Machine Learning for Predictive Analytics Mini Sprint

Predicting Attrition rate for Bain and Company

from the HR Dataset

Saima Ahmed
(Individual Contribution)

My objective is to develop a machine learning model to predict employee attrition rates. This model will minimize overfitting and achieve good performance metrics. By providing HR with accurate insights into employee departure risks, the model will empower them to implement effective retention strategies and improve employee satisfaction.

Data Analysis Summary

Employees statistics who leave the company

1. Age - 25-35 years
   a. 63.3% of men left the job as compared to 36.7% women.
2. Most people who leave a job have a low salary.
   a. Monthly income bracket (20- 70K) - As income increases, attrition decreases.
3. Percentage Salary Hike - less than or equal to 14%
4. Stock Option Level- 0 and 1 level
5. Job Level - Junior to mid-Junior
6. Education - College to Bachelor
7. Number of years worked - one year
   a. Most people who work in the company had less than 1 year experience with the current manager.
8. Total Working Years- 1, 5 and 10 years
9. Training Times Last Year - 2 and 3 times
10. Years since Last Promotion-0-2 years
11. Job Involvement- 2-3 levels
12. Performance Rating - 2-3 rating
13. Work Life Balance - 3 rating
14. Most people leave the company - who travel rarely.
15. Most people who leave the job come from - research and development department, 63.7%.
16. Most people who leave their job were working - as a sales executive , research scientist and laboratory technician.
17. Most people who leave the company studied life sciences
18. Most people who leave the job aren't satisfied with the work environment.

Feature : Age - 25-35 years.

Feature: Gender - 63.3% of men left the job as compared to 36.7% women.

