EDA & Predictive Analysis of Healthcare Employee Attrition

Bianca Ashar

1. Introduction
2. Dataset
3. Exploration
4. Model
5. Limitations
6. Insights

What is Healthcare Attrition?

When an employee leaves the company through any method, including voluntary resignations, layoffs, failure to return from a leave of absence, or even illness or death

1 Management Issues

2 Workplace Toxicity

3 Personal Problems

4 Workforce Demographics

5 Business Relocation

6 COVID-19 Restructuring

Why is it important?

According to the 2022 NSI National Healthcare Retention & RN Staffing Report, the average hospital turnover rate in 2021 was

25.9%

revealing a 6.4% increase over the prior year which was approx 19.5%

- Medication Errors
- Hospital Readmission
- Quality of Care

Intro to Dataset

IBM Watson Health™

# Rows

1676

# Features

35

Example Features

Background Demographics
- Marital Status
- Age
- Gender

Career Information
- Job Satisfaction
- Job Level
- Monthly Income

Educational Background
- Education Level
- Educational Field
- Department

Advantages

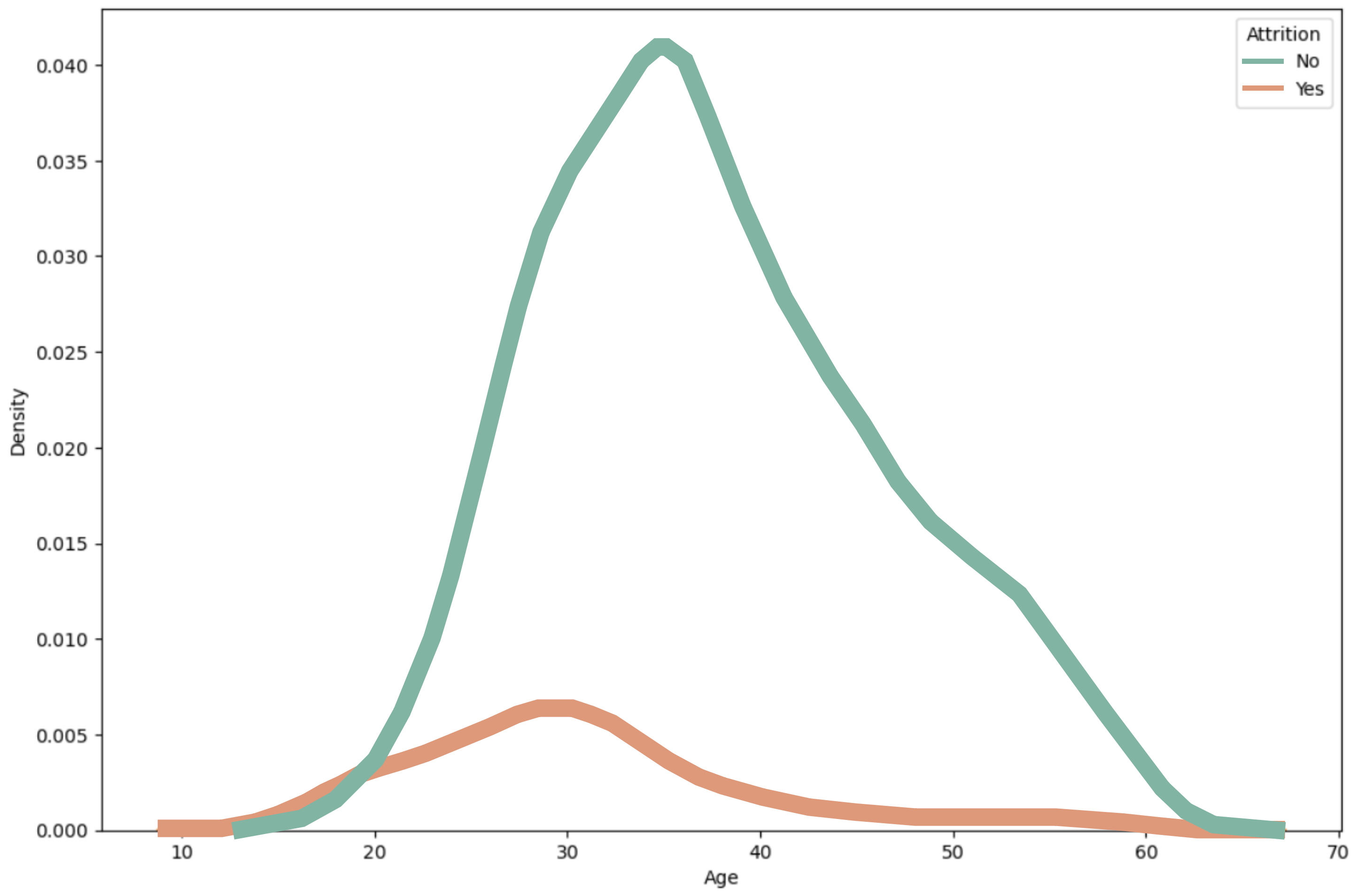
1. High Level of Detail (35 Features)
2. Includes a Variety of Demographics
3. IBM Watson: Reputable natural language processing machine

Disadvantages

1. Using Synthetic Data that may be unrepresentative
2. Lack of standardization in qualitative data
3. No Time Period Given (pre or post COVID-19)

