Attrition Analysis

Alistair Garioch

Attrition Analysis Overview

Unwanted attrition leads to loss of talent & high costs of replacing employees.

HR teams are piloting machine learning techniques to predict the likelihood of employees leaving so they can take pro-active measures.

Take away our top 20 employees and we become a mediocre company

Bill Gates

The only way for businesses to consistently succeed is to attract & retain smart creative employees

Eric Schmidt

Project Outline

Objective

Build a binary classification model using employment factors (e.g. ‘time in role’, ‘age’, ‘job title’) to predict whether an employee will quit

Secondary

Understand which employment factors are strongest predictors of attrition, e.g. would investing in training or benefits reduce attrition?

Data

Attrition Analysis Dataset from IBM Watson Data Science Team

Evaluation

Recall: ability to correctly identify attrition
Area under ROC curve: true positives versus false positives

Data Overview

1,470 rows (employees), 35 features | Target: attrition (16% yes, 84% no)

Continuous	mean	std	median
Age	37	9	36
DailyRate	802	404	802
EmployeeNumber	1,025	602	1,021
MonthlyIncome	6,503	4,708	4,919
MonthlyRate	14,313	7,118	14,236
NumCompaniesWorked	3	2	2
PercentSalaryHike	15	4	14
TotalWorkingYears	11	8	10
TrainingTimesLastYear	3	1	3
YearsAtCompany	7	6	5
YearsInCurrentRole	4	4	3
YearsSinceLastPromotion	2	3	1
YearsWithCurrManager	4	4	3
DistanceFromHome	9	8	7

Categorical	values		
Attrition	Yes / No		
BusinessTravel	Non_Travel / Travel_Rarely / Travel_Frequently		
Department	Sales / R&D / HR		
EducationField	5 fields (e.g. Life Sciences / Medical)		
Gender	Female / Male		
JobRole	9 job roles (e.g. Sales Executive)		
MaritalStatus	Single / Married / Divorced		
OverTime	Yes / No		
Ordinal	**mean**	**min**	**max**
Education	3	1	5
EnvironmentSatisfaction	3	1	4
JobInvolvement	3	1	4
JobLevel	2	1	5
JobSatisfaction	3	1	4
PerformanceRating	3	3	4
RelationshipSatisfaction	3	1	4
StockOptionLevel	1	-	3
WorkLifeBalance	3	1	4

