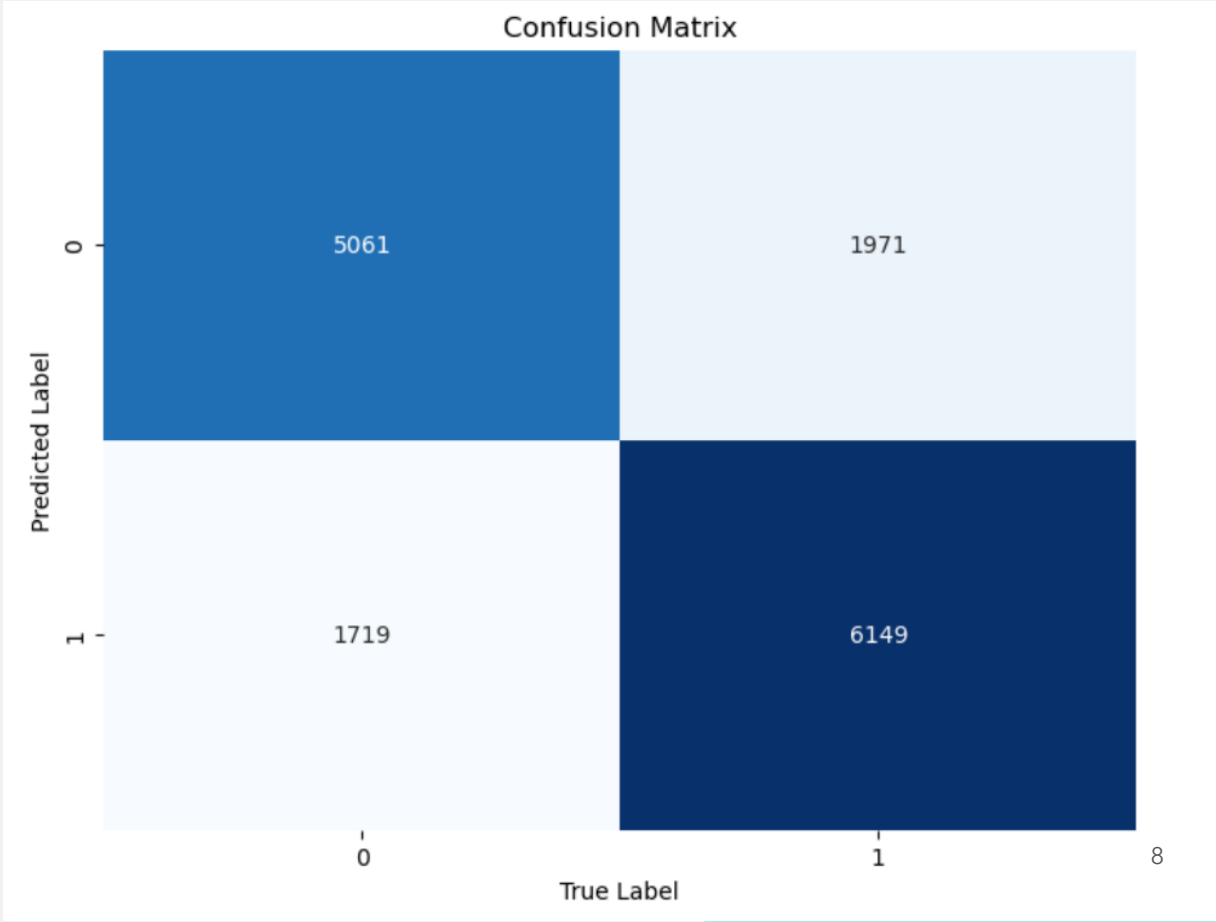
Model#1: Naïve Bayes

Variable Binning for Analysis:

- Continuous variables (Age, Monthly Income, Distance from Home, Company Tenure) were binned into 5 quantiles.
- Years at Company was binned into specific intervals to capture key career stages and tenure-related trends.