EDA Of Age, Distance from Home, and Marital Status

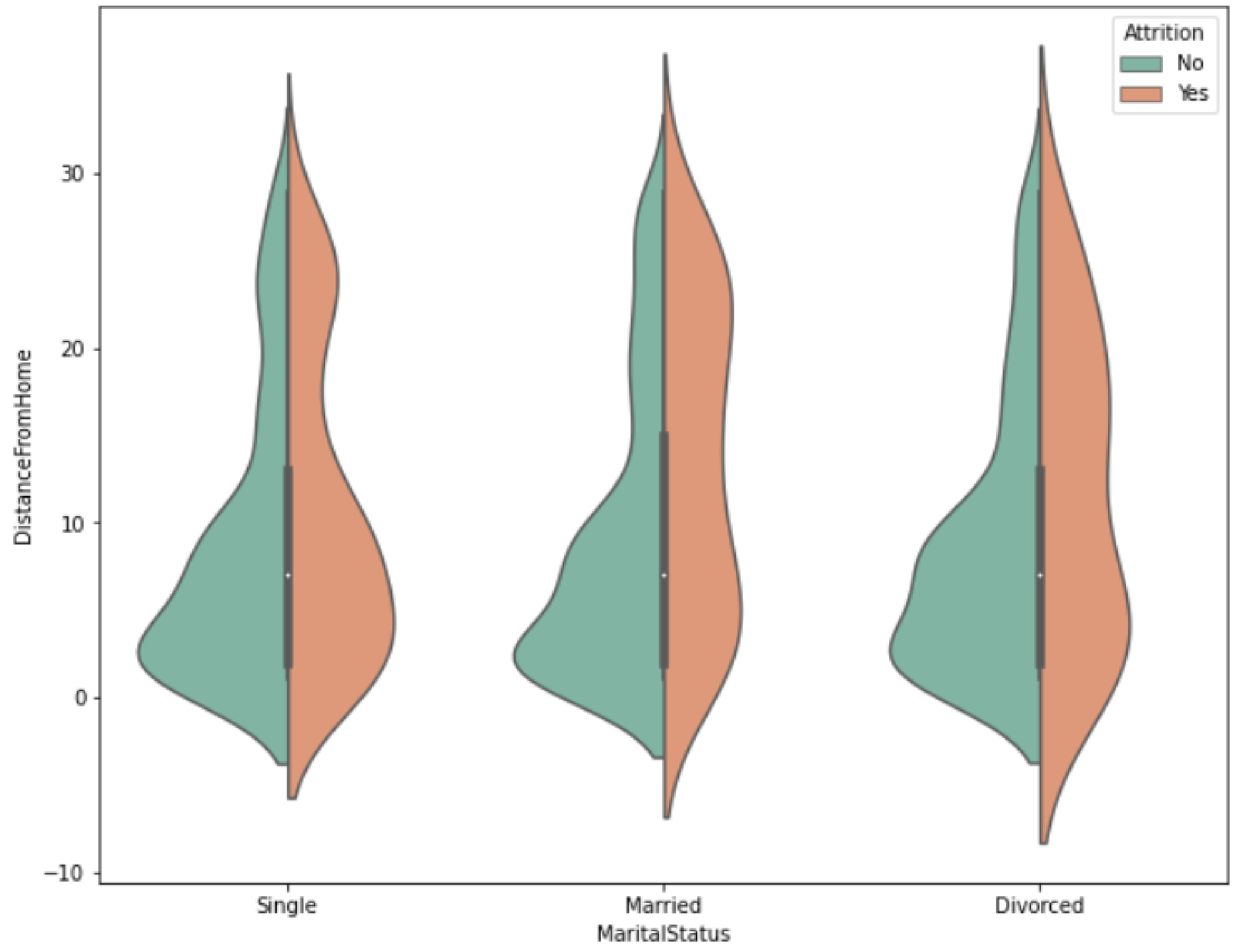
Probability Density (KDE) of Age



Ages ~ 18-35 have the highest rates of Attrition due to opportunities for pivoting

- Mean (Yes): 30.899 yrs
- Mean (No): 37.670 yrs

Distance from Home vs Marital Status (Violin Plot)

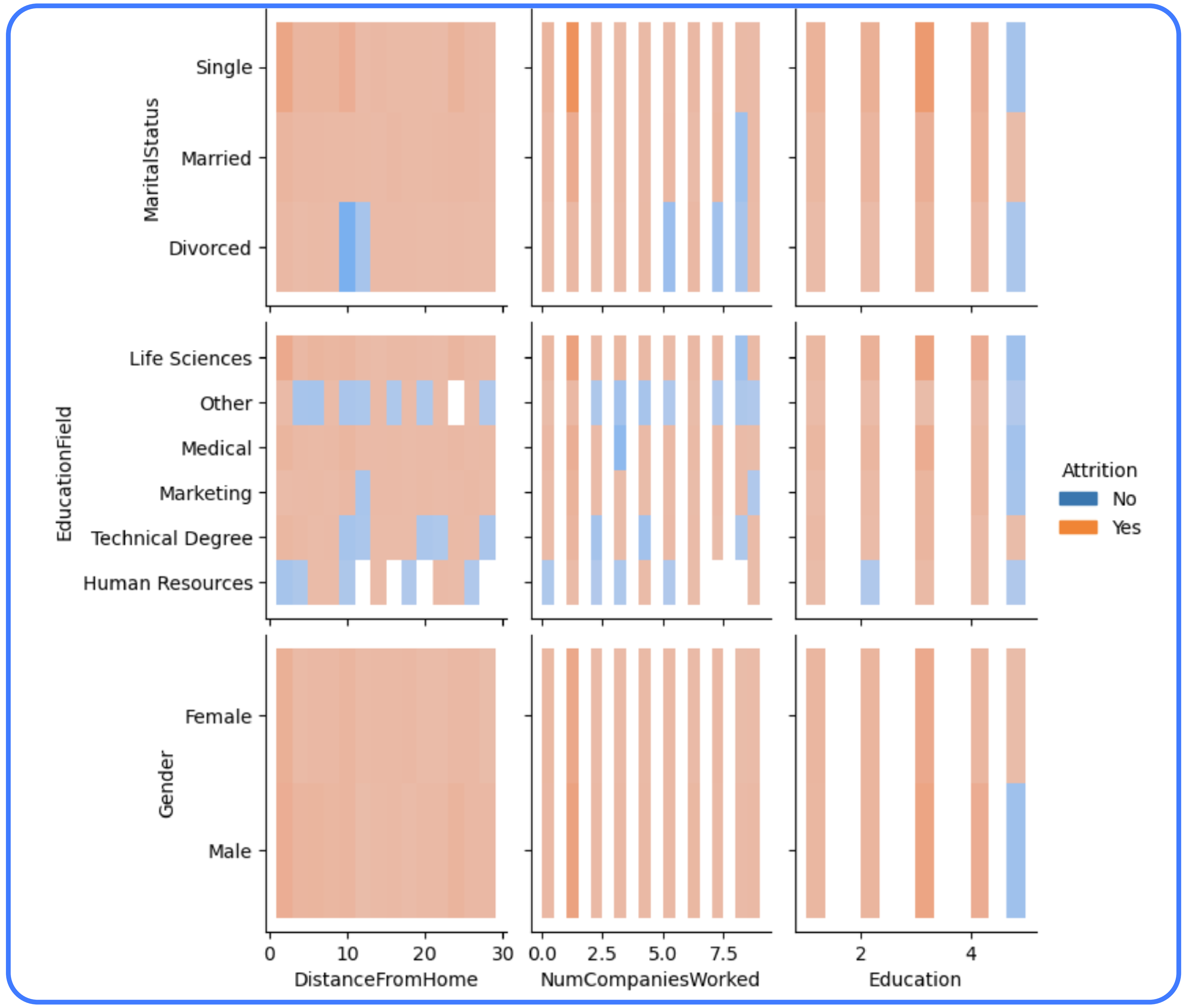


Higher distance from home results in greater attrition

- Mean (Single): 10.614
- Mean (Married): 13.361
- Mean (Divorced): 11.542

Introduction Dataset Exploration Model Limitations Insights

Background History of Employees Analysis



Observations:

1. Number of Companies Worked
   - Mean (Yes): 2.647 & Mean (No): 2.779
2. Gender
   - 60% Male & 40% Female
3. Education Level
   - Mean (Yes): 2.798 & Mean (No): 2.922
4. Educational Field
   - Roughly equal across all departments

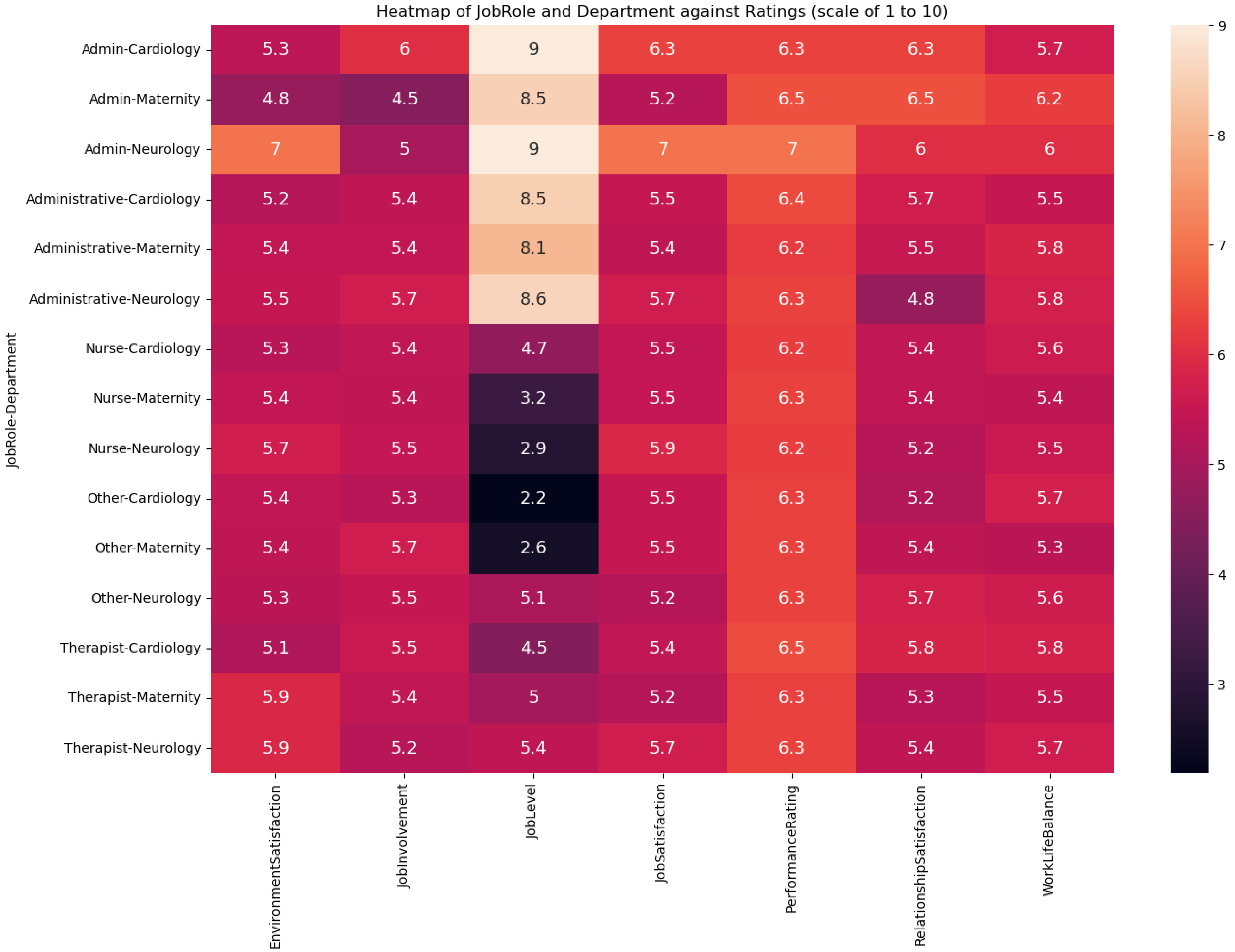
Insight: Each of these features show minimal effect on overall employee attrition based on the dataset tested as the average values are relatively similar.

Analysis of Work Engagement based on Job Role

Features: Job Involvement, Job Satisfaction, Environment Satisfaction, Relationship Satisfaction, Work Life Balance, Performance Rating, Job Level

Observations:

1. Environment, Relationship, and Job Satisfaction have minimal difference in means
2. Job Roles of Nurses & Others associated with low Job Level
3. Association between Lower Means (below 5) & Low Attrition



Heatmap of JobRole and Department against Ratings (scale of 1 to 10)