Modelling Approach

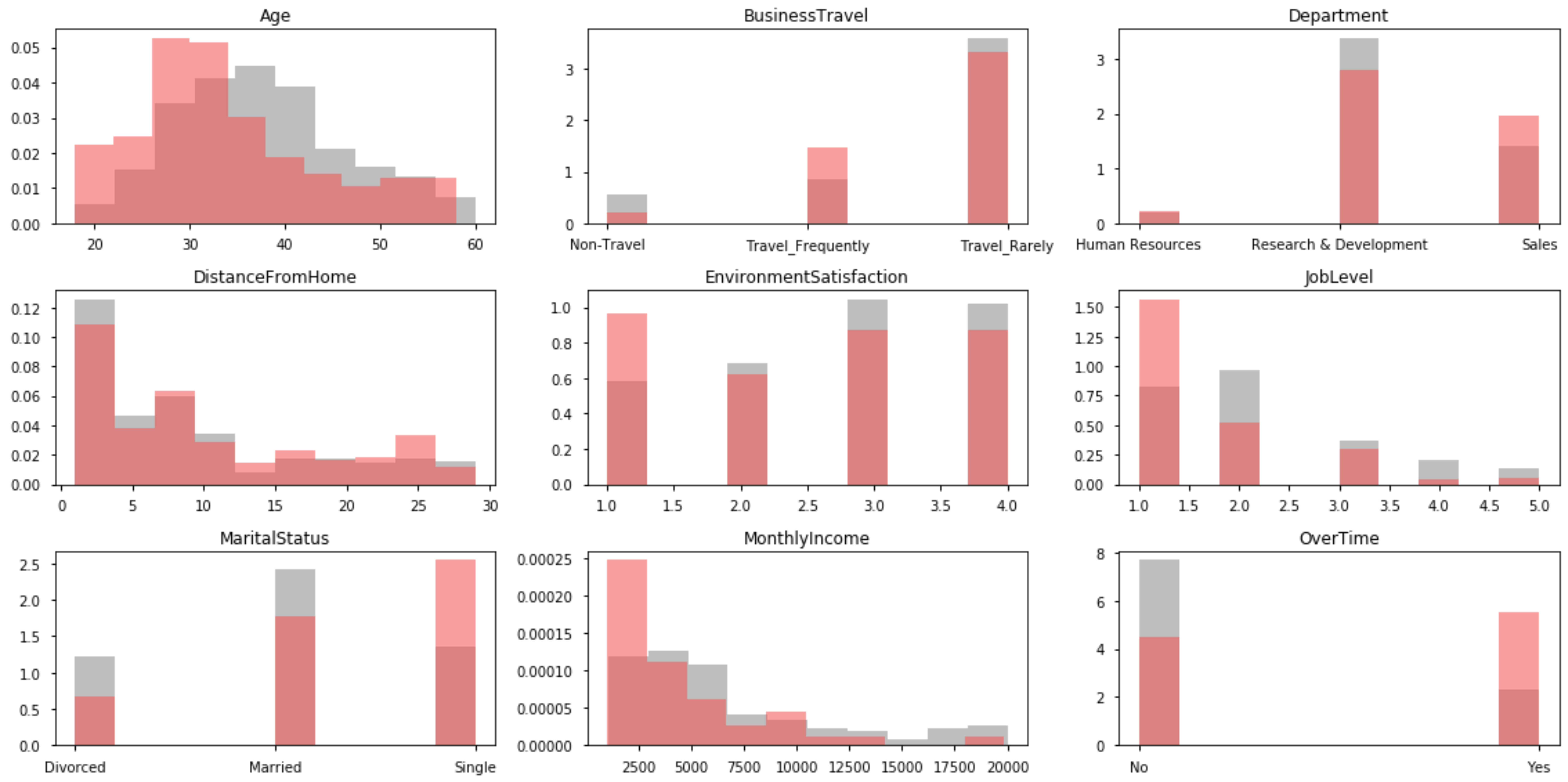
Parse, Mine & Refine Data

- **Literature:** reviewed 4 relevant papers
- **Train/test:** 90:10 train/test split
- **Data dictionary:** created definitions for all fields
- **Statistics:** ran descriptive statistics for all variables
- **Distribution:** compared attrition/non-attrition
- **Data quality:** no missing values and few outliers
- **Categorical:** created 13 dummy variables
- **Feature engineering:** add ‘manager’ feature
- **Relationships:** created correlation matrix
- **Other issues:** drop features with multi-collinearity
- **Feature selection:** created feature matrix for models

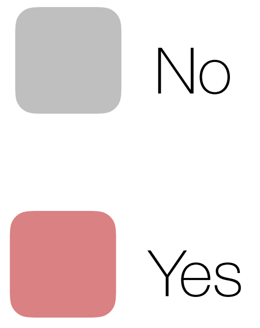
Select, Tune & Test Model

Model	select	tune	test
Decision tree	✔	✔	✔
KNN classifier	✔		
SVM classifier	✔	✔	
Logistic regression	✔	✔	
Naive bayes	✔		
Ada boost	✔	✔	✔
Random forest	✔	✔	
Voting classifier	✔	✔	✔

Insights: Attrition vs. Retention Population Comparisons



Attrition Key:



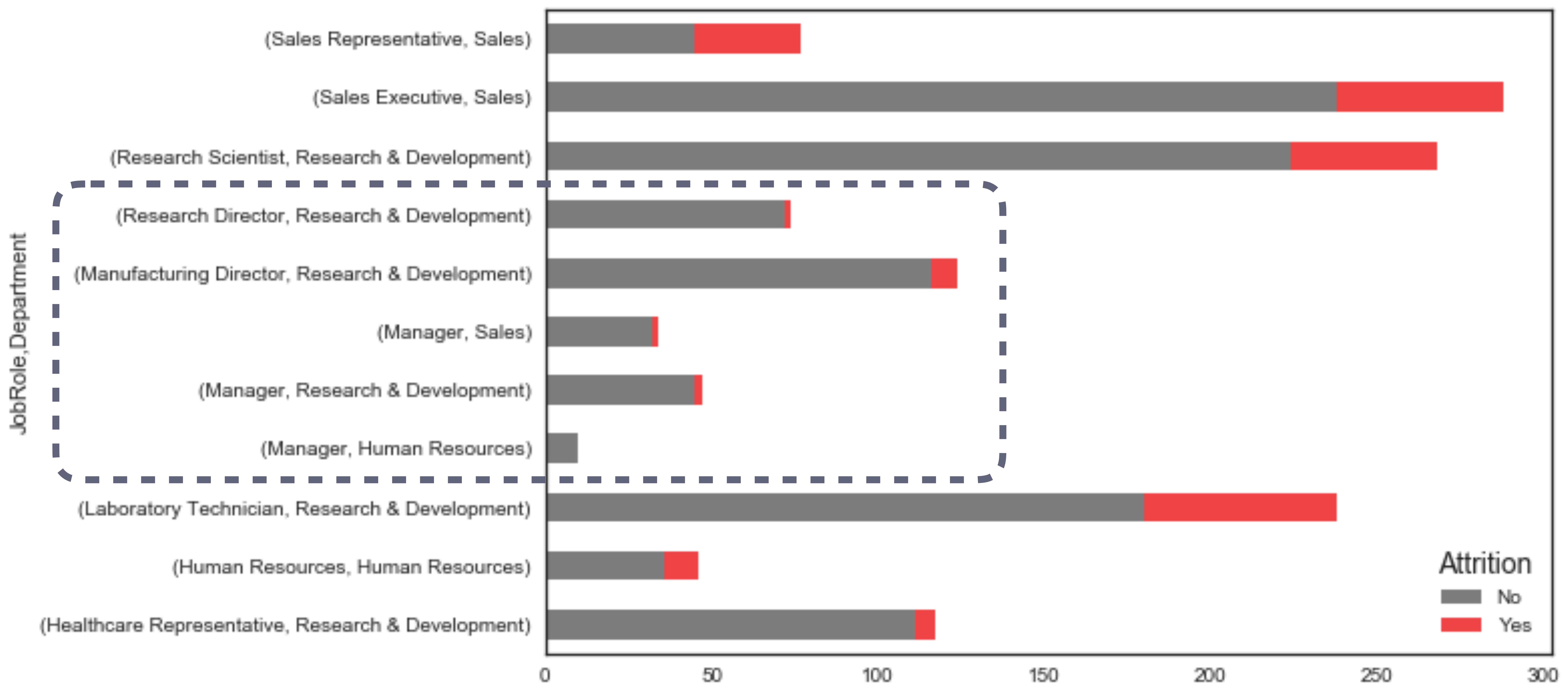
T-Test Note:

p-value <0.04 for all features displayed

Can reject the null hypothesis that the mean for attrition is the same for non-attrition populations at 95% confidence

Insights: Job Roles ➛ Managers/Employees

Create a feature to identify managers & directors. Less likely to quit than others



Managers:

5% Attrition

Non-Managers:

19% Attrition

Insights: Population Comparison Summary

- Employees who work **overtime** are more likely to leave (stress?)
- **Younger, single** employees are more likely to leave (more expendable? more flexible personal situations?)
- **Frequent travel** is linked to attrition (stress?)
- **Sales** has higher attrition than **R&D** (Sales skills transferable, R&D more company specific?)
- Employees **living < 5 miles from office** are more likely to stay (low stress?)
- **Low satisfaction** (job, work-life, environment, involvement, relationships) is linked with attrition
- **Lower job level** or **low income level** or **no stock options** employees are more likely to leave
- **Less time in the role**, with the company or current manager is linked to higher churn

Can we quantify how important these insights are? Can they predict attrition?

