Analyzing Employee Attrition

36-602 Final Project

Caleb Peña, Daniel Nason

Executive Summary

- *Introduction*: Analyze employee attrition by building a model to predict whether or not an employee will leave
- *Data*: Investigate variables related to employee attrition using data from HR (Kaggle dataset)
- *Methods*: Fit multiple model types and used EDA and cross validation to identify appropriate variables to include and choose appropriate values of the hyperparameters
- *Results*: Logistic regression performed best both in terms of accuracy and sensitivity for predicting attrition
- *Next Steps*: Automate analysis and develop mitigation strategies for employees with high risk of attrition

Introduction/Motivation

- Why are we here?
  - Since the start of the pandemic, our company's turnover has exceeded the historical average from the previous 10 years
  - HR has tasked us with identifying staff at a higher risk of leaving to determine if further intervention is necessary and worthwhile to reduce this risk
- What is the goal of the project?
  - Build a predictive model that correctly identifies whether an employee is about to leave
  - Discuss next steps for how to reduce attrition

Data - Where is it from and what is it like?

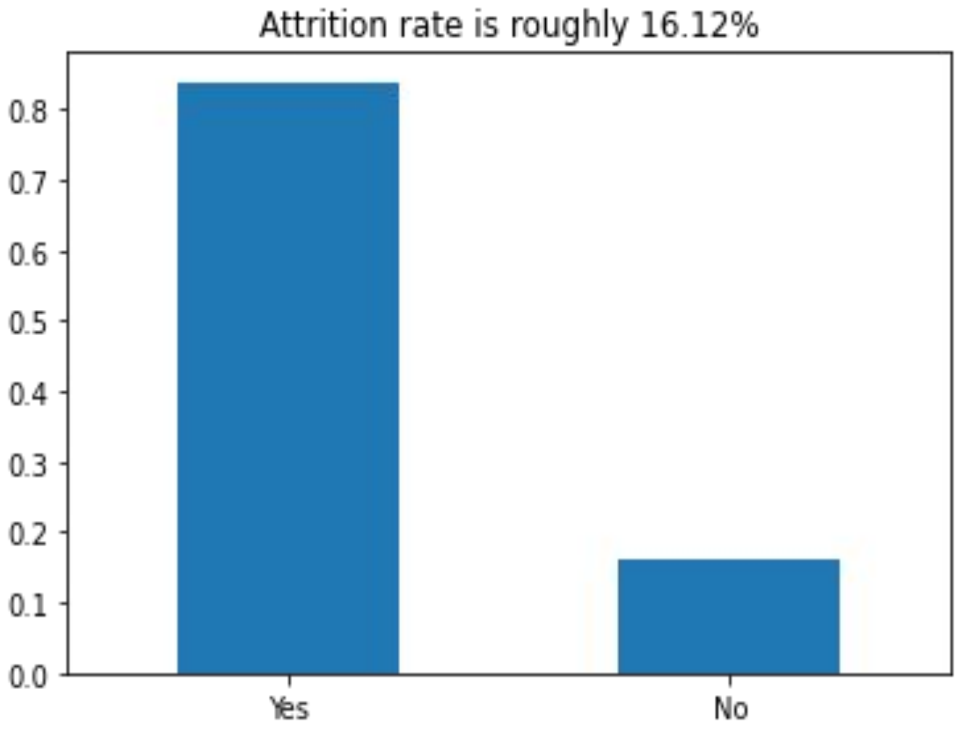
Data

- Collected by HR for a sample of the workforce
- 1,470 employees with 35 attributes related to those employees (31 used)