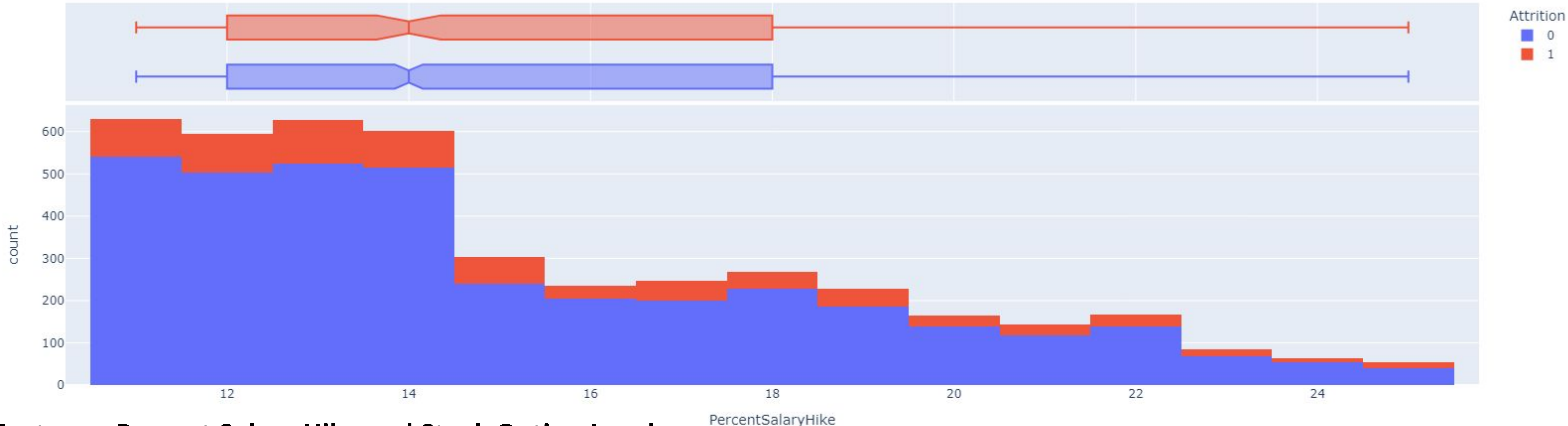
Attrition ->1

No Attrition -> 0

Feature: Monthly Income

Attrition ->1

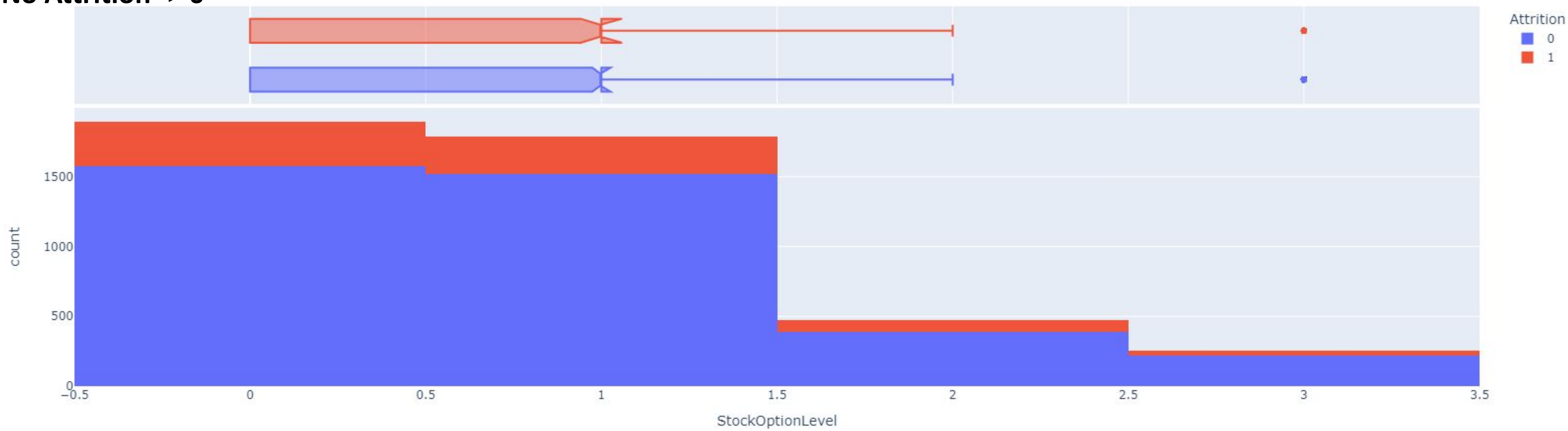
No Attrition -> 0



Features: Percent Salary Hike and Stock Option Level

Attrition ->1

No Attrition -> 0



Takeaways

Data Insights

- Most people who leave a job have a low salary.
    - Monthly income bracket (20- 70K) - As income increases, attrition decreases.
- Percentage Salary Hike - less than or equal to 14%.
- Stock Option Level- 0 and 1 level.

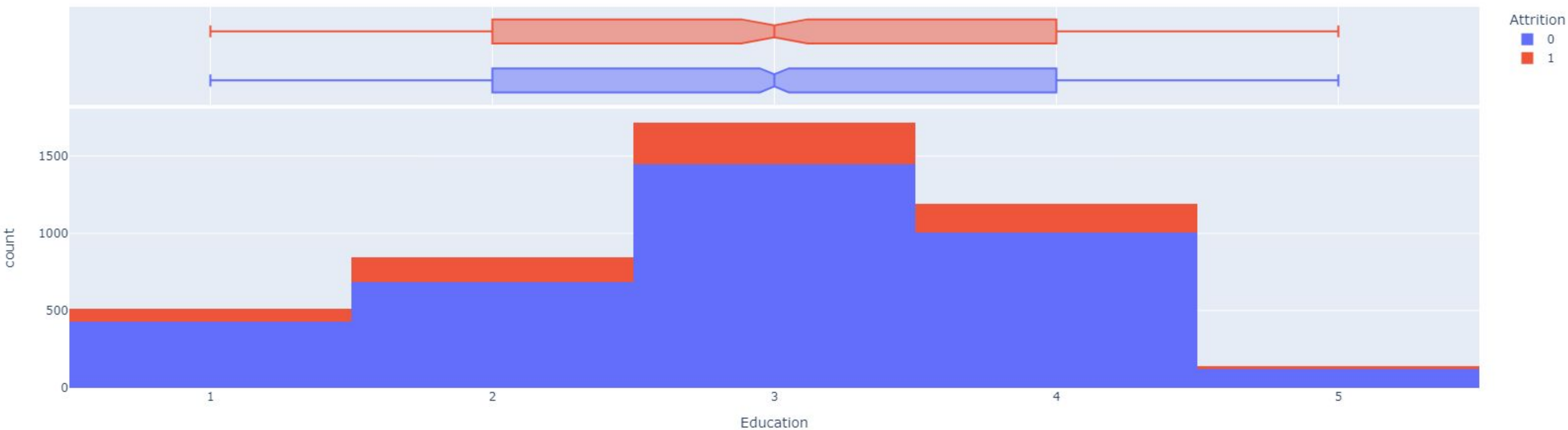
Suggestions

- **Conduct regular salary reviews**
    - **Compensated competitively**.
- **Evaluate benefits package**
    - **Flexible work arrangements**, **wellness programs**, or **educational reimbursements**.
- **Performance based Rewards**
    - **Performance-based bonuses** incentives to motivate employees.