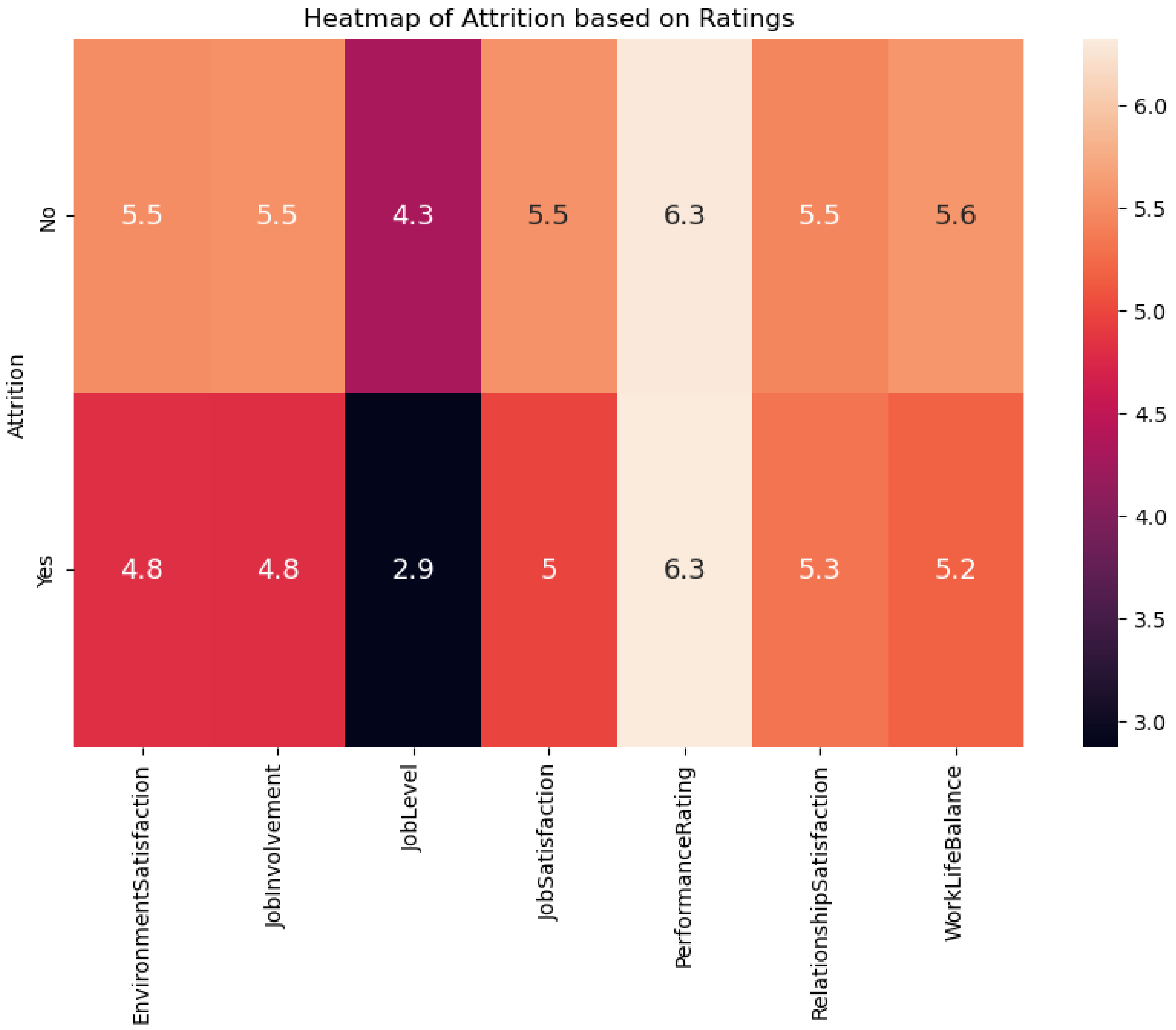
	EnvironmentSatisfaction	JobInvolvement	JobLevel	JobSatisfaction	PerformanceRating	RelationshipSatisfaction	WorkLifeBalance
Admin-Cardiology	5.3	6	9	6.3	6.3	6.3	5.7
Admin-Maternity	4.8	4.5	8.5	5.2	6.5	6.5	6.2
Admin-Neurology	7	5	9	7	7	6	6
Administrative-Cardiology	5.2	5.4	8.5	5.5	6.4	5.7	5.5
Administrative-Maternity	5.4	5.4	8.1	5.4	6.2	5.5	5.8
Administrative-Neurology	5.5	5.7	8.6	5.7	6.3	4.8	5.8
Nurse-Cardiology	5.3	5.4	4.7	5.5	6.2	5.4	5.6
Nurse-Maternity	5.4	5.4	3.2	5.5	6.3	5.4	5.4
Nurse-Neurology	5.7	5.5	2.9	5.9	6.2	5.2	5.5
Other-Cardiology	5.4	5.3	2.2	5.5	6.3	5.2	5.7
Other-Maternity	5.4	5.7	2.6	5.5	6.3	5.4	5.3
Other-Neurology	5.3	5.5	5.1	5.2	6.3	5.7	5.6
Therapist-Cardiology	5.1	5.5	4.5	5.4	6.5	5.8	5.8
Therapist-Maternity	5.9	5.4	5	5.2	6.3	5.3	5.5
Therapist-Neurology	5.9	5.2	5.4	5.7	6.3	5.4	5.7

Analysis of Work Engagement based on Job Role

Features: Job Involvement, Job Satisfaction, Environment Satisfaction, Relationship Satisfaction, Work Life Balance, Performance Rating, Job Level

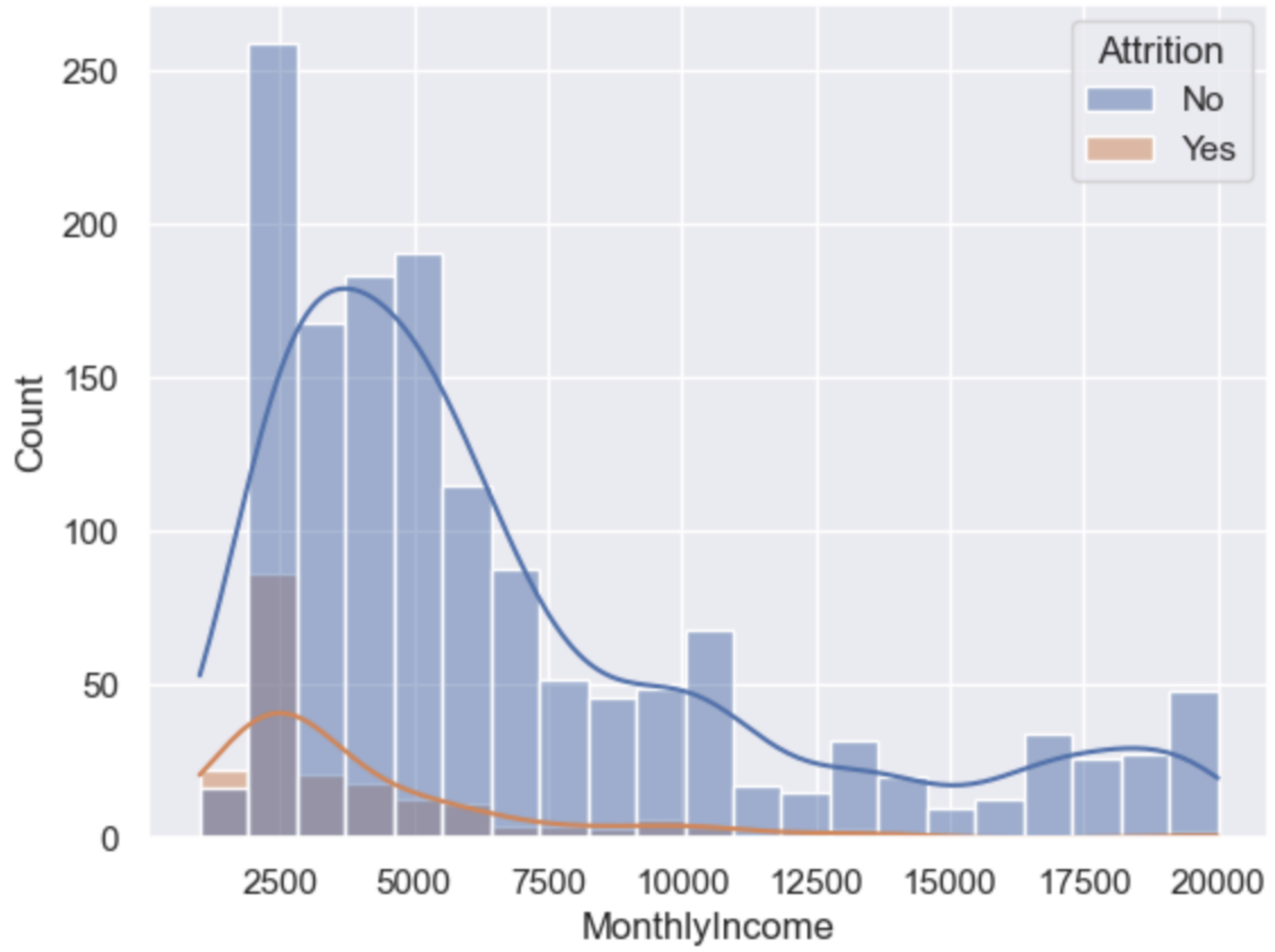
Observations:

1. Largest difference in means of Job Level (1.4) between Attrition categories
2. No difference in means within Performance Ratings
3. Mild difference (~0.5-0.7) seen in Environment Satisfaction, Job Involvement, and Job Satisfaction



Analysis of Work Compensation on Attrition

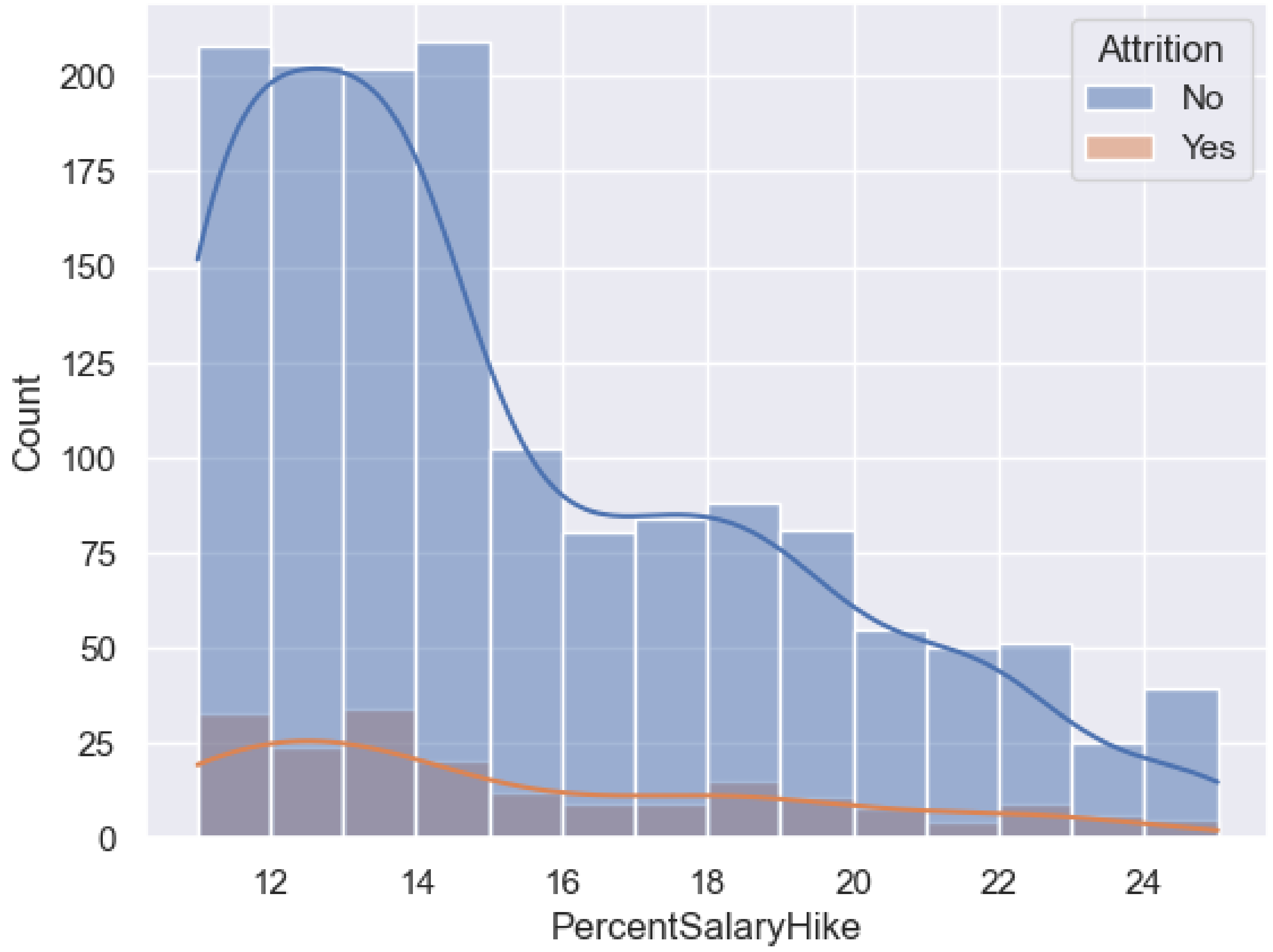
Distribution of Monthly Income ($)



Lower Monthly Income directly correlates to higher chance of Attrition

- Mean (Yes): $4,024.246
- Mean (No): $6,852.302

Distribution of % Salary Hike



Lower Percent Salary Hike doesn’t signal higher chance of Attrition

- Mean (Yes): 15.226%
- Mean (No): 15.193%

Prediction Model Overview

1 Column Selection & Splitting the Data

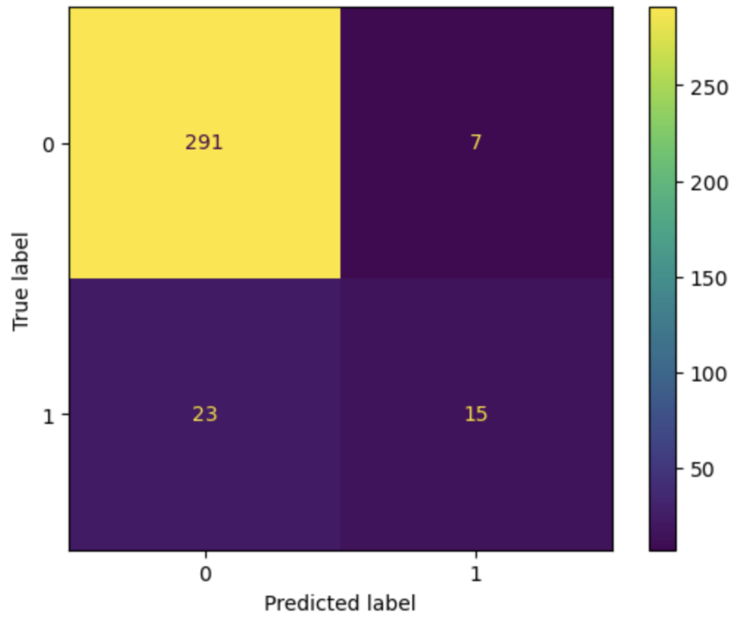
- Over Time
- Age
- Distance From Home
- Marital Status
- Monthly Income
- Job Involvement
- Environment Satisfaction
- Job Satisfaction