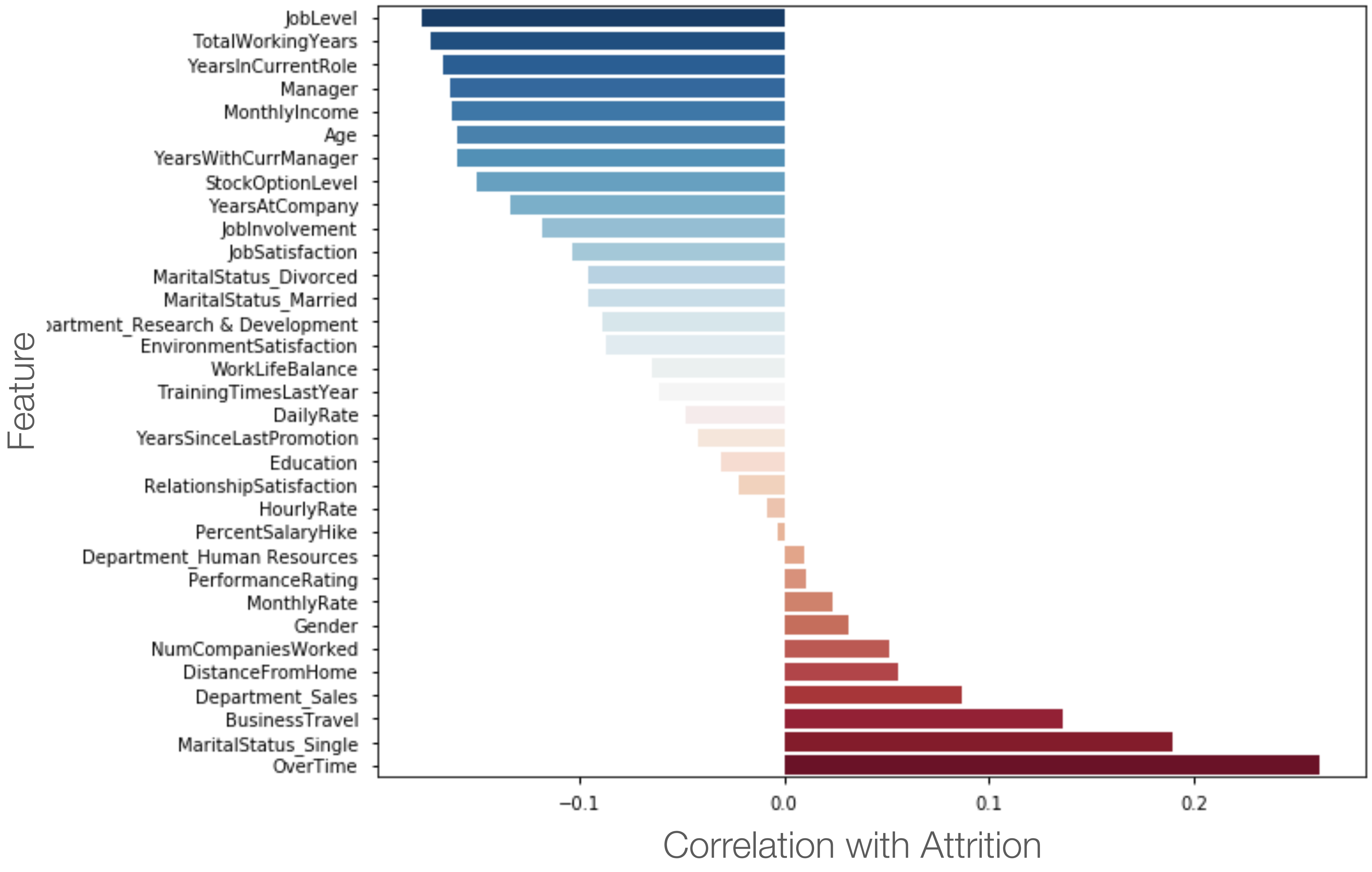
Insights: Correlation Between Features and Attrition