Feature: Job level and Education

Attrition ->1

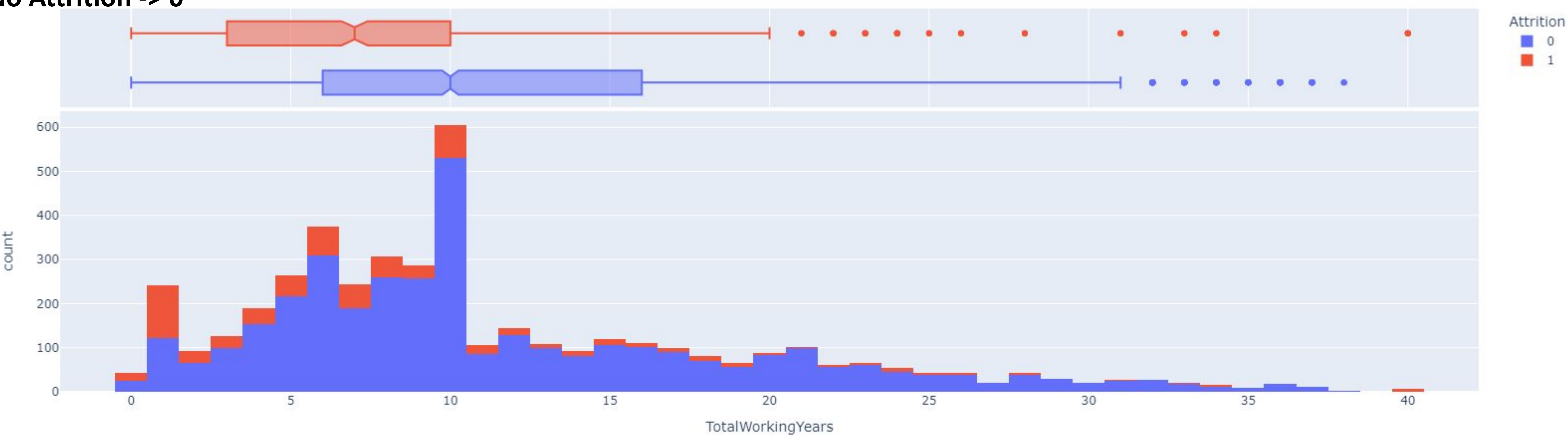
No Attrition -> 0



eature: Num Companies Worked and Total Working Years

ttrition ->1

lo Attrition -> 0



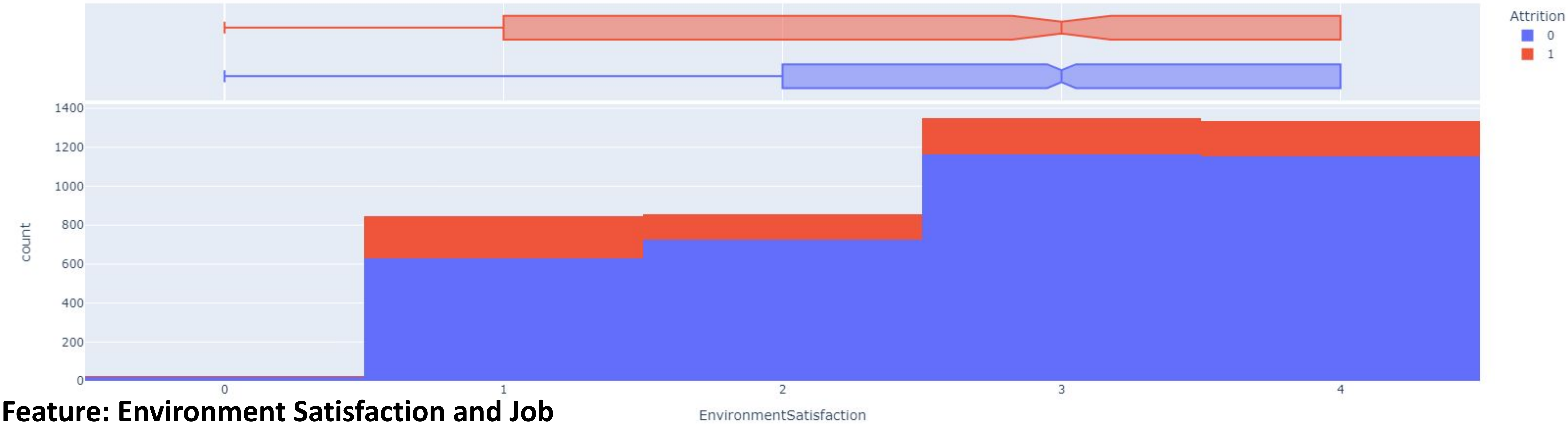
Takeaways

Data Insights

- Most people who leave a job has job level- - Junior to mid-Junior- 1 and 2 level.
- Education - College and Bachelor.
- Total Working Years- 1, 5 and 10 years.
- Number of years worked - one year.
  - Most people who work in the company had less than 1 year experience with the current manager.

Suggestions

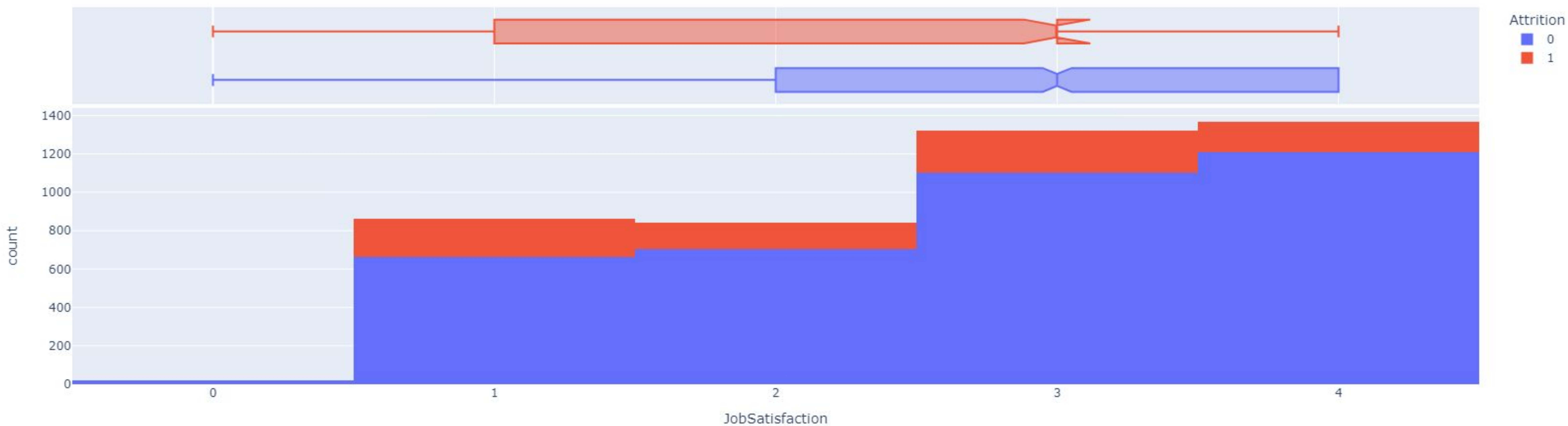
- **Career Path Visibility**
  - Clearly outline **potential career paths** within the company. Show junior staff how their current role can lead to **future opportunities for advancement**.
- **Mentorship Programs**
  - Establish **mentorship programs that pair junior staff with experienced colleagues** who can provide guidance and support.
- **Internal Job Postings**
  - Give **junior staff priority access to internal job postings** before opening positions to external candidates. This demonstrates your commitment to their growth within the company.
- **Tuition Reimbursement**
  - **Offer tuition reimbursement programs** for relevant academic or professional development courses.