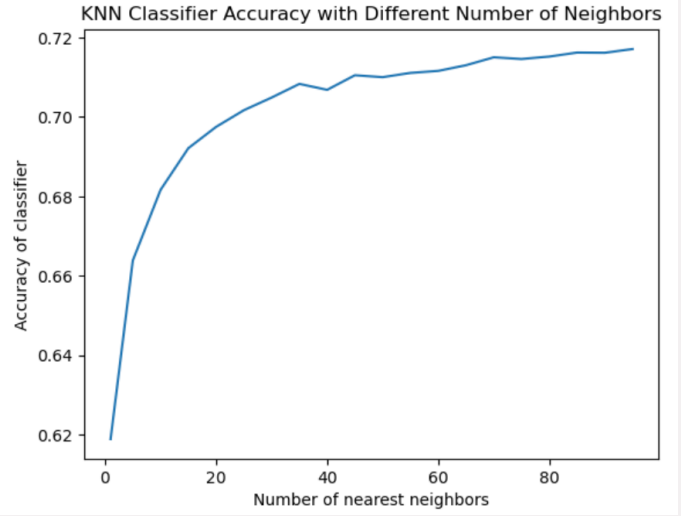
Model Performance:

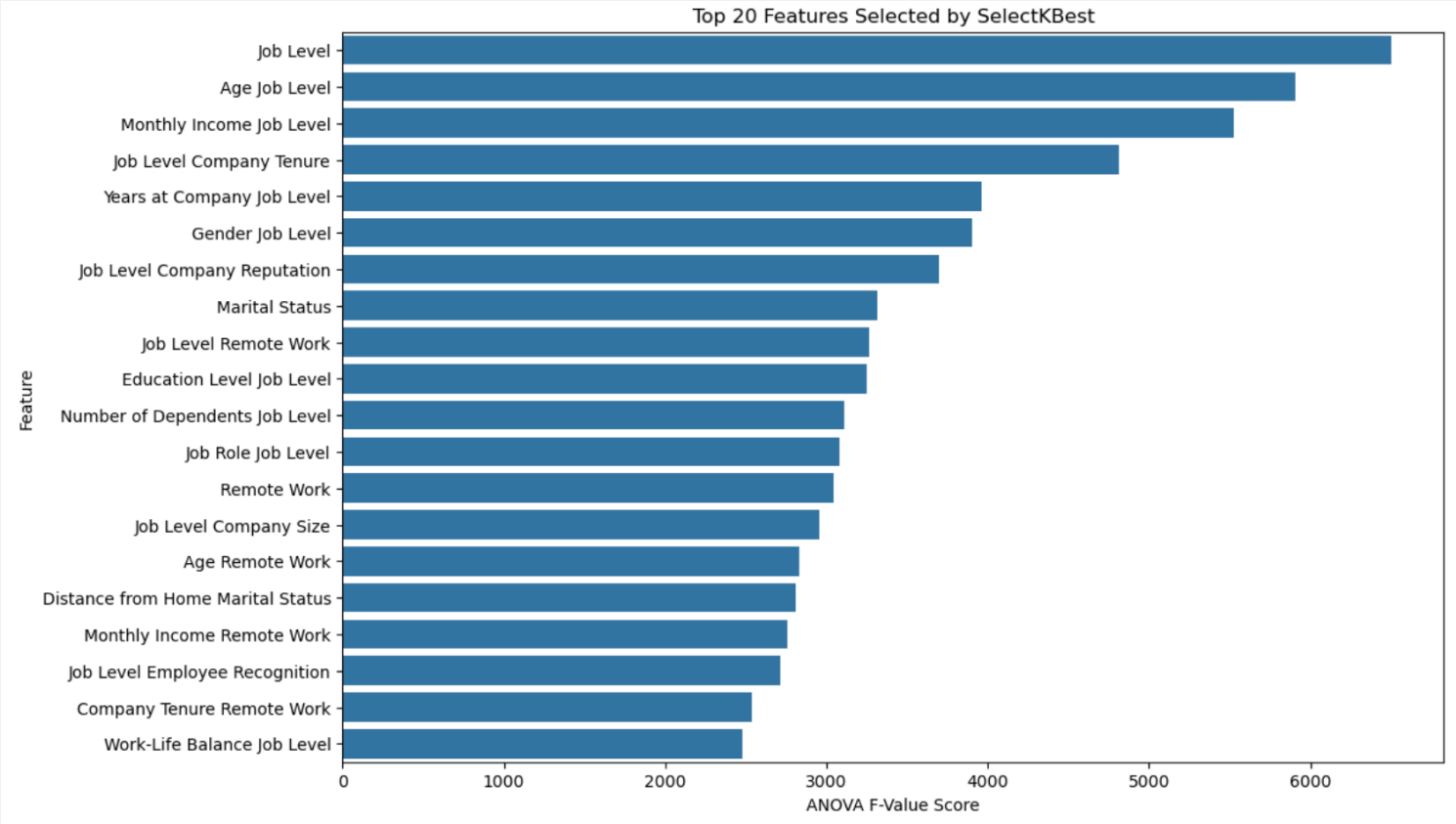
- Accuracy: 75%



Model #2: KNN

	Accuracy
Initial model	67.3%
Cross Validation	71.4%
Selected Features	71.8%





Important features

- Job level
- Age

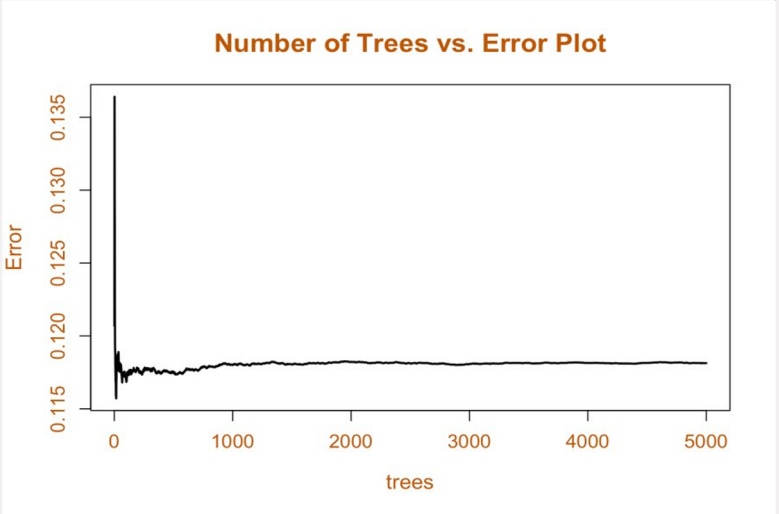
Optimal Number of Neighbors

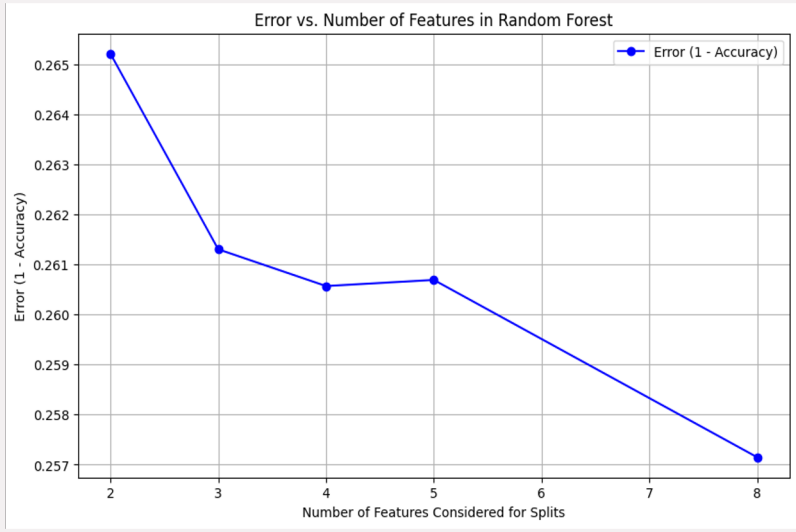
- K=95 for initial cross validation
- K=50 for selected features

Model 3: Random Forests

Number of Trees Selection

- Used error vs number of trees plot to determine optimal tree count to
- The error stabilizes starting with 1000 trees
- The selected tree count is 1500, which balances accuracy and computational efficiency