1. No feature > 30% correlation

2. Strongest correlations:
- OverTime … 26%
- Single … 19%
- JobLevel … -18%
- WorkingYears … -17%
- YearsInRole … -17%
- Manager … -16%
- MonthlyIncome … -16%

3. Surprising that pay and benefits are not the #1 driver of attrition

4. Stress (overtime, travel, commute) has strong correlation

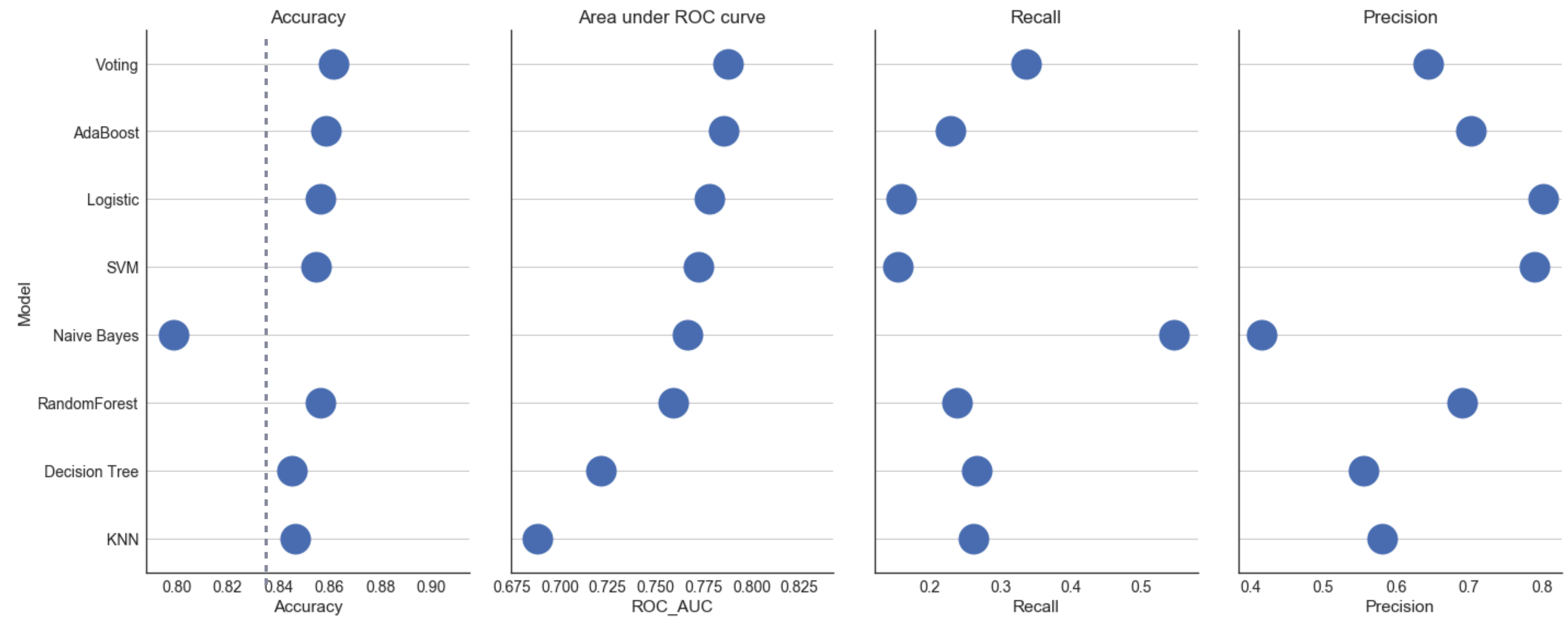


Modelling Approach