Attributes

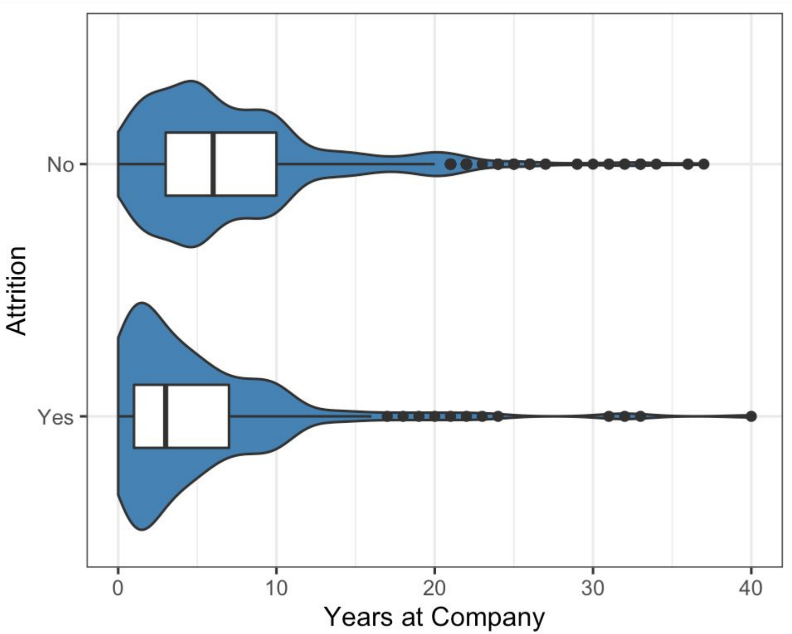
- Variable we want to predict: Attrition (did the employee resign?)
  - 1 if yes, 0 if no
- Variables we consider to predict the response (and examples):
  - Demographics: Age, education, gender, marital status, etc.
  - Compensation: Daily rate, monthly income, stock options level, etc.
  - Job-related items: Job role, job satisfaction, performance rating, etc.
  - Miscellaneous: Total working years, work life balance, etc.

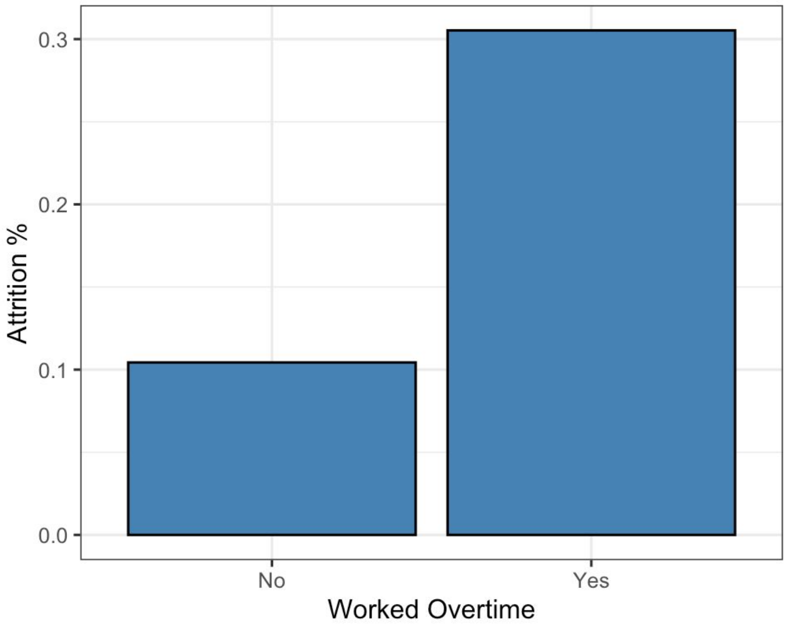
Exploring the response variable: Attrition