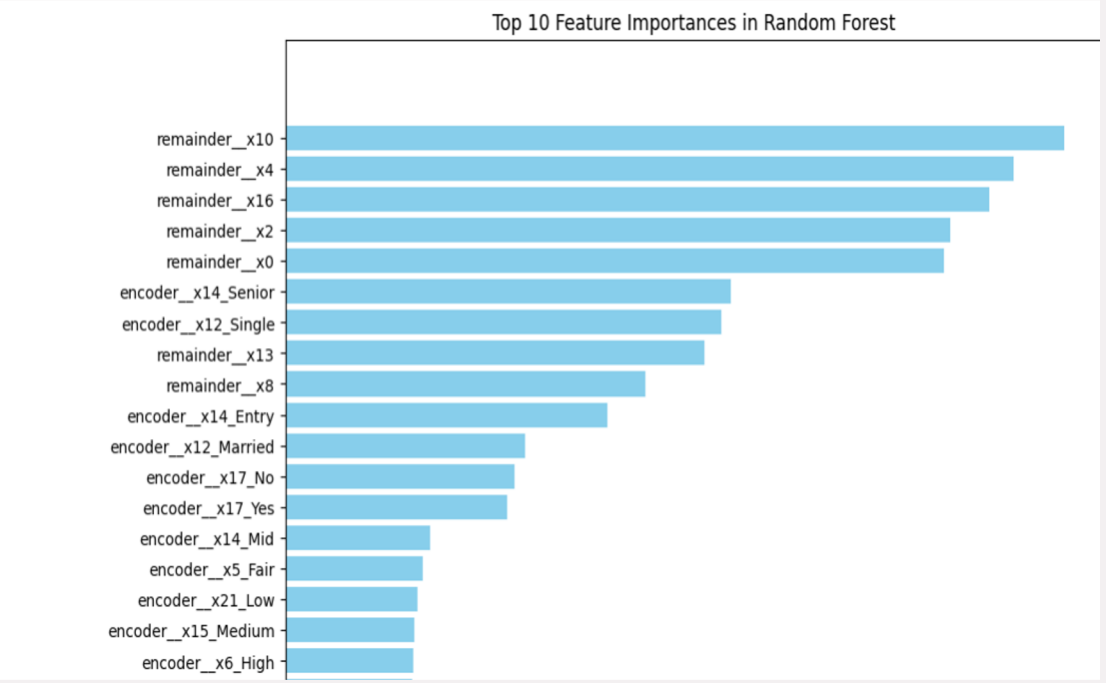
Number of Features

- Used error vs number of features plot to determine optimal feature value
- Selected 8 features to be sampled in random forest calculation

Model 4: Random Forests

Feature Importance

- As per Random Forest variable importance plot the top features were as same as expected in EDA:
  - Age
  - Years at Company
  - Job Level
  - Marital Status