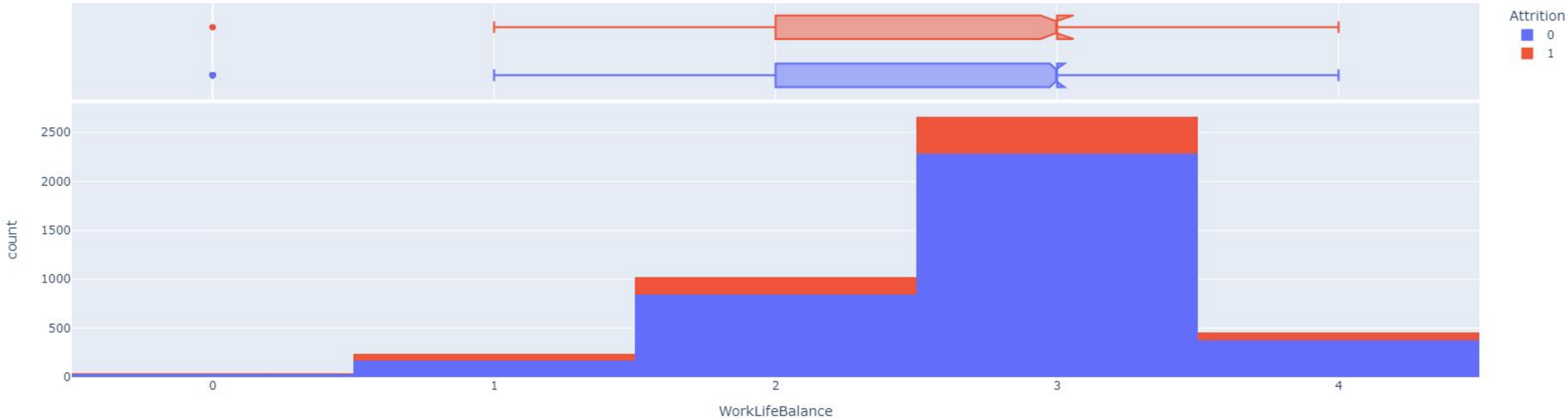


Feature: Environment Satisfaction and Job Satisfaction

Attrition ->1

No Attrition -> 0





Feature: Work Life Balance and Job Involvement

Attrition ->1

No Attrition -> 0

Takeaways

Data Insights

- Job Involvement- 3 level.
- Work Life Balance - 3 level.
- Job Satisfaction -1, 3 and 4 levels.
- Environment Satisfaction-1,2,3 4 levels.
    - Most people who leave the job aren't satisfied with the work environment.

Suggestions

- **Regular Feedback and Recognition**
    - Provide **regular feedback and recognition** to staff to keep them motivated and engaged.
- **Flexible Work Arrangements**
    - Offer **flexible work schedules, remote work options**, or compressed workweeks.
- **Discourage Overtime**
    - Set **clear expectations about working hours** and discourage excessive overtime.
- **Open Communication**
    - **Foster a culture of open communication** where employees feel comfortable voicing their concerns and suggestions.
- **Positive Workplace Culture**
    - Promote a positive and respectful work environment. **Address conflicts promptly**, encourage collaboration, and celebrate team achievements.

Takeaways

Suggestions

- **Employee Well-being**
  - Invest in employee **well-being initiatives** such as on-site fitness programs, wellness workshops, or Employee Assistance Programs (EAPs).
- **Meaningful Work**
  - Employees crave a sense of purpose. Ensure employees understand **how their role contributes** to the **company's overall goals.**
- **Recognition and Rewards**
  - Publicly acknowledge and reward employee achievements. This can be through verbal praise, bonus programs, or promotion opportunities.
- **Growth and Development**
  - Provide opportunities for continuous learning and development. Offer training programs, mentorship opportunities.

Feature: Business Travel and Department

Attrition ->1

No Attrition -> 0

Feature: Education Field and Job Role

Attrition ->1

No Attrition -> 0

Correlation Matrix

Correlation Between Different Features

There are high correlations among some features:

- **PercentsalaryHike** and **PerformanceRating.**
- **YearsatCompany, YearsSinceLastPromotion**, and **YearsWithCurrManager**.

Feature Selection

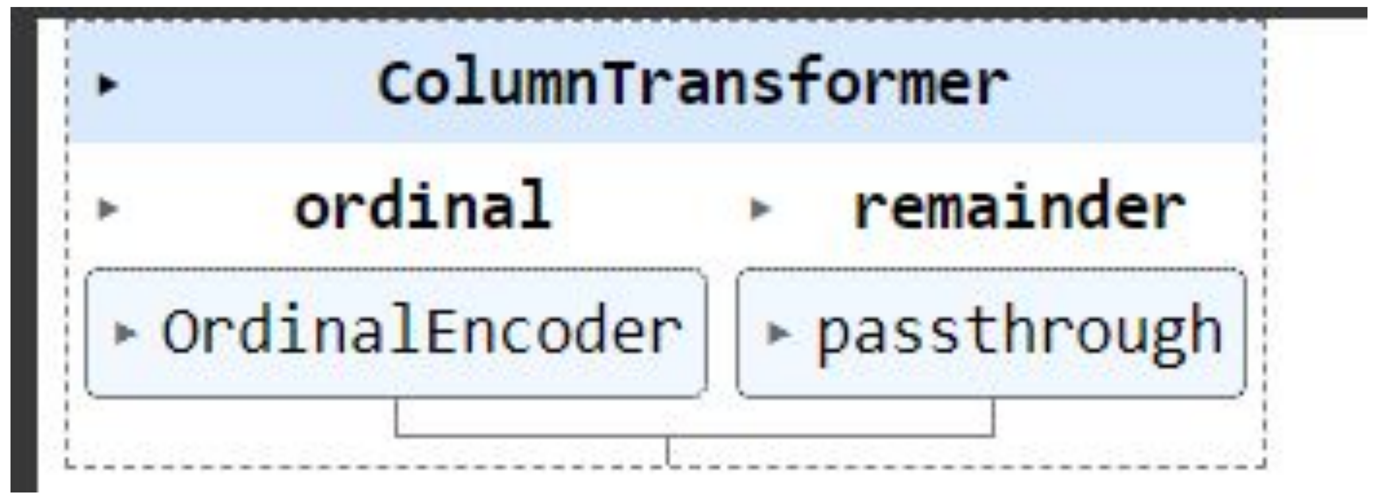
- Correlation matrix is used to identify highly correlated features. Removing highly correlated features can help improve ML model performance by avoiding multicollinearity.
- Correlation doesn’t imply causation.

