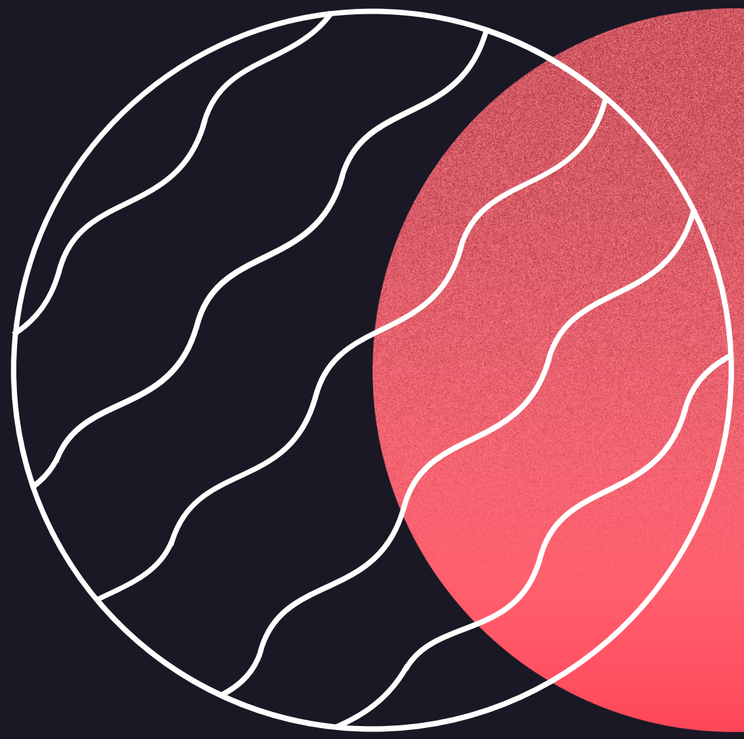


IBM HR Analytics

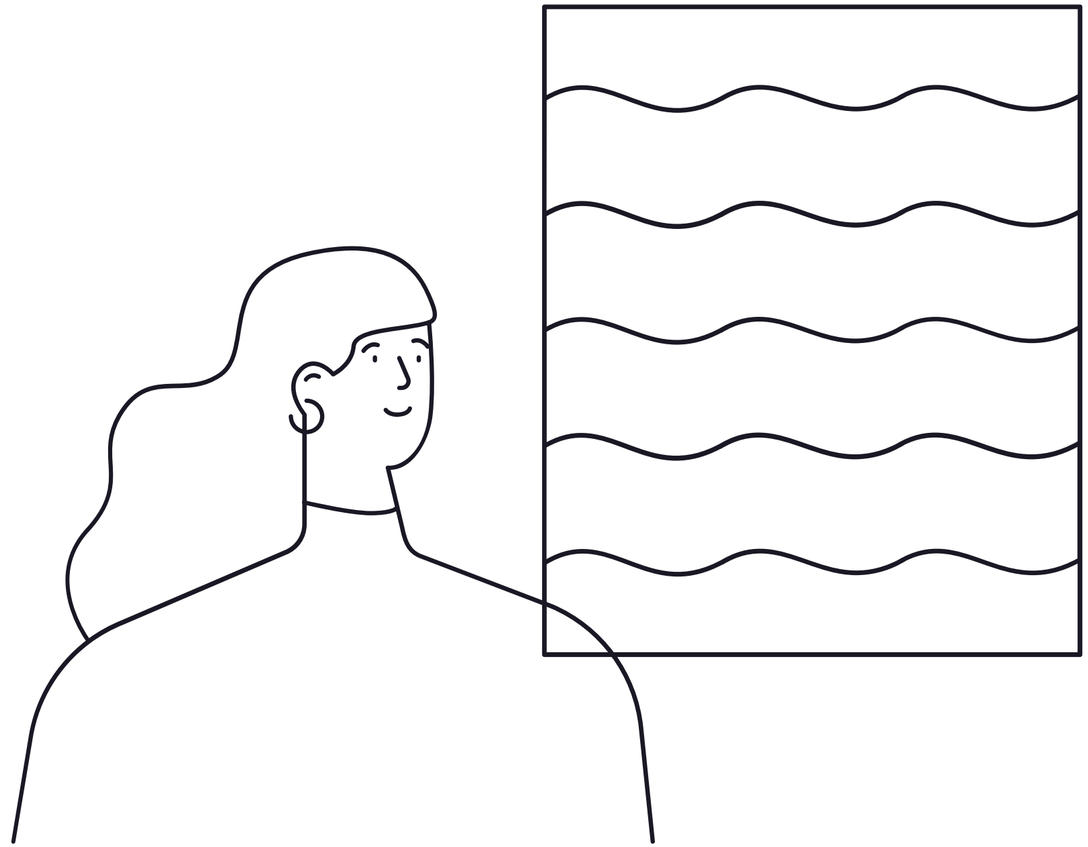
Team 4: Emory's Next Top Models
Neha Bansal
Emmy Fortunato
Boping Zhang