2 Training/Testing Data to Optimize Model

Split data into Training & Test Sets & Tested different models

- Model Accuracy = 0.9107
- Precision Score = 0.6818

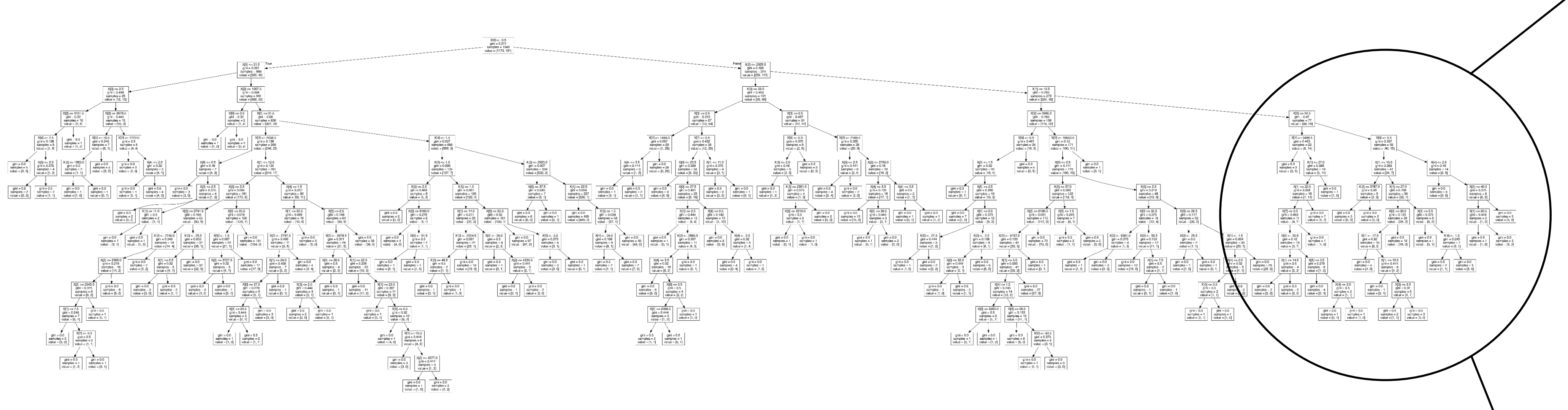
Confusion Matrix

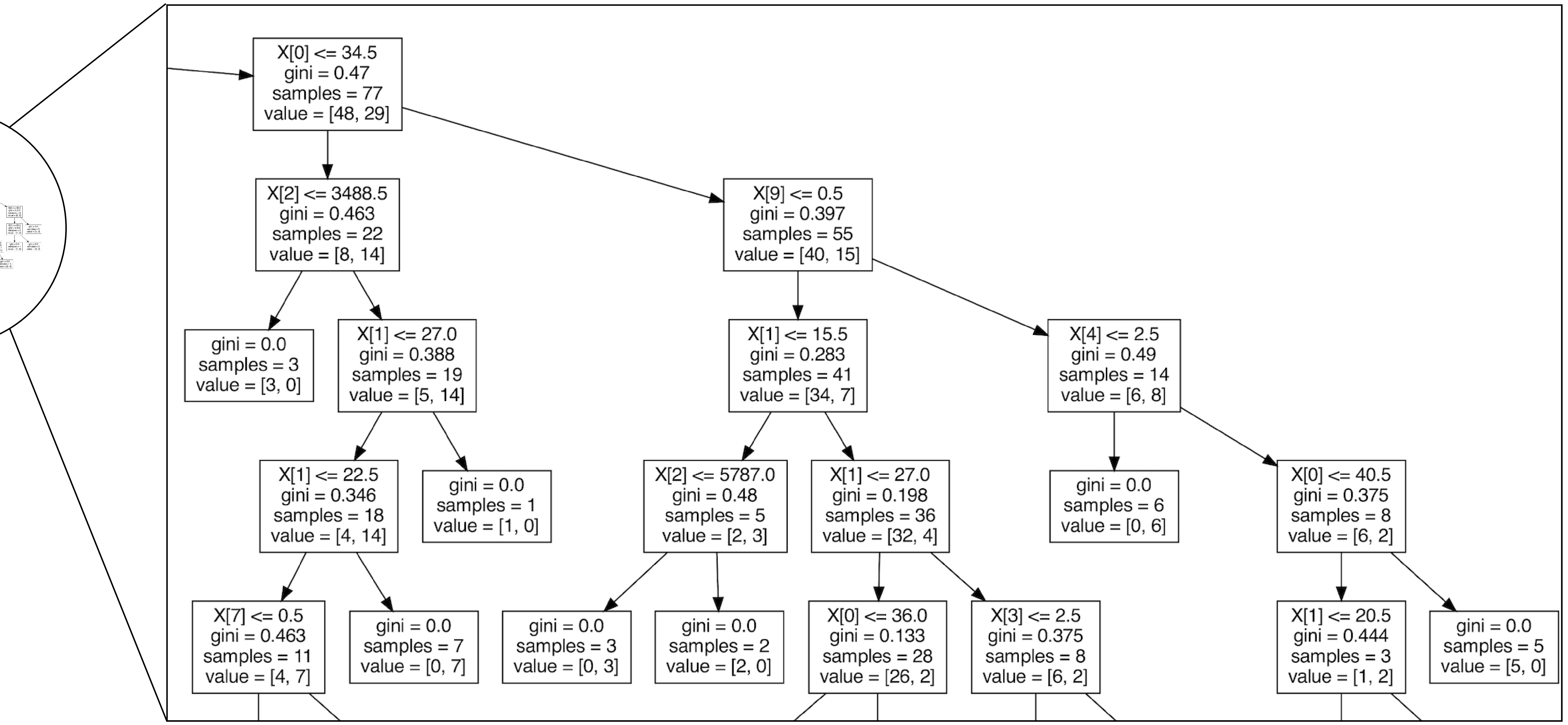


3 Model Finalization & Pruning

- Decreased number of features that were looked at (to prevent overfitting)
- Optimized tree depth = 3
- Limitations
  - Model Accuracy
  - Underfitting