All Features :

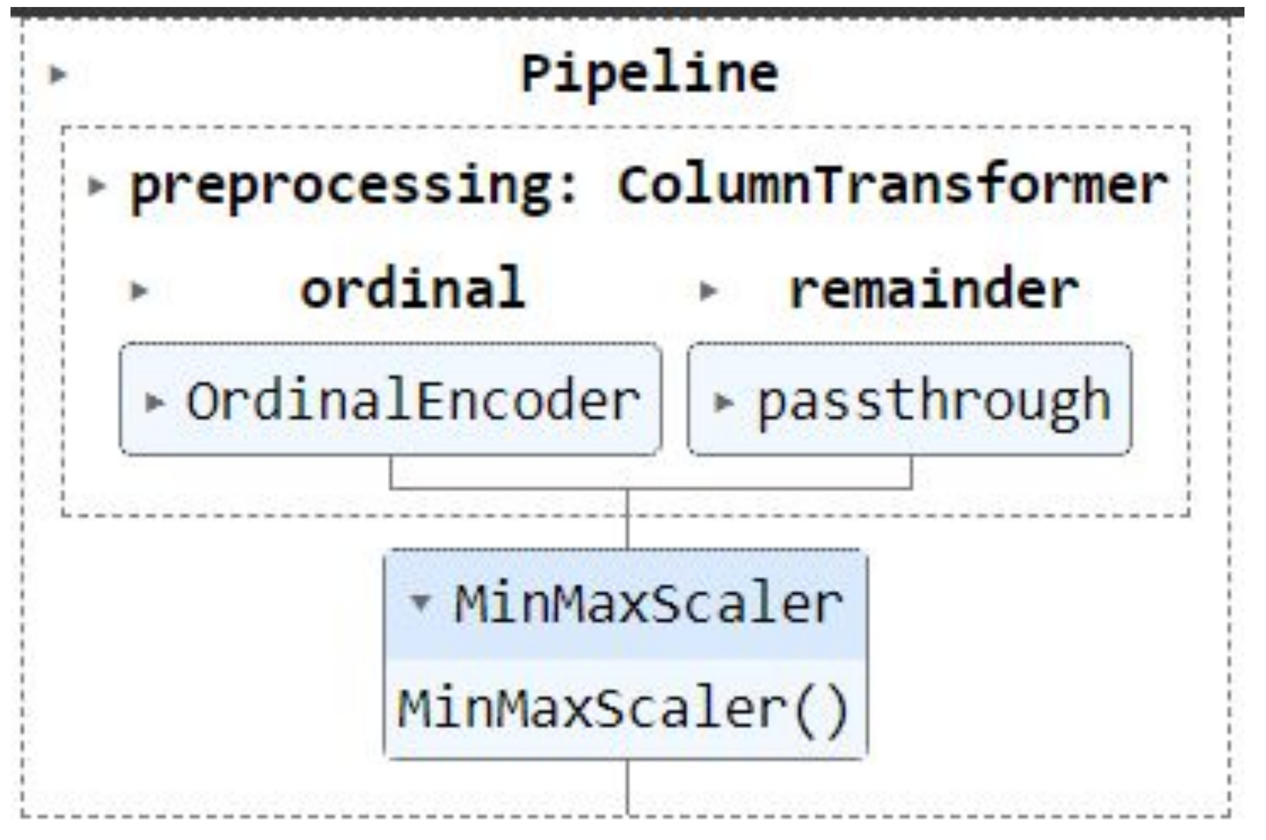
```
Features : 'Age', 'BusinessTravel', 'Department', 'DistanceFromHome', 'Education',
    'EducationField', 'Gender', 'JobLevel', 'JobRole', 'MaritalStatus',
    'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike',
    'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
    'JobInvolvement', 'EnvironmentSatisfaction', 'JobSatisfaction',
    'WorkLifeBalance'],
   dtype='object')
```

Workflow - ML Model

- Data Loading
    - Problem Scoping
- Data
    - Inspect
    - Data Acquisition and Understanding
        - Numerical and Categorical Features
    - Analysis - Univariate, Bivariate
    - Merging Datasets - general_data, manager_survey_data and employee_survey_data
    - Missing, NA Values, Outliers
- Data Analysis
    - Feature Engineering
- Categorical Features - Encoding



- Feature Scaling
    - Min-Max Scale or Standardize



- Resampling
    - SMOTE - Synthetic Minority Oversampling Technique
        - Unfortunately, no major improvements using SMOTE
- Data split - Training- Testing Data
- Classification
    - Confusion Matrix, Classification Report

Classifiers

Logistic Regression

Binary label either 0 or 1
Balanced data
Accuracy = 67%

Decision Tree

criterion : Entropy
max_depth : 15
min_samples_leaf : 1
min_samples_split : 2
splitter : best

Binary label either 0 or 1
Balanced data
Accuracy =98%

Random Forest

criterion : gini
max_depth : None
min_samples_leaf : 1
min_samples_split : 2

Binary label either 0 or 1
Balanced data
Accuracy = 100%

Performance of Classifiers

Prediction	Data	0	1
Precision	Test	0.92	0.26
	Training	0.92	0.30
Recall	Test	0.67	0.67
	Training	0.68	0.69
F1-score	Test	0.78	0.38
	Training	0.78	0.42