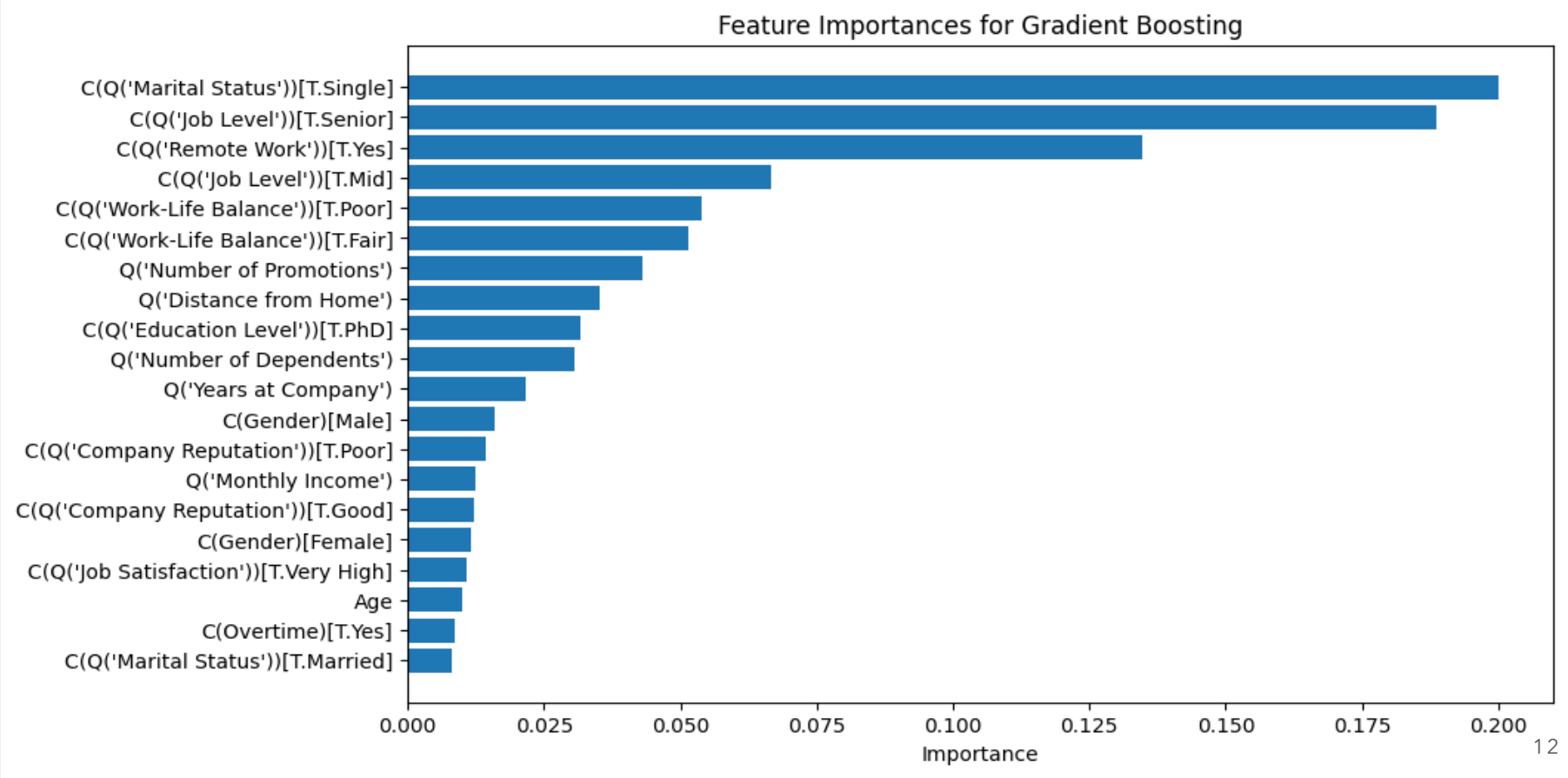


Baseline Accuracy	52.81%
Model Accuracy	75.2%
Precision	76%
Recall	76%

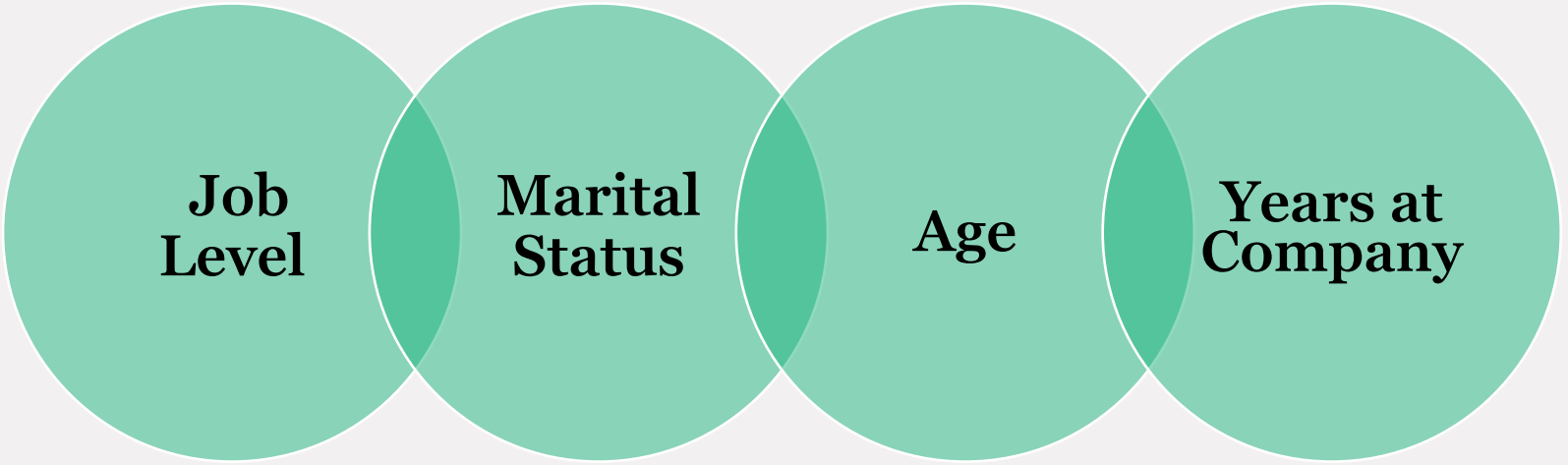
Model 5: Boosting