Unpruned Decision Tree





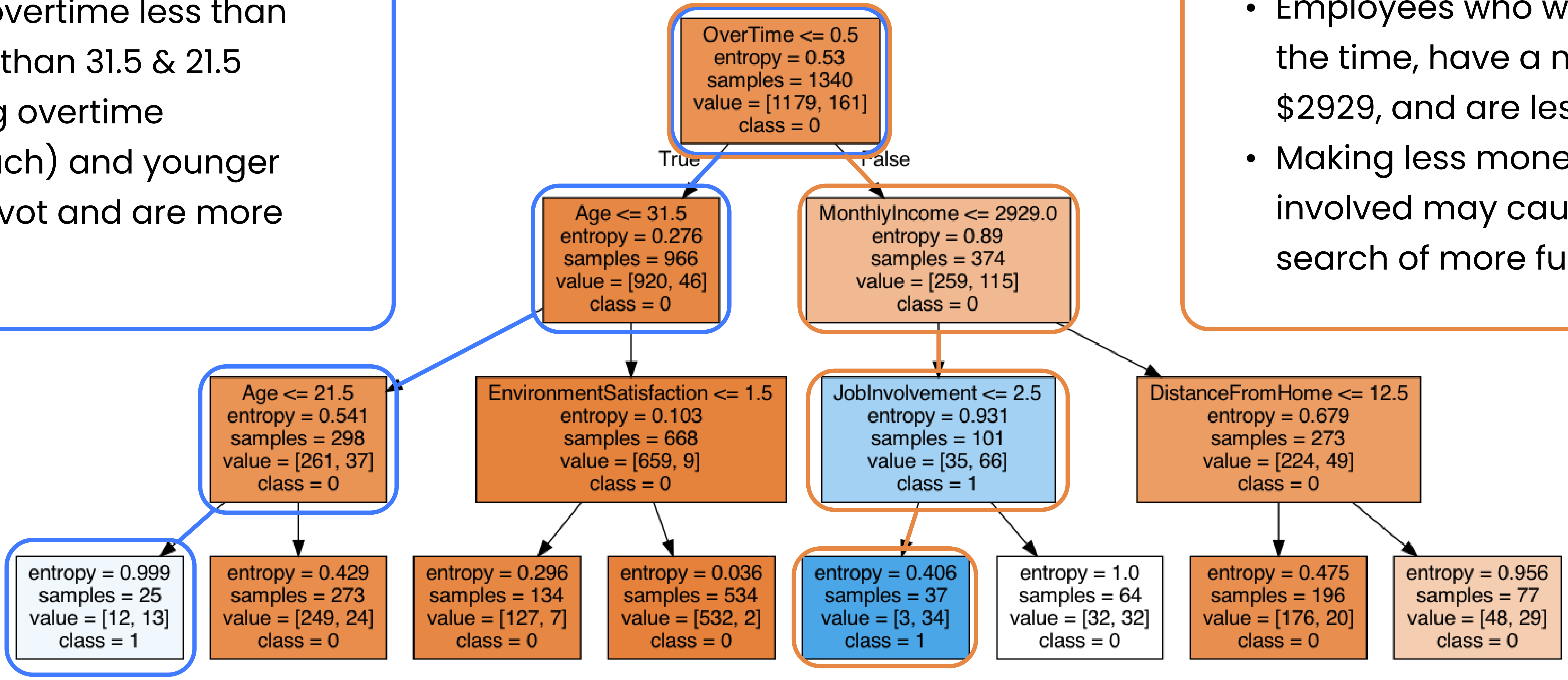
ex) X[0] = Age, Gini = 0.47, Samples = 77, Value = [48, 29]

- Lower Gini score: Lower chance of misclassification
- Samples: # of employees in that category
- Value: Tells how many values fall into each category [No Attrition (0), Attrition (1)]

Final Pruned Decision Tree Model

Cluster 1:

- Employees who work overtime less than half the time, younger than 31.5 & 21.5
- Those who are working overtime (regardless of how much) and younger may find it easier to pivot and are more likely to quit



Cluster 2:

- Employees who work overtime less than half the time, have a monthly income of less than $2929, and are less involved in their job
- Making less money and not being actively involved may cause employees to quit in search of more fulfilling, higher paying roles

Limitations

Unstable Nature of Decision Tree

Slight changes to data can completely change the tree construction

- Unbalanced dataset

Lack of Various Datasets

Same dataset was split into both training and testing datasets which could potentially skew results

Loss of Prediction Model Accuracy

Pruning process could result in underfitting of data

- Removed Marital Status feature: prior EDA Analysis showed its importance

Key Insights

1 People of younger ages are more likely to leave the workplace, especially those with less years working in the hospital

2 Working overtime is a common factor in almost all attrition clusters as it reduces work life balance and overall satisfaction

3 Given increasing inflation and cost of living, a lower monthly income has a high correlation with rising levels of attrition

4 Educational background doesn’t have any noticeable effect or correlation with attrition & there are equal amounts of attrition across all education levels

Business Recommendations

Improve Recruiting & Onboarding: Introducing sign on bonuses, tangible benefits, wellness perks, and well-organized onboarding and training

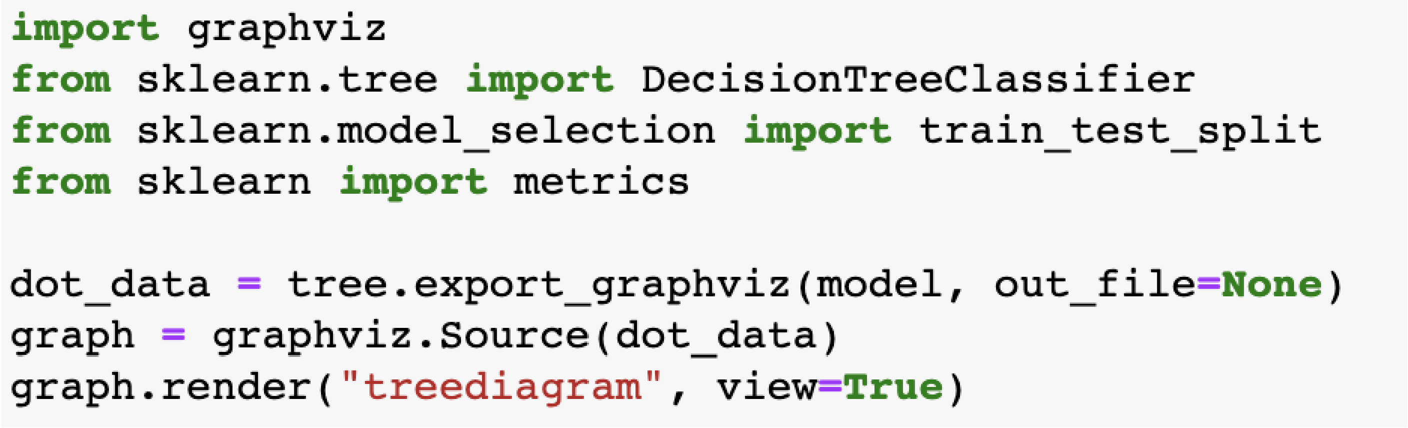
Build Community Engagement: Establish positive hospital culture, promote work life balance, and encourage open communication between doctors & nurses

Invest in Employee Engagement: Organizing mentoring programs and require Continuing Medical Education (CME) & Professional Dev (CPD)

Questions?

