Predicting Employee Turnover with Machine Learning Classification Models

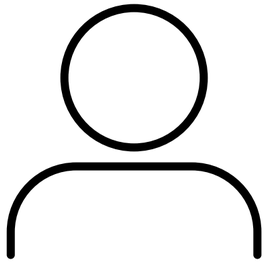
RUBY JANG

Cost of Turnover

Total cost to hire and train a new employee, to the same level of productivity as the employee who left.

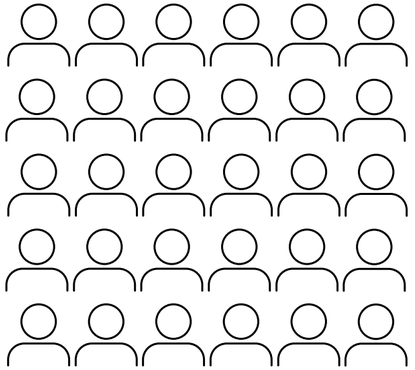
Example

- An employee is paid **$90,000** per year
- Assume it costs **33%** of salary to replace.[1]



1 employee

~$30,000



30 employees

~$900,000

1. Reported by Employee Benefit News (EBN) based on Work Institute US 2017 Retention Report.

BUSINESS QUESTION

Can we **identify** employees who are likely to leave, so we can take **actions** to improve retention and **save costs**?

DATA QUESTION

Can we accurately **predict** whether an employee will leave, by building a machine learning **model** based on company's employee **data**?