ATTRITION

- Predict the likelihood that an existing employee will quit

BUSINESS VALUE

- Expensive in terms of money and time to train new employees
- Loss of experienced employees
- Impact on profit because of business disruption



THE DATA

PREDICTIVE MODEL

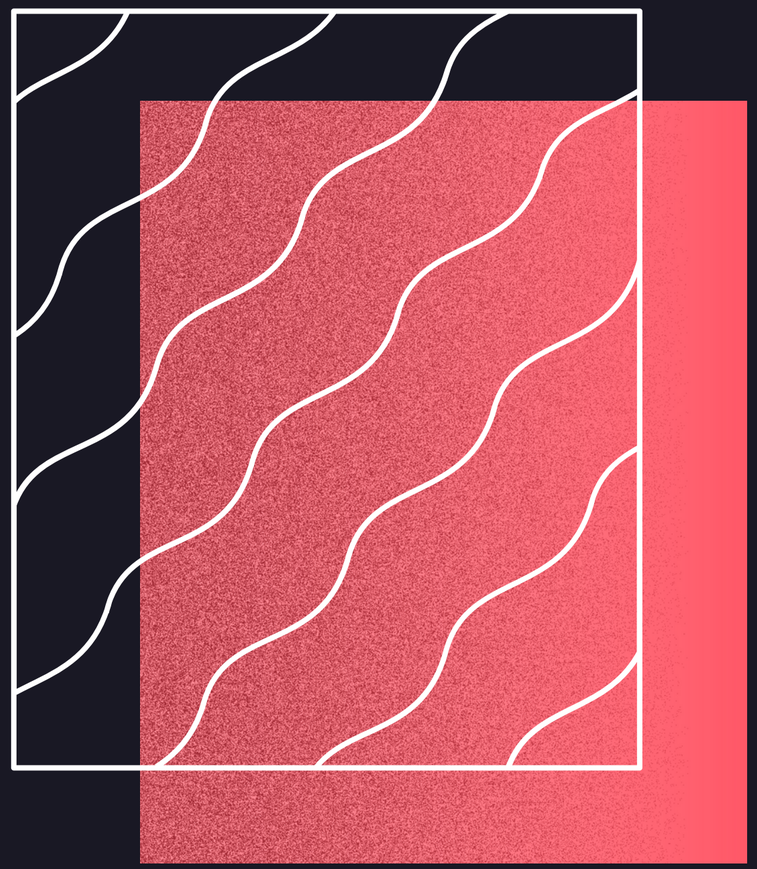
The business question that we want to answer is:

- Will employee "X" quit?

So the data science questions that we want to answer are:

- Which variables are the best predictors of employee attrition at IBM?
- Based on these variables, can we expect that employee "X" will voluntarily leave IBM?

To answer these questions, we will use the data available to us to build a predictive model.



TARGET VARIABLE

In the case of our dataset, Attrition is a binomial variable, with a value of 0 indicating that the employee has not left the company and a value of 1 indicating that the employee has voluntarily left the company.