Prediction	Data	0	1
Precision	Test	1	0.87
	Training	1	0.96
Recall	Test	0.97	0.97
	Training	0.99	0.99
F1-score	Test	0.99	0.93
	Training	1	0.98

Prediction	Data	0	1
Precision	Test	0.99	1
	Training	1	1
Recall	Test	1	0.98
	Training	1	1
F1-score	Test	0.99	0.99
	Training	1	1

Close to 1 -> best, close to 0-> worse | Test data is unseen data given to ML model to predict attrition | Training data, on which ML model is trained

Performance of Classifiers

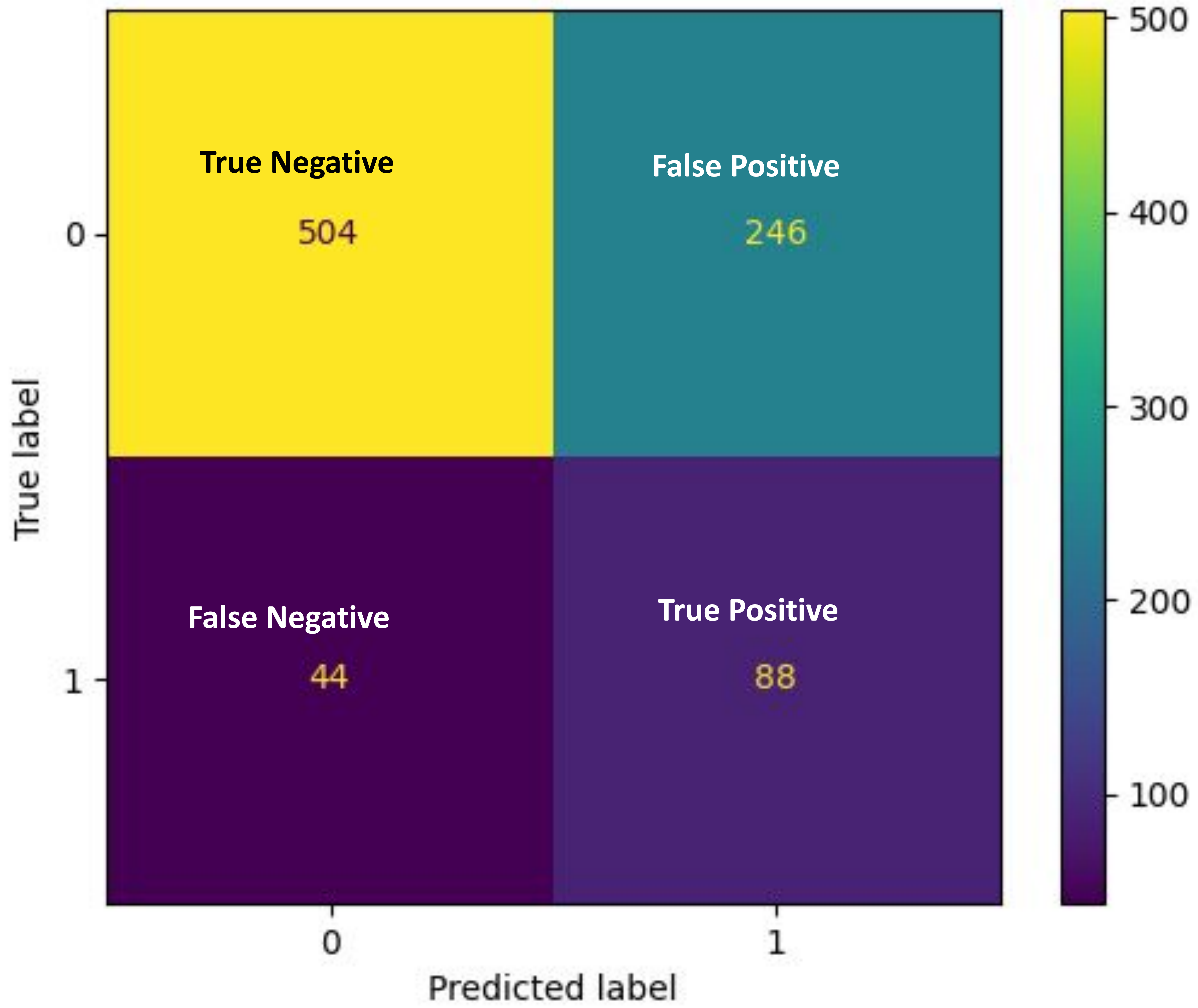
Precision, recall, and F1 score are all metrics used to evaluate the performance of a classification model.

Precision:
Focuses on positive predictions: Measures the proportion of predicted positive cases that are actually correct.
Calculation: Precision = True Positives / (True Positives + False Positives)
Interpretation: A high precision indicates that most of the positive predictions made by the model were accurate. However, it doesn't tell you anything about the negative predictions.