BINARY CLASSIFICATION PROBLEM

Logistic Regression
Support Vector Machine (SVM)
Naive Bayes

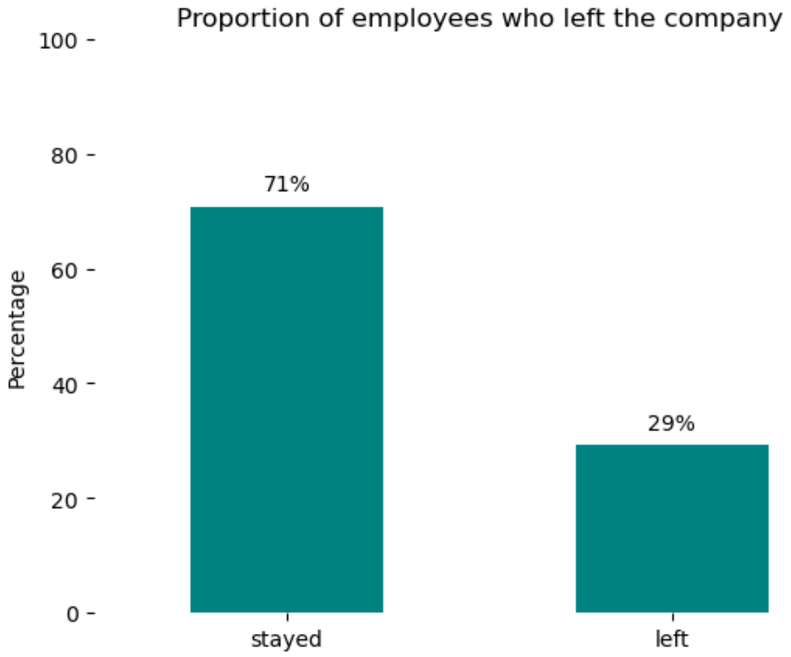
Data

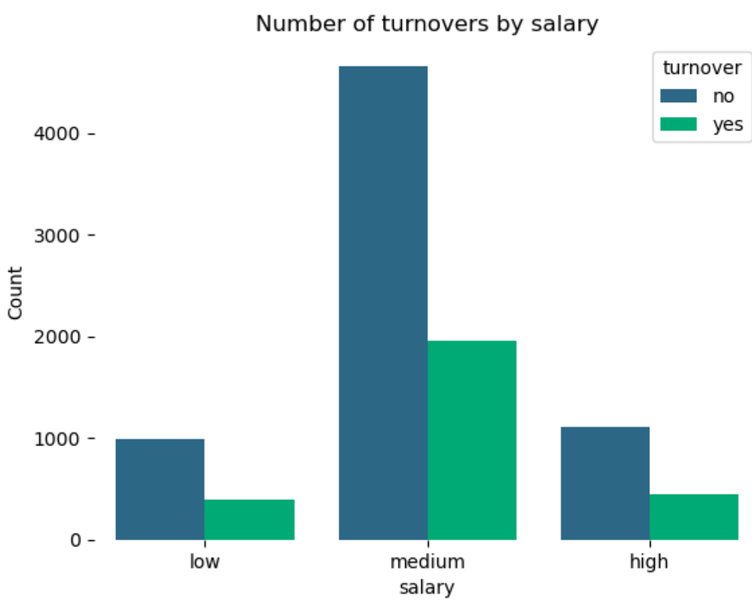
- Employee turnover dataset from Kaggle
- Anonymous US company over 2016 -2020
- 9540 employee records, 10 columns

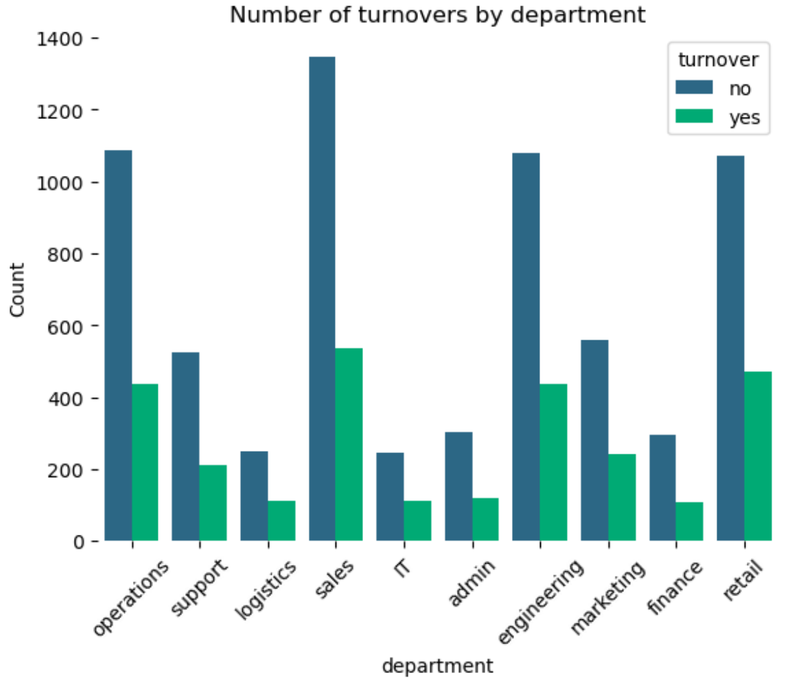
Feature Type	Feature Description
Category	**Turnover**, department, salary band
Numeric - Continuous	Average hours worked per month, performance score, employee satisfaction score
Numeric - Discrete	Bonus, Promotion, Tenure, Number of projects involved

ASSUMPTION ➡ Records are collected in regular intervals (e.g. monthly).

Exploratory Data Analysis (EDA)