Accuracy	75.7%
Precision	76.5%
Recall	76.8%

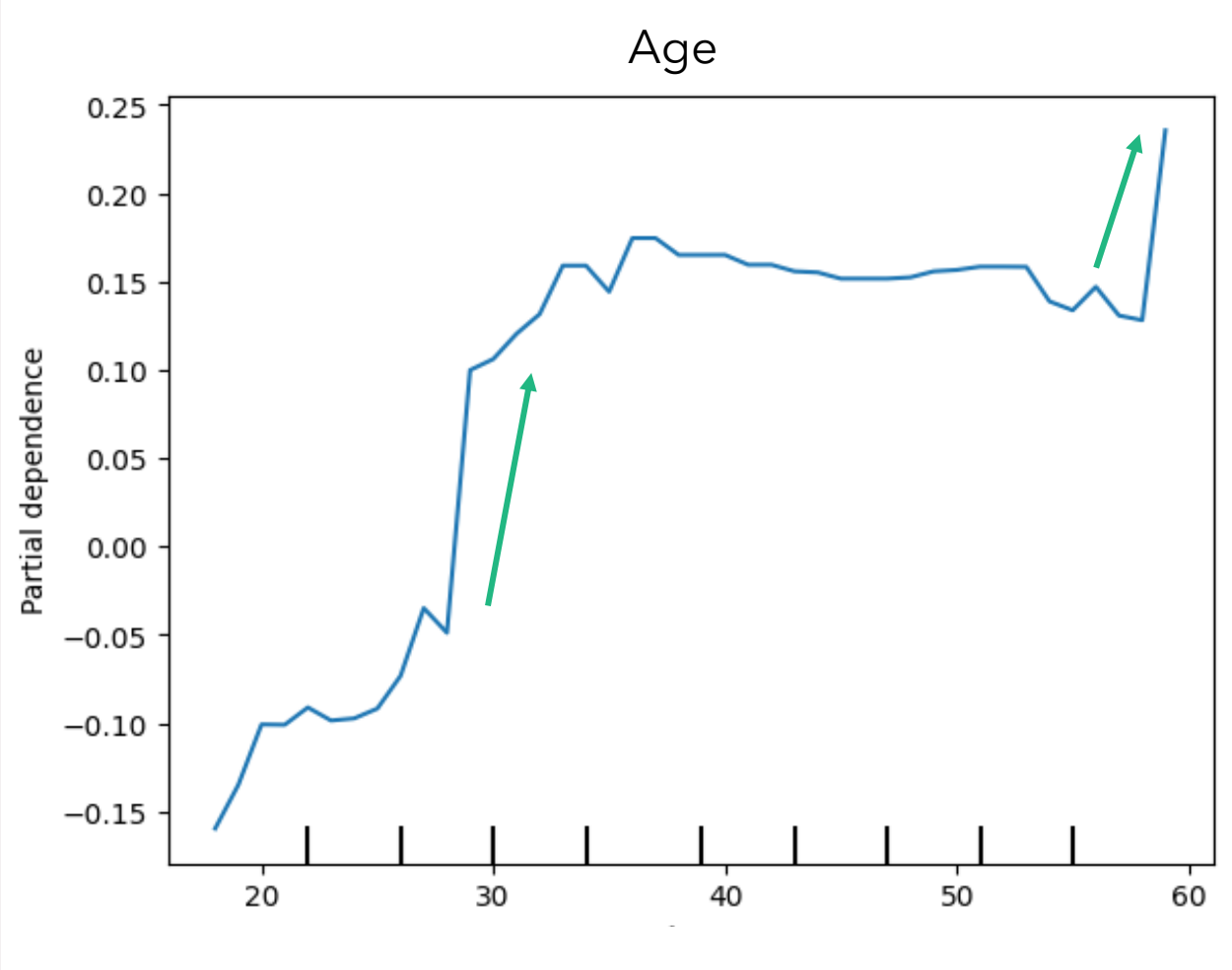
- Marital Status, Job Level, Work Environment
- Important predictors not consistent