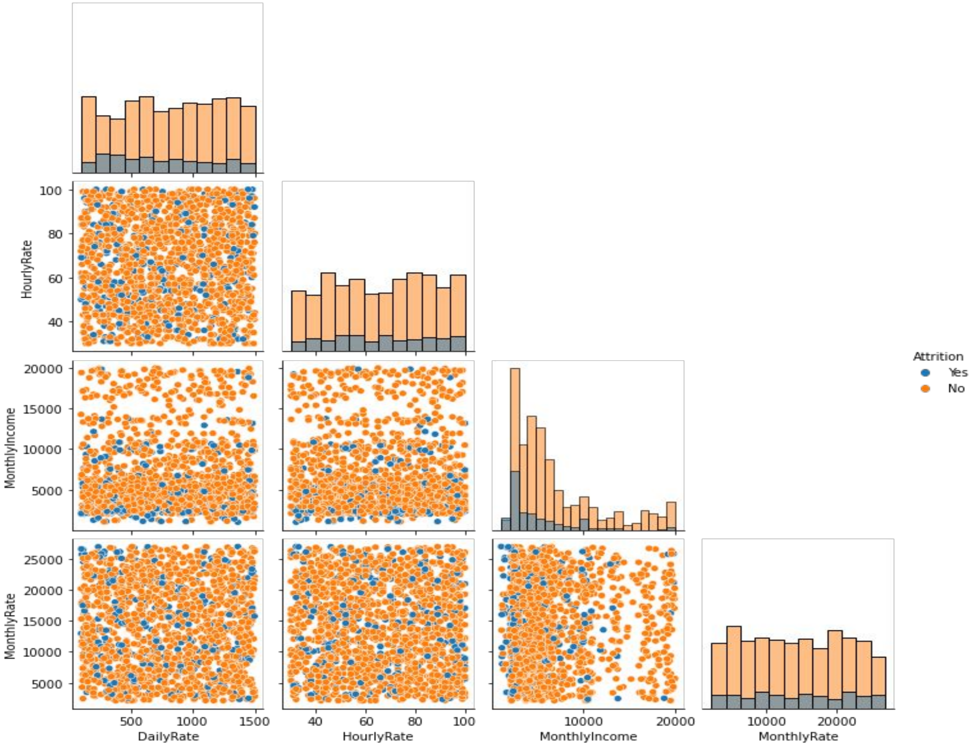
- Binary outcome suggests classification analysis is appropriate
- Imbalanced classes: non-attrition (Yes) is the more dominant occurrence in the data

Employees with shorter tenures and those who work overtime are more likely to leave





Income measures display no clear relationship with attrition



Histograms (diagonal plots):

- Distribution for each variable is not influenced by whether the employee departs

Scatterplots (off-diagonal plots):