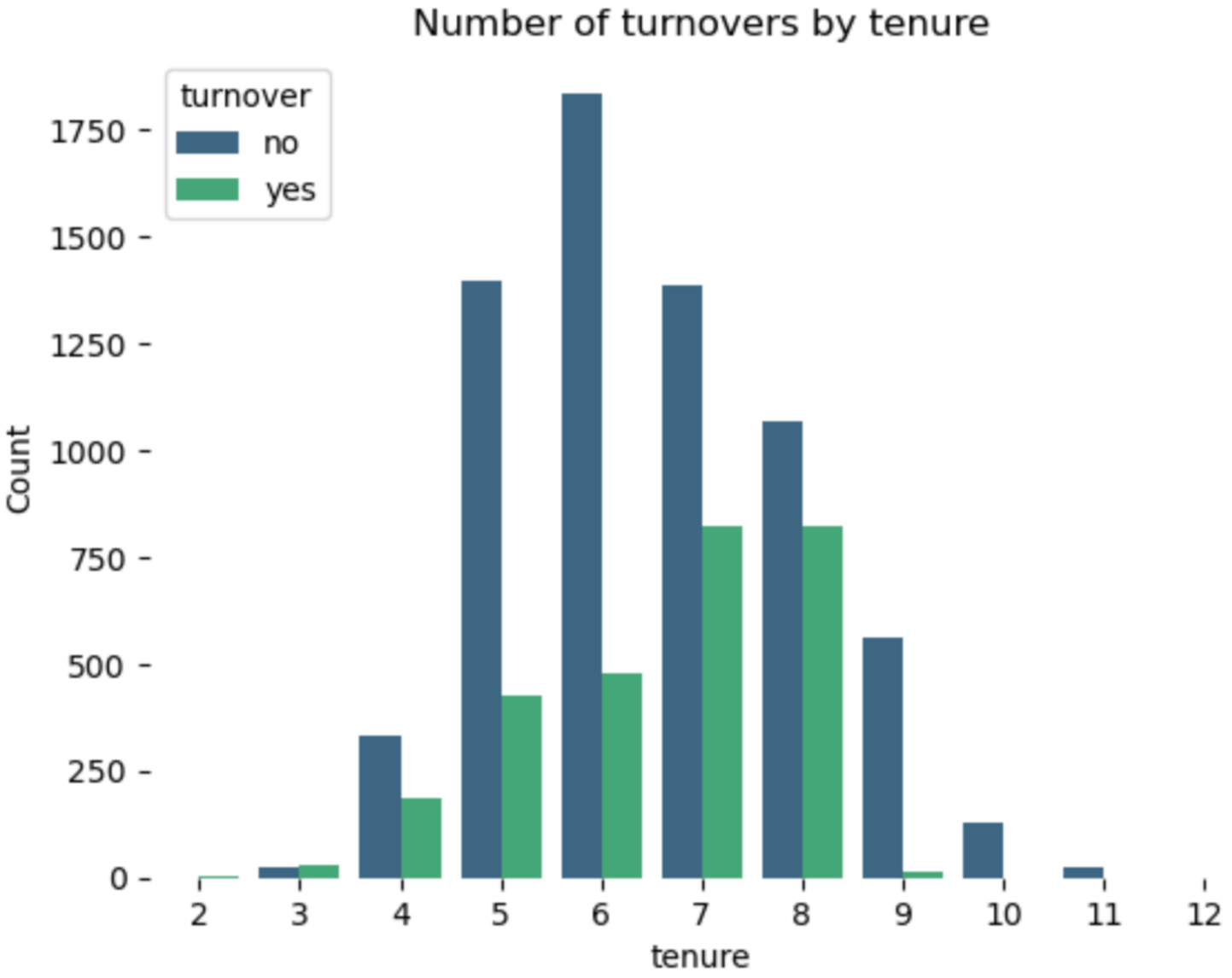




71% of employees stayed and **29%** left, over the intervals data was collected.

By salary band and department, proportions of turnover were similar across categories.