1 Evaluate multiple models with cross validation

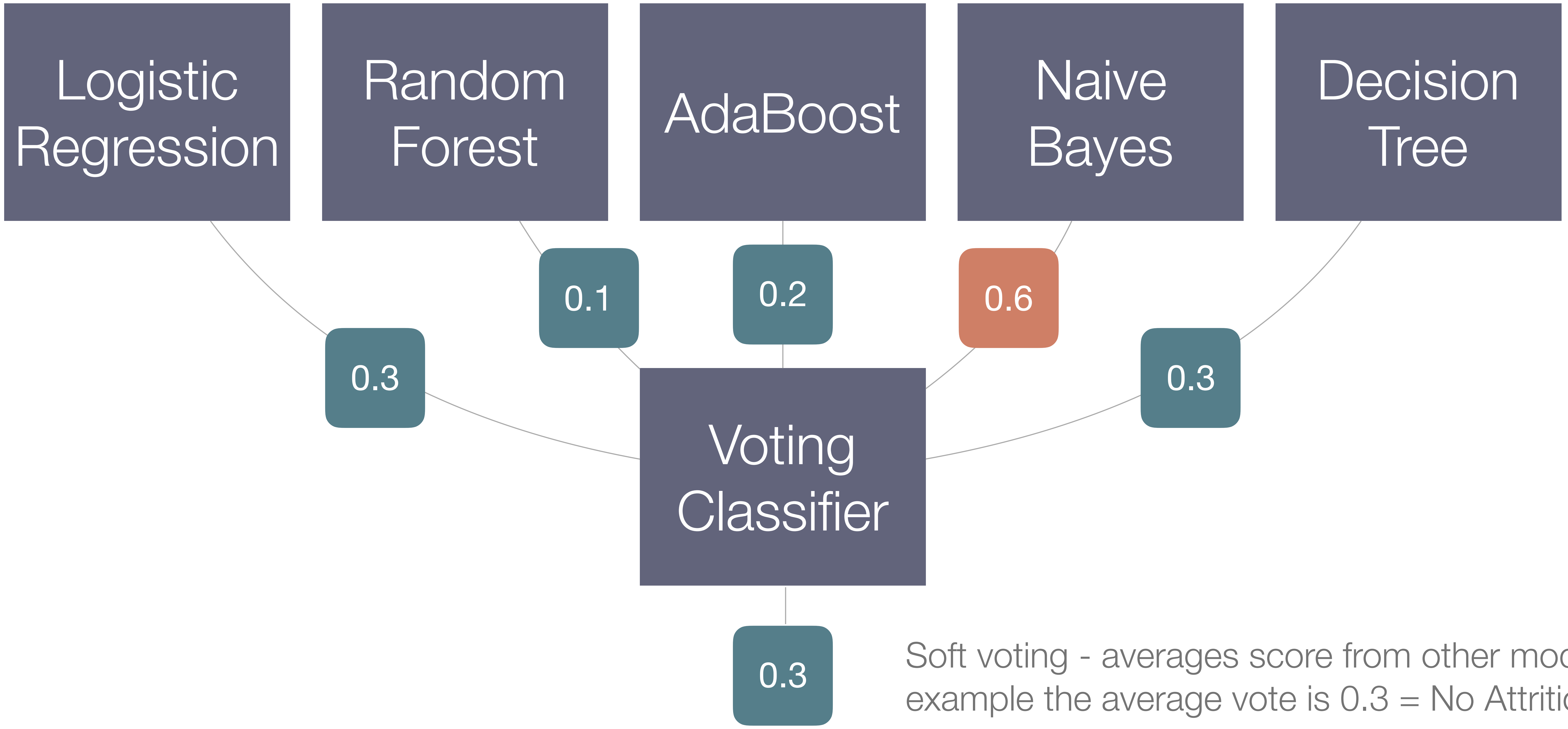
2 Select top models, tune parameters via grid search

3 Select final model and test with holdout dataset



Voting Classifier

The best results came from combining the other models in a voting classifier



Soft voting - averages score from other models. In this example the average vote is 0.3 = No Attrition

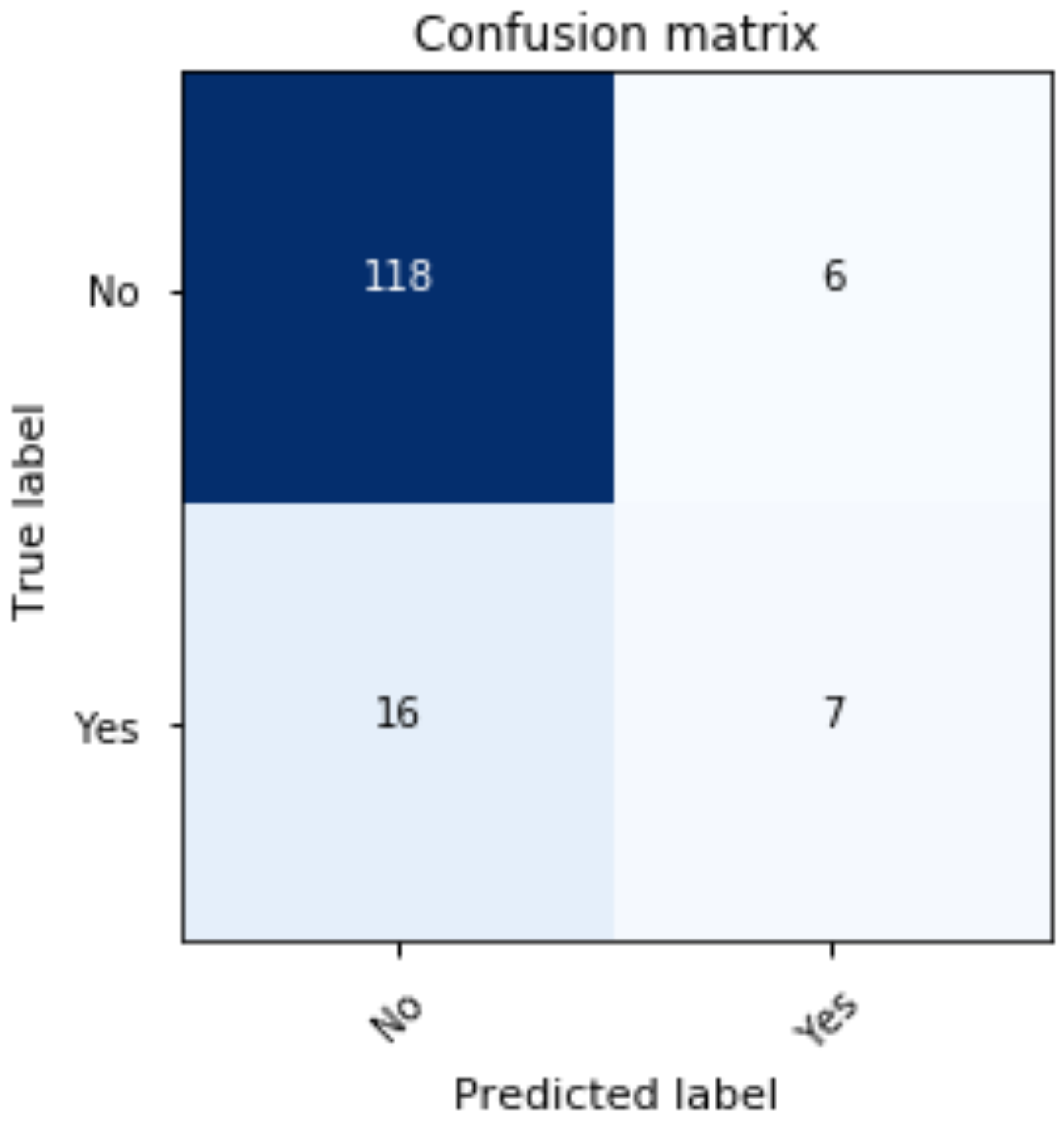
Modelling Results

The **voting classifier** gave the best results under 3 fold cross validation, so it was selected as the final model and tested.

In the 10% data held back for testing there were 147 employees. 23 left the company (16% attrition).

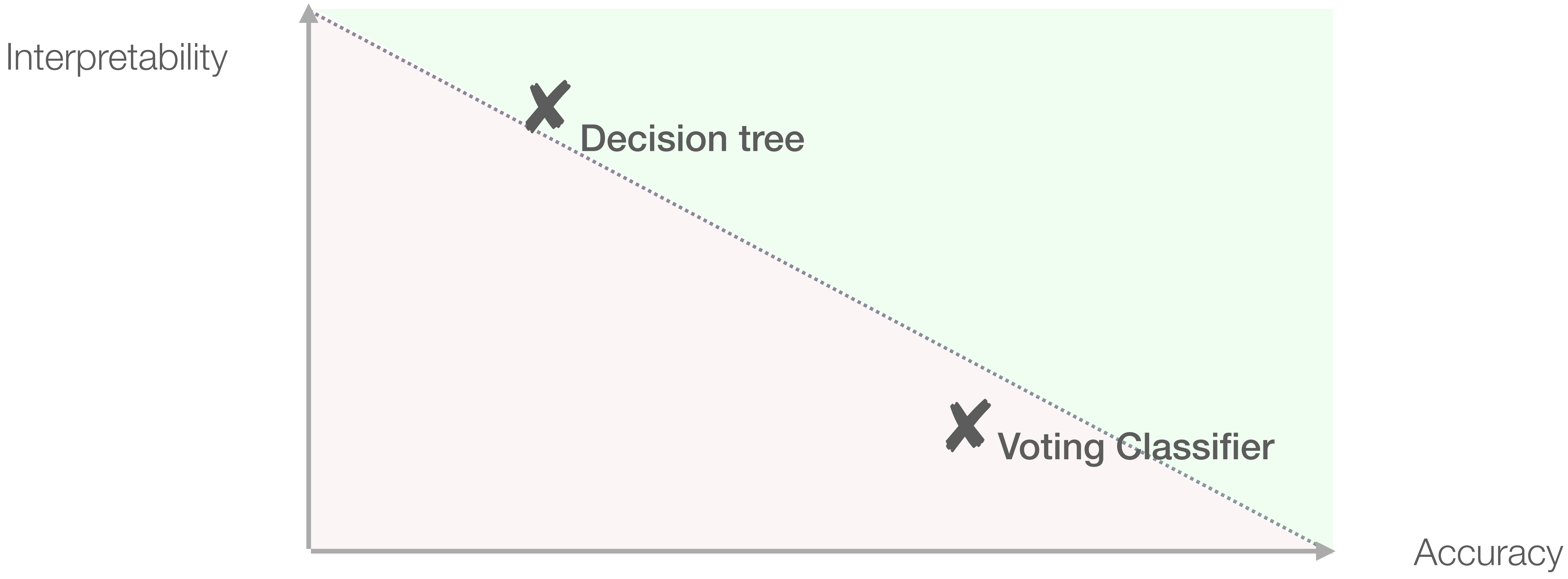
The model identified 7 of those employees (30% recall) but incorrectly labelled 6 employees who remained in the company as attrition (54% precision).

```
Accuracy: 0.850
ROC_AUC: 0.628
Precision: 0.538
Recall: 0.304
```



Modelling Summary

- The most predictive model was the **voting classifier**
- On the other hand, a voting classifier is difficult to interpret and the **decision tree** may provide a better aid to support human led decision making



Conclusions

Accuracy of HR models will always be limited. There are too many important variables we cannot measure accurately e.g. work relationships, family situation, employee personality

That said; analysing attrition using employment data can add insight beyond the traditional approach of tracking an attrition metric by department / manager

Helping employees to manage their workload and stress is likely to be a cost effective way of reducing turnover (free yoga classes? time management seminars? simplify processes?)

Employees newly entering the workforce are more likely to leave than experienced employees, this company should try to understand if they need to adapt their practices for millennials

Next Steps

Larger sample. Model on all company employees vs. only one division

Add features. Capture new features e.g. location, size of team, creative work

Oversampling. Capture or create more ‘yes’ attrition data points

More benchmarking. Lots of companies & researchers looking at this problem