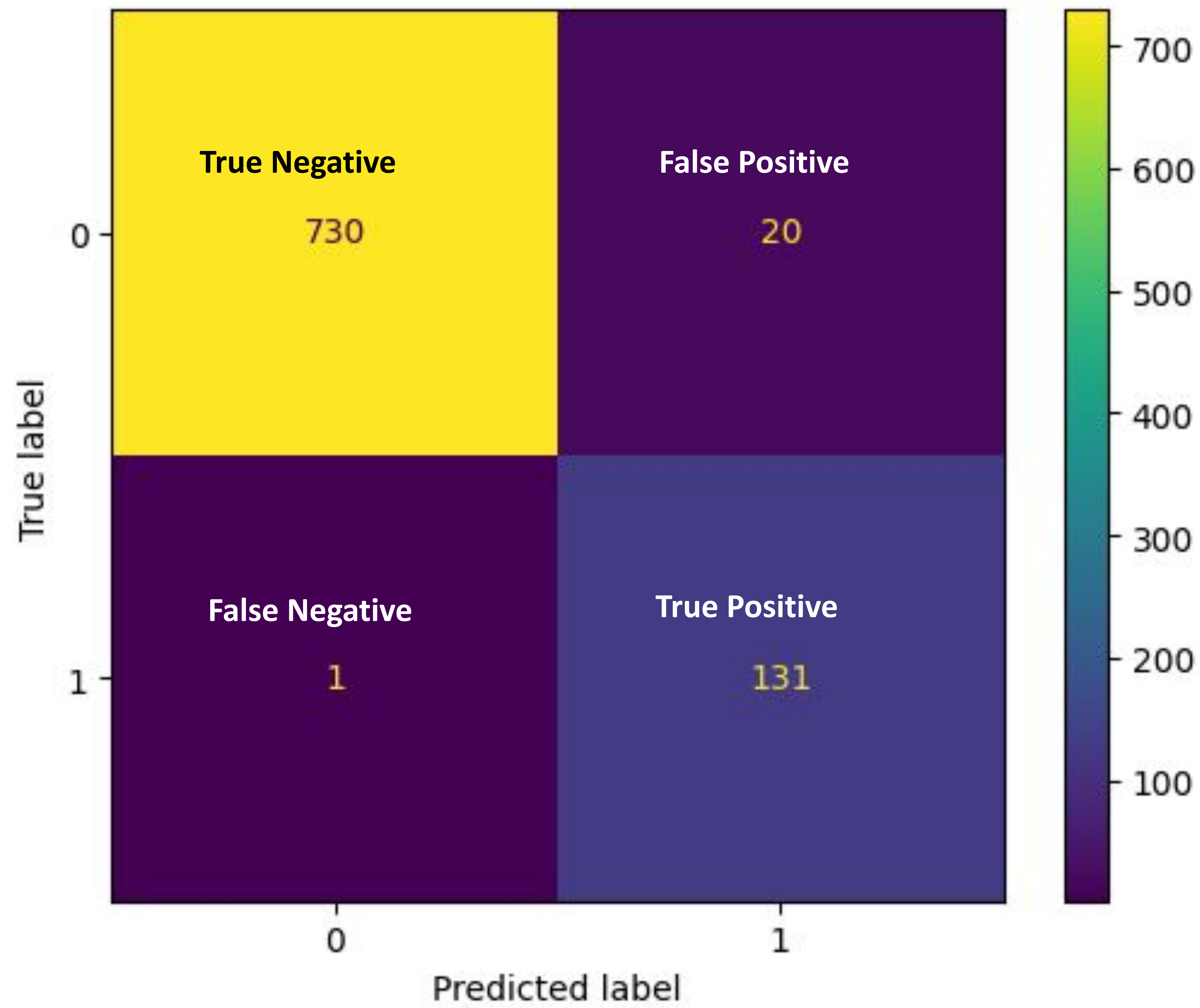
Recall:
Focuses on true positives: Measures the proportion of actual positive cases that were correctly identified by the model.
Calculation: Recall = True Positives / (True Positives + False Negatives)
Interpretation: A high recall indicates that the model identified most of the actual positive cases. However, it doesn't tell you anything about the false positives (incorrectly predicted positive cases).

F1 Score:
Combines precision and recall: It's a harmonic mean of precision and recall, providing a single metric to balance both aspects.
Calculation: F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
Interpretation: A high F1 score indicates that the model is performing well at both identifying true positives and minimizing false positives. It's a good overall measure for imbalanced datasets where one class might be more important than the other.

Confusion Matrix



Logistic Regression



Decision Tree

- Both **Recall (True Positives)** and **F1 Score** are important metrics to consider in attrition prediction.
- F1 score combines precision and recall.
- Type-II Error (False Negative) and Type-I Error(False Positive) of Decision Tree classifier are better than Logistic Regression Classifier.

Conclusion

- Decision trees are often better suited than logistic regression where we have complex and non-linear relationship between features and the target variable.

- Whereas logistic regression models the relationship between the features and the outcome variable as a linear function.

In this HR dataset , we will go with the **decision trees**.

- Better Recall and F1 Score.
- It can approximate non-linear relationship between features and the target variable.
- Feature selection is done to remove multicollinearity among features for ML model to perform best.
- Precision ->87%, Recall -> 97% and F1 Score ->93%.
    - Here we will consider **Recall** and **F1 Score** metrics more important to predict the attrition.

Conclusion (Contd.)

By leveraging a robust and generalizable machine learning model, HR can:

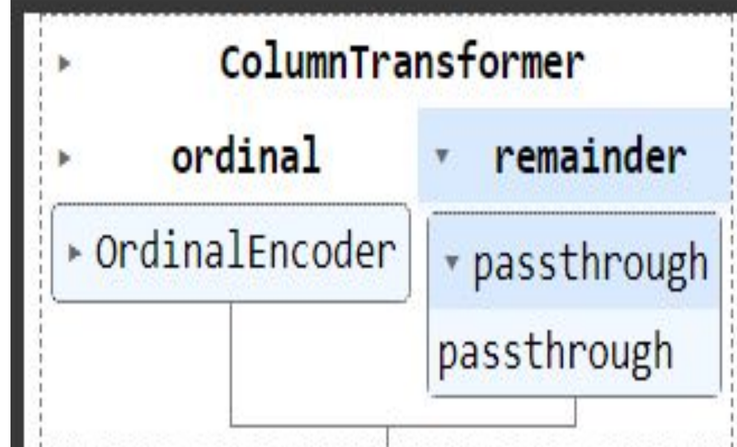
- Proactively identify employees at risk of leaving.
- Develop targeted interventions and retention programs to address specific employee needs.
- Improve overall employee satisfaction and morale.
- Reduce the negative impacts associated with employee churn.