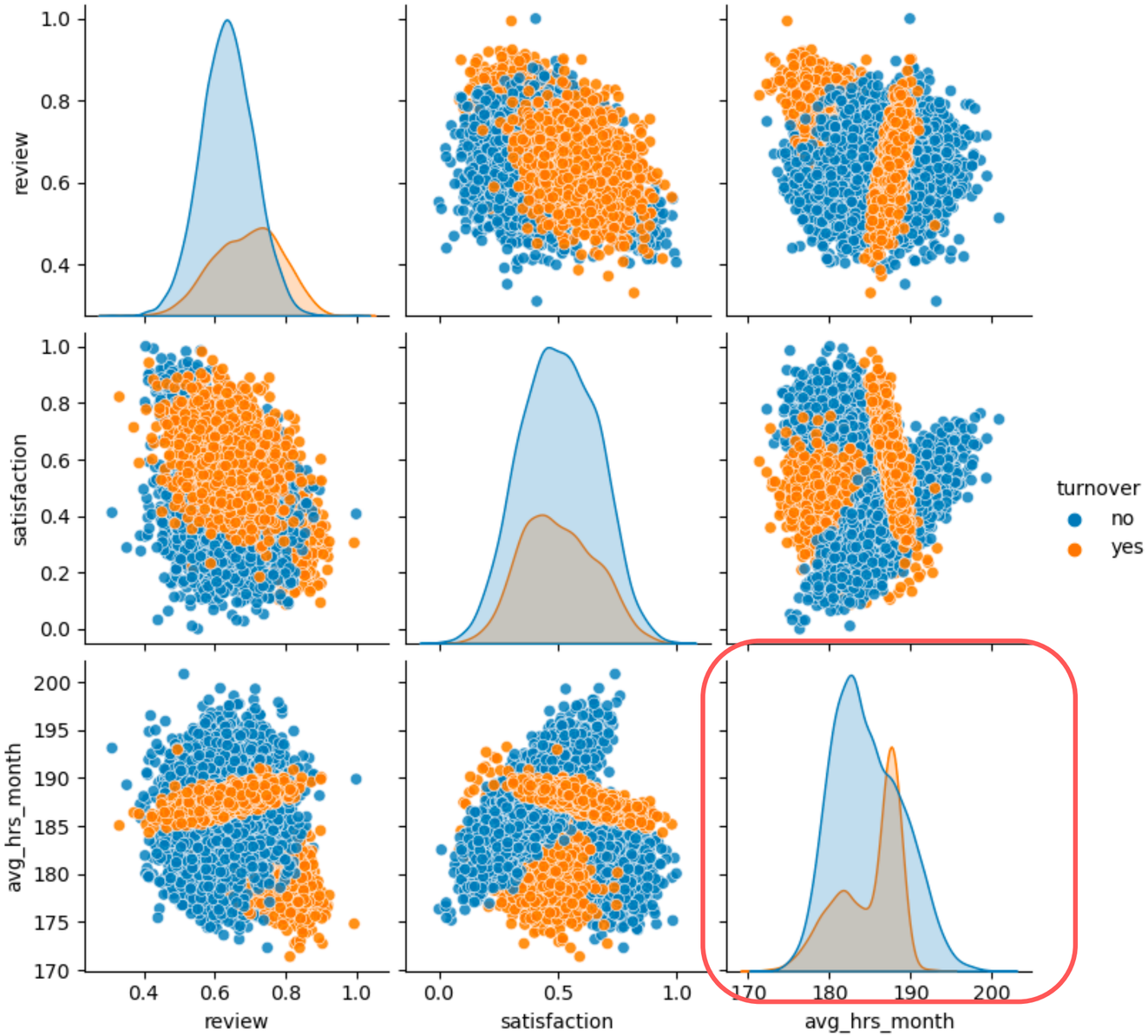
Exploratory Data Analysis (EDA)