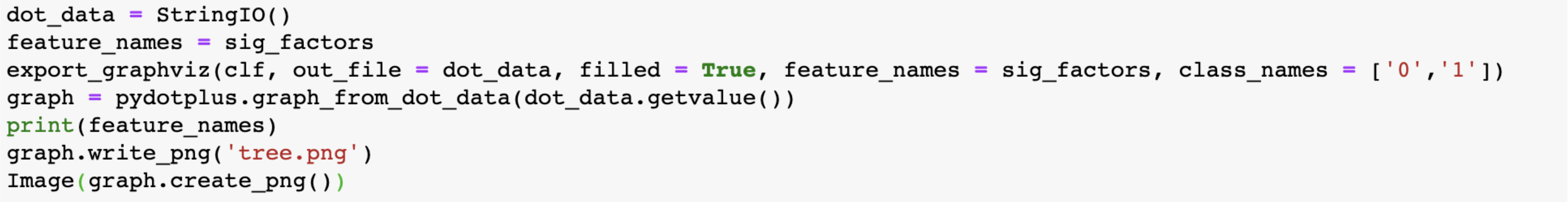
Appendix

Decision Tree Classifier



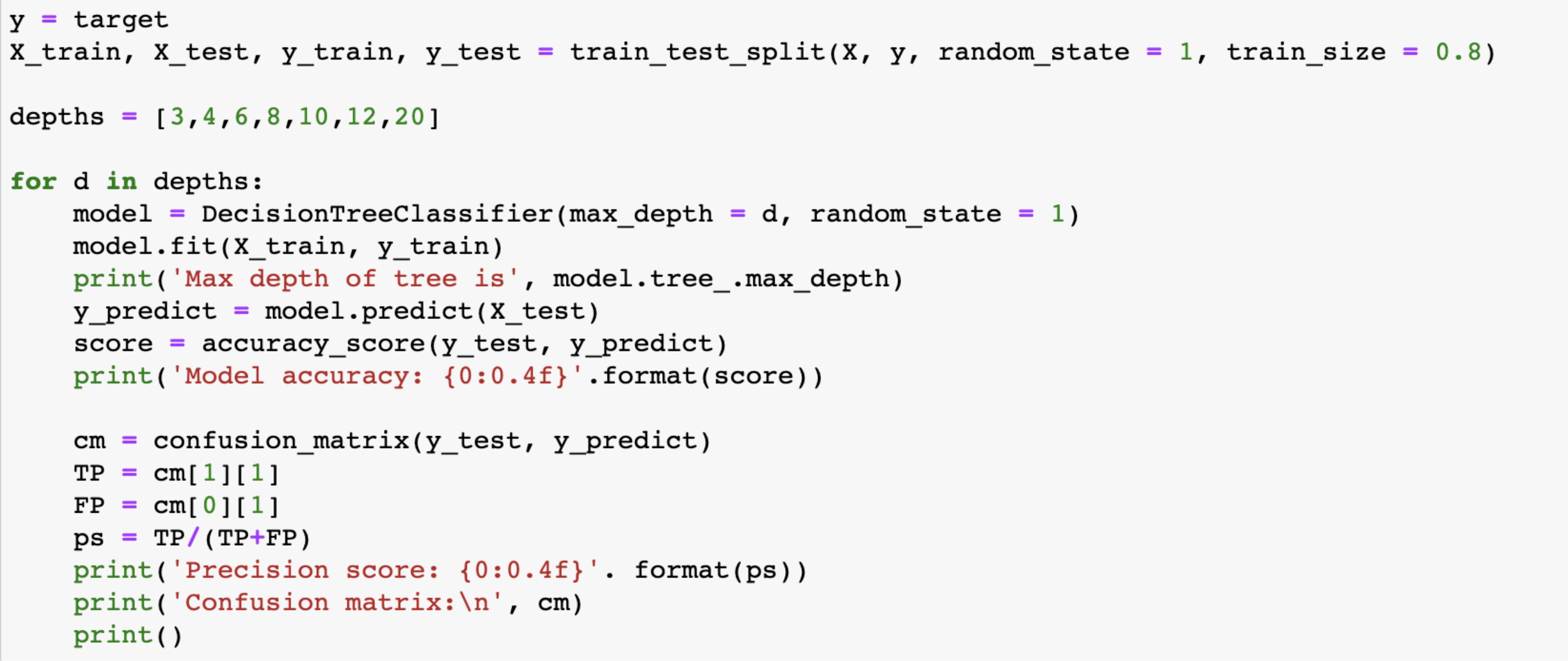
```
import graphviz
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn import metrics

dot_data = tree.export_graphviz(model, out_file=None)
graph = graphviz.Source(dot_data)
graph.render("treediagram", view=True)
```

Post-pruning



```
dot_data = StringIO()
feature_names = sig_factors
export_graphviz(clf, out_file = dot_data, filled = True, feature_names = sig_factors, class_names = ['0','1'])
graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
print(feature_names)
graph.write_png('tree.png')
Image(graph.create_png())
```

Splitting Data & Testing Models



```
y = target
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 1, train_size = 0.8)

depths = [3,4,6,8,10,12,20]

for d in depths:
    model = DecisionTreeClassifier(max_depth = d, random_state = 1)
    model.fit(X_train, y_train)
    print('Max depth of tree is', model.tree_.max_depth)
    y_predict = model.predict(X_test)
    score = accuracy_score(y_test, y_predict)
    print('Model accuracy: {0:0.4f}'.format(score))

    cm = confusion_matrix(y_test, y_predict)
    TP = cm[1][1]
    FP = cm[0][1]
    ps = TP/(TP+FP)
    print('Precision score: {0:0.4f}'. format(ps))
    print('Confusion matrix:\n', cm)
    print()
```