Backup Slides

Workflow - ML Model

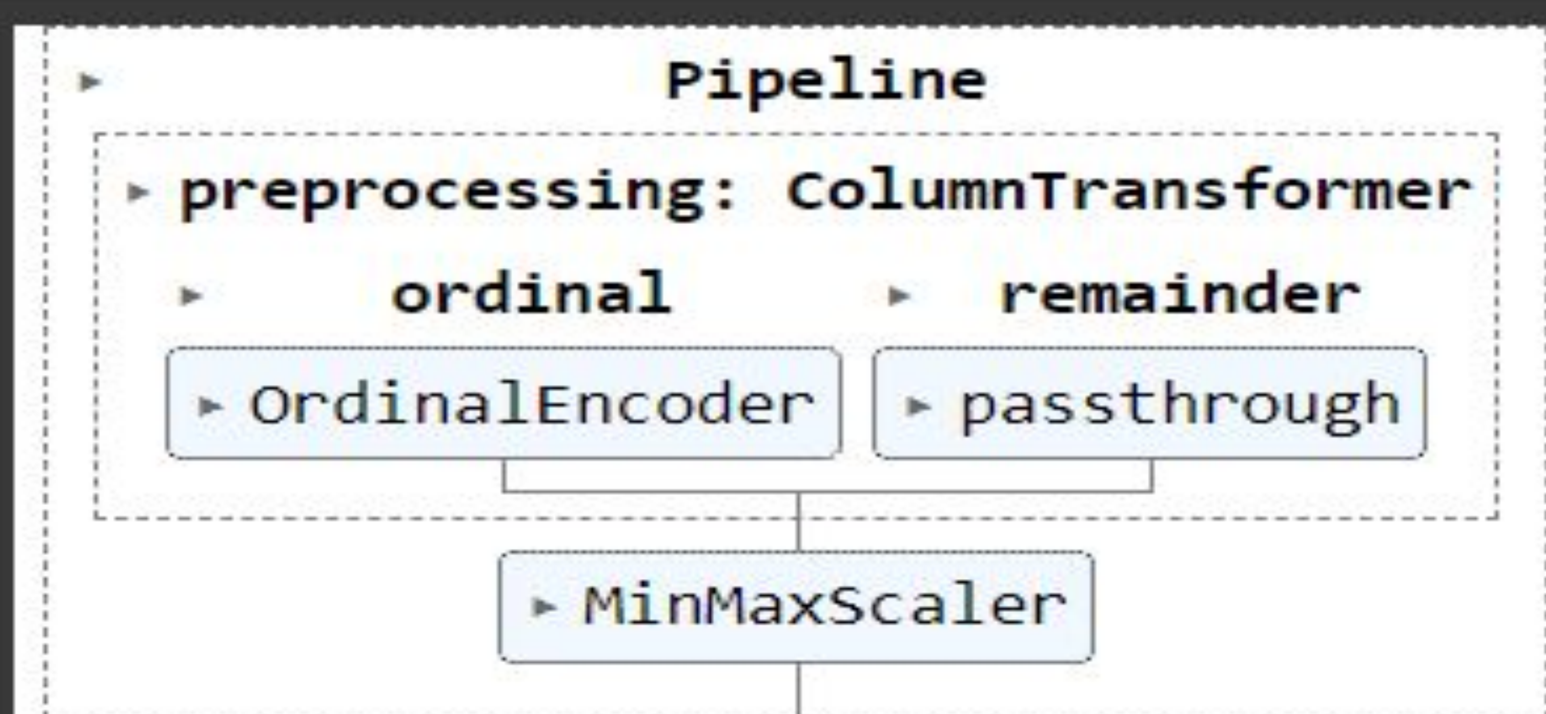
```python
encoder = ColumnTransformer (transformers=[
        #('ohe', OneHotEncoder(drop='first', sparse=False),
         ('ordinal', OrdinalEncoder(),['BusinessTravel', 'Department', 'EducationField', 'Gender','JobRole', 'MaritalStatus']),  # One-hot encoding with drop_first=True for 'Gender'
    ],
    remainder='passthrough'  # Keep the other columns unchanged
)

# setting to get a pandas df
encoder.set_output(transform='pandas')
```




```python
# Define the pipeline
pipe = Pipeline([
    ('preprocessing', encoder),  # Assuming `encoder` is your previously defined encoder
    ('scaling', MinMaxScaler()),  # Scaling step
    #('feature_selection', SelectKBest(score_func=chi2, k=15)),  # Feature selection step
])
```

```python
pipe.fit(X_train_copy, y_train_copy)
```




```python
[ ]  # Transform both the training and testing data
X_train_transformed = pd.DataFrame(pipe.transform(X_train_copy))
X_test_transformed = pd.DataFrame(pipe.transform(X_test))
```

Classifiers

K-Nearest Neighbor (KNN)

weights : uniform
n_neighbors : 15
scaler : min-max

Binary label either 0 or 1
Balanced data

Prediction	Data	0	1
Precision	Test	0.85	0.63
	Training	0.87	0.65
Recall	Test	0.99	0.13
	Training	0.99	0.13
F1-score	Test	0.92	0.22
	Training	0.91	0.21

Logistic Regression

Binary label either 0 or 1
Balanced data

Prediction	Data	0	1
Precision	Test	0.92	0.26
	Training	0.92	0.30
Recall	Test	0.67	0.67
	Training	0.68	0.69
F1-score	Test	0.78	0.38
	Training	0.78	0.42