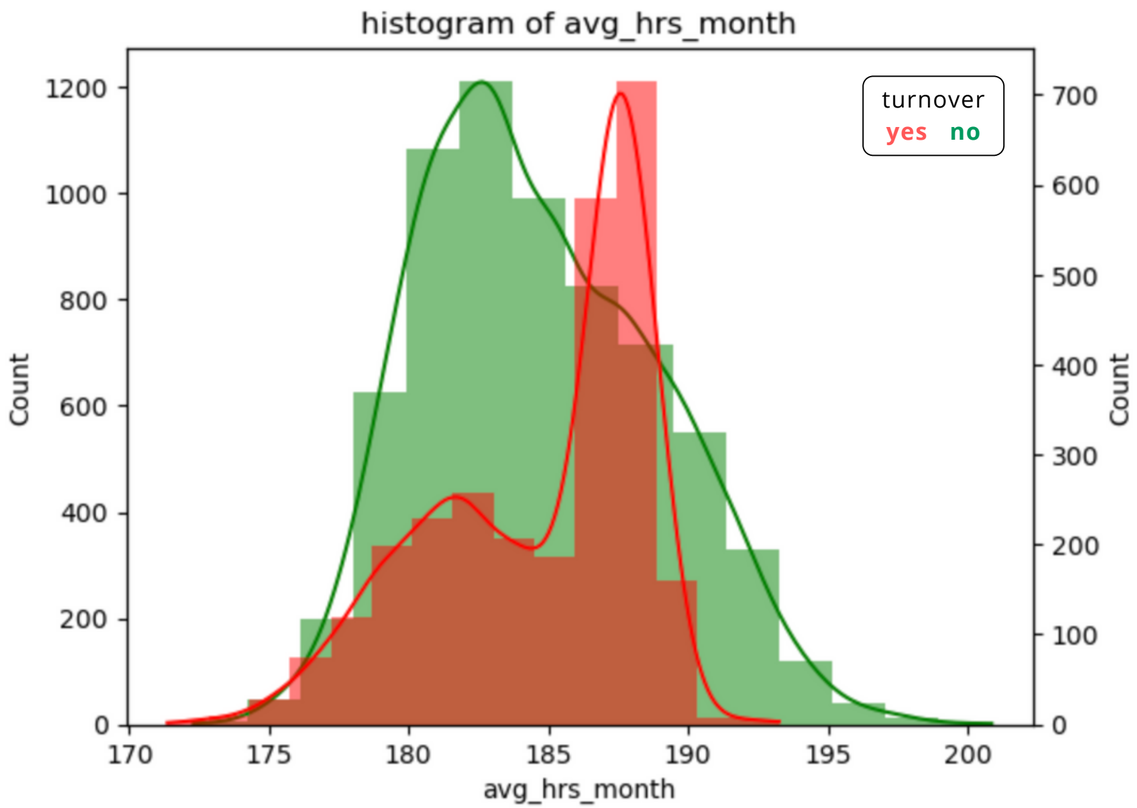
EDA - Continued

- Comparing continuous features
- Clustering present
- Possibly not linearly separable

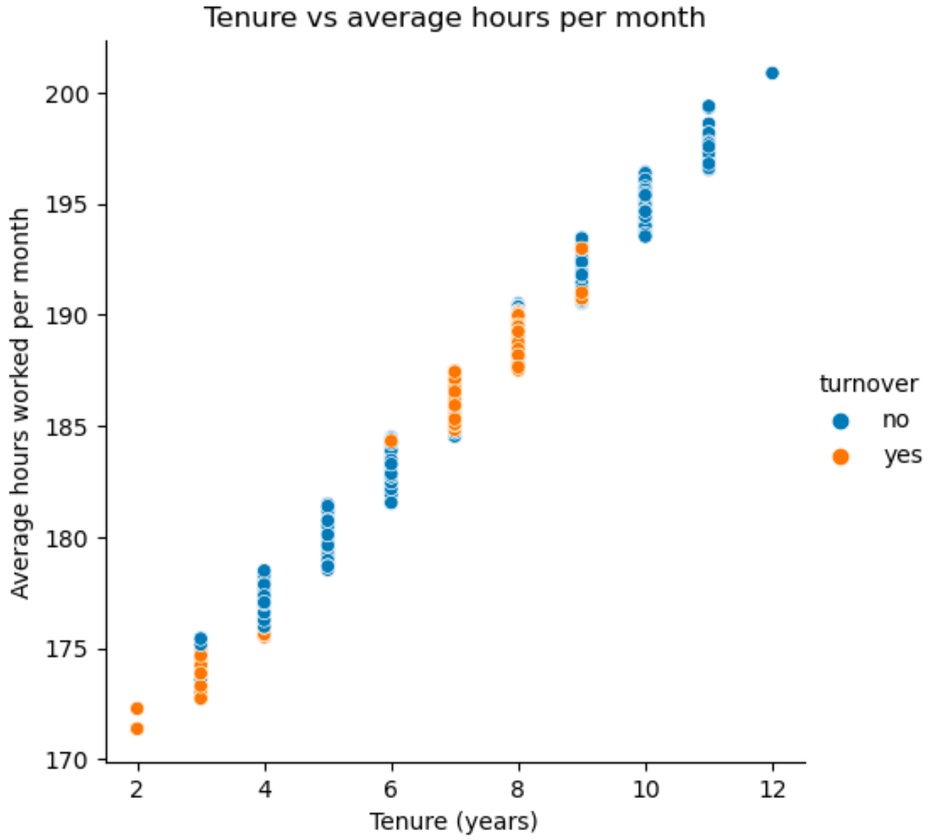
If not linearly separable, performance of some models will be affected.