- Interestingly, no obvious relationship between income measures

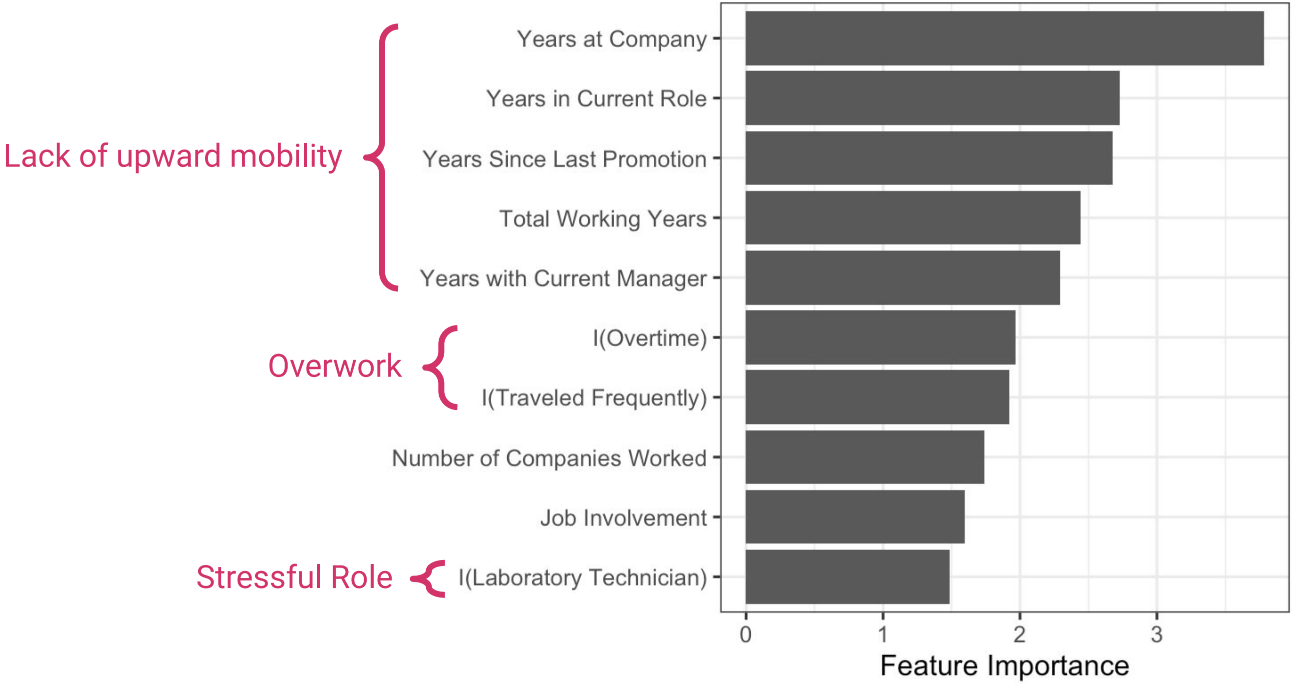
Methods

- Since identifying at-risk employees is our primary goal, we mainly considered ML predictive engines
    - Random Forests, XGBoost, and Logistic Regression
- Due to the noisiness identified in the EDA stage, we omitted measures of income from our analysis
- We selected models using prediction accuracy
    - If models reported similar levels of accuracy, those with higher sensitivity were preferred
- Model hyperparameters were chosen using grid search and 5-fold cross validation

Logistic Regression was selected based on best performance in both accuracy and sensitivity

Model Performances on Testing Data			
Model	**Accuracy**	**Sensitivity**	**AUC**
Baseline	83.88	0	0.50
Logistic Regression	86.36	39.76	81.04
Random Forest	85.17	27.59	78.53
XGBoost	85.25	33.65	78.13

Results - Feature Importance*



* To estimate feature importance we applied min-max normalization to the data and refit the model. The estimated coefficients tell us the relative importance of each variable.

Insights and Limitations

Insights

- The distribution of the income variables suggests the presence of some kind of data abnormality (*Note: Per Kaggle, this data is not real but simulated*)

- Factors influencing attrition: length of time at the company, presence of opportunities to advance, and stress of the position

Limitations

- For certain job roles, the model struggles to correctly identify whether an employee will leave
    - Accuracy of the model is substantially lower for sales representatives (77.3%) and lab technicians (80.3%) than for the remaining positions

Next Steps

- Automating analysis to create a monthly report for HR to identify potential high risk employees
- Develop mitigation strategies for talent flagged as likely to leave
  - Plan can include changes in compensation, travel, OT, work-life balance etc.
- Inference: identify variables and quantify relationships that could help to determine which employees will risk
  - Avoid putting employees in these situations to retain top talent

Thank you!

Time Spent: Expected versus Actual

Phase	**Subtask**	**Expected (Dan)**	**Actual (Dan)**	**Expected (Caleb)**	**Actual (Caleb)**
1: Preparation	Aggregate & Ingest Data	1 hour	0.5 hours	*NA*	*NA*
	Clean Data	*NA*	*NA*	1 hour	0.5 hours
	Perform EDA	2 hours	1.5 hours	2 hours	1.5 hours
2: Build models		3 hours	5 hours	3 hours	3 hours
3: Extract insights		2 hours	2 hours	2 hours	1.5 hours
4: Generate work product		1 hour	2 hours	1 hour	2 hours
Total		10 hours	11 hours	10 hours	8.5 hours