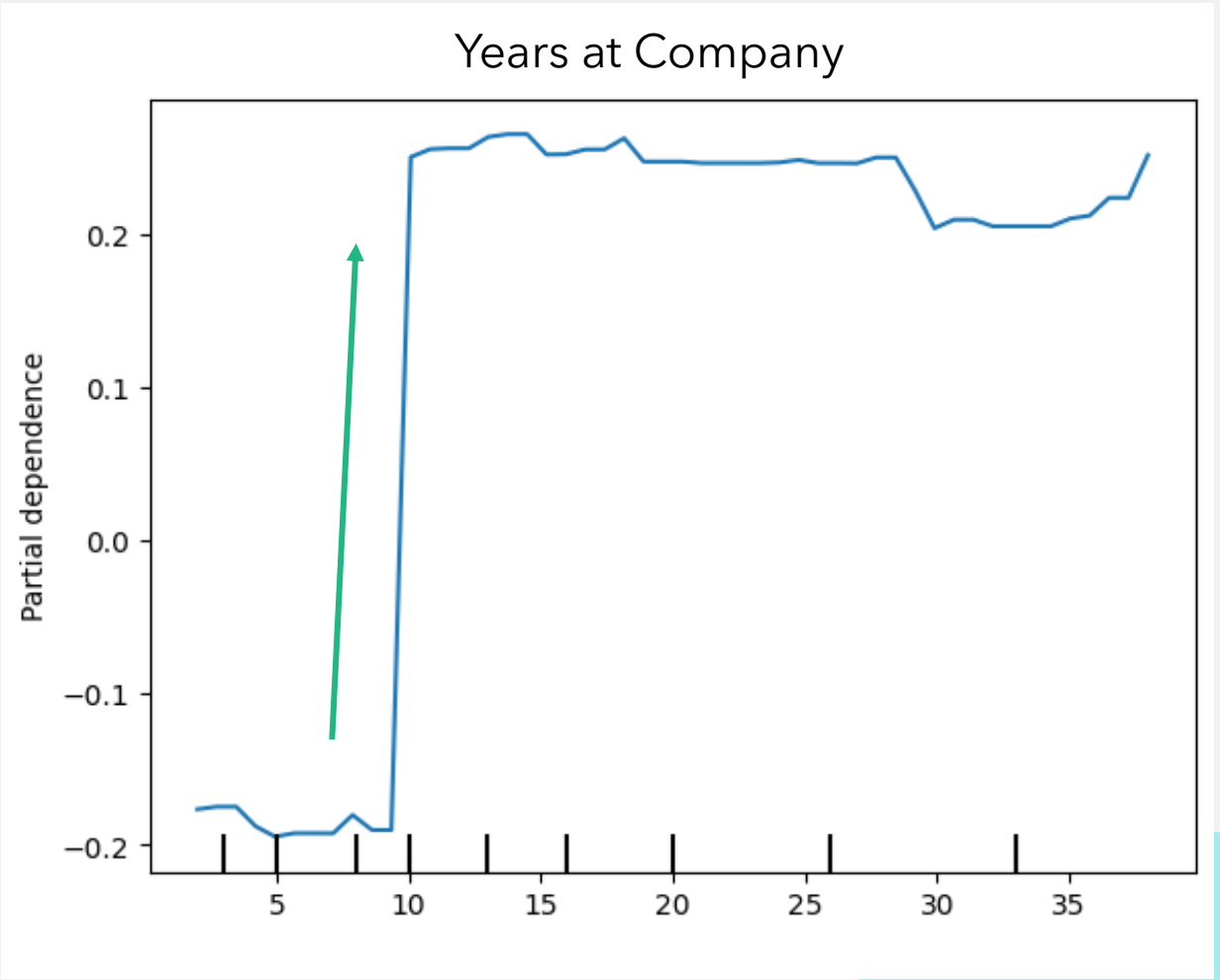


Best Predictors of Attrition



Partial Dependence





Thank You!