EDA - Continued

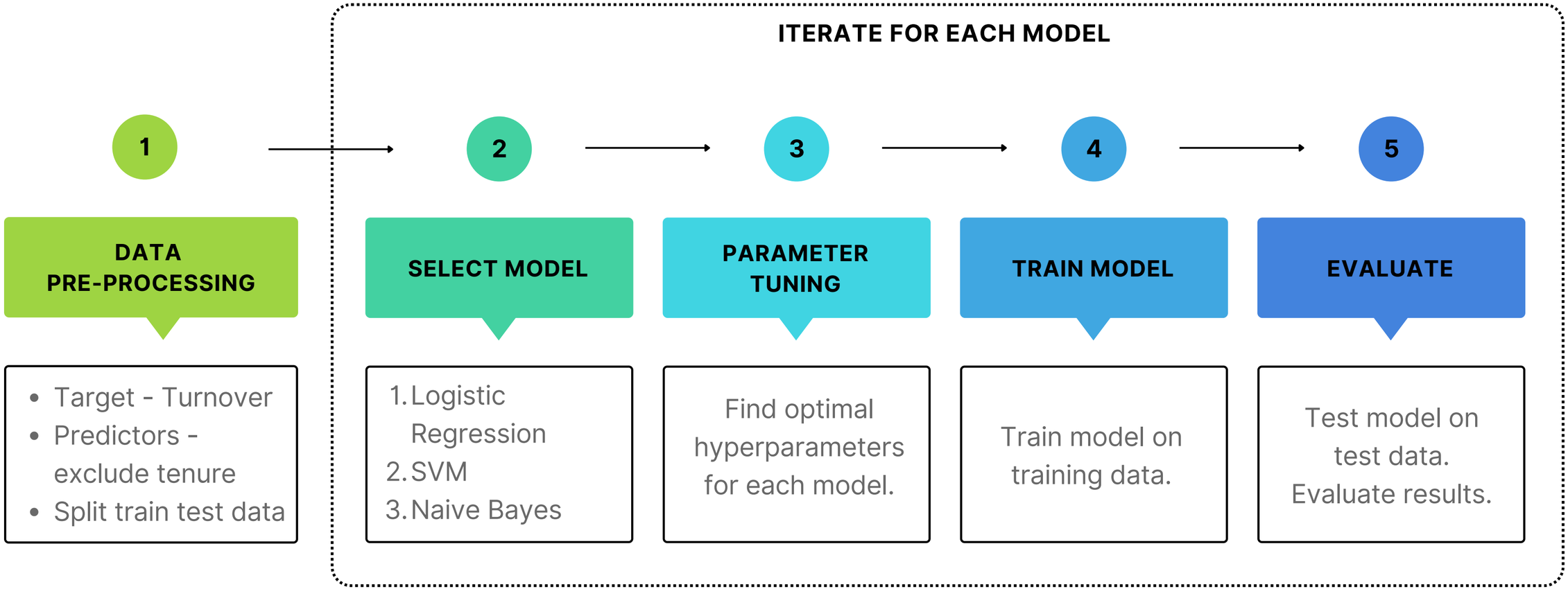


Average monthly hours worked appears higher on average for employees who left.



Very strong positive correlation between average hours per month and tenure.

Model Development



Evaluation

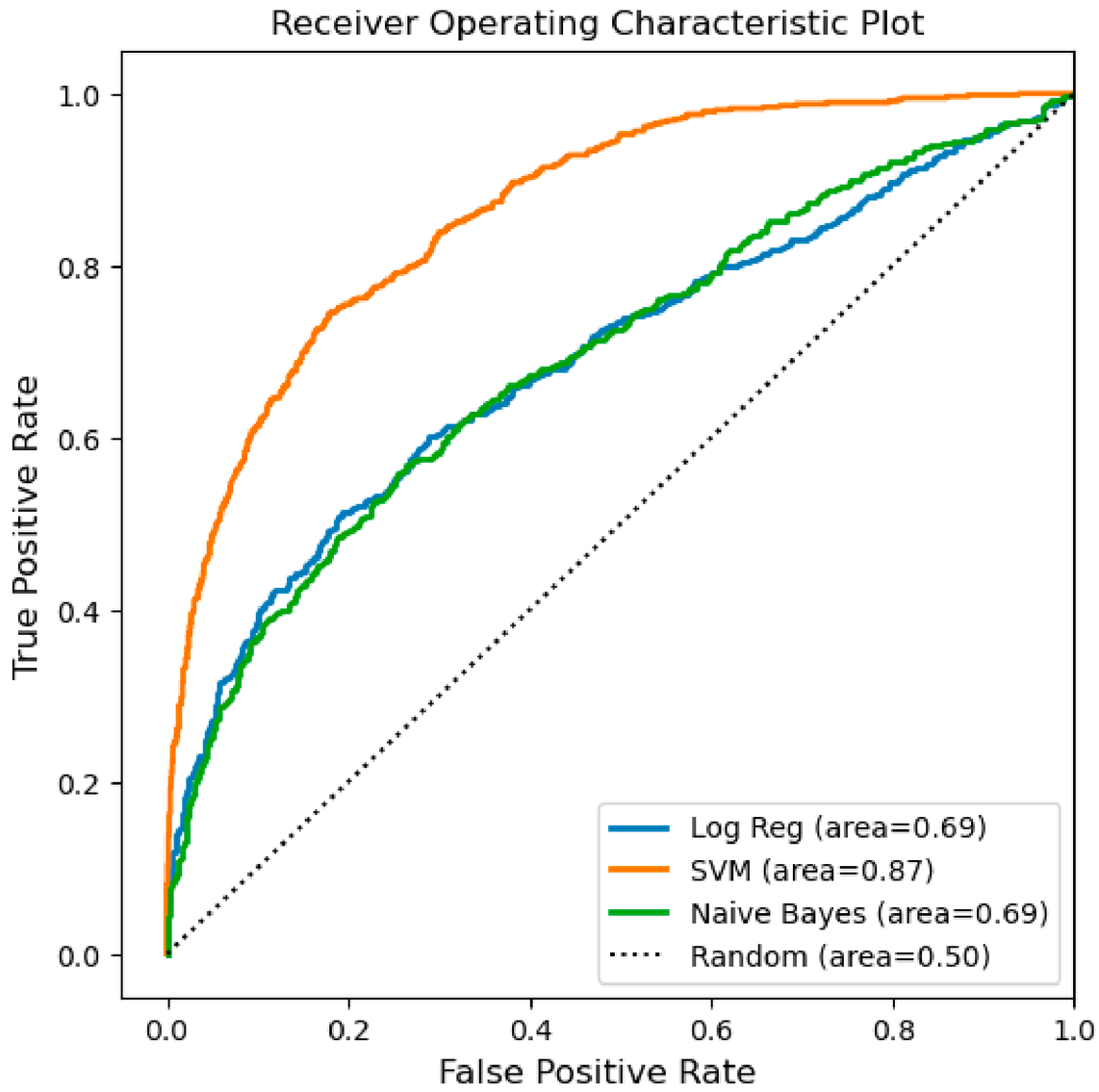
Score	Description	Logistic Regression	SVM	Naive Bayes
Accuracy	Percentage of **all** correct predictions.	74%	**81%**	74%
Precision	Percentage of **positive predictions** (predicted that employee left) that are actually correct.	**79%**	77%	64%
Recall	Percentage of **actual positives** (employee left) that are correctly predicted. *Low recall = Many employees actually left, but model predicted they didn't.*	18%	**55%**	31%

SVM model had the best scores overall, with the highest accuracy (81%) and recall (55%).

Receiver Operating Characteristic (ROC)

- Area under curve (AUC)
- Measure of classification performance
- Higher is better
  - Area=1, model correctly classify 100% of employees
  - Area=0.5, similar to random classifier

SVM model had the highest AUC, indicating best performance at classifying employees that left.



Conclusion

- **SVM** model was the most effective model for predicting employee turnover
- Achieved **81%** accuracy in classifying employee turnover in the test data
- Can be used as a preliminary tool for targeted employee retention

Next Steps

- Explore ways to gather more information about employees
- Improve recall score
- Explore feature selection techniques to identify key factors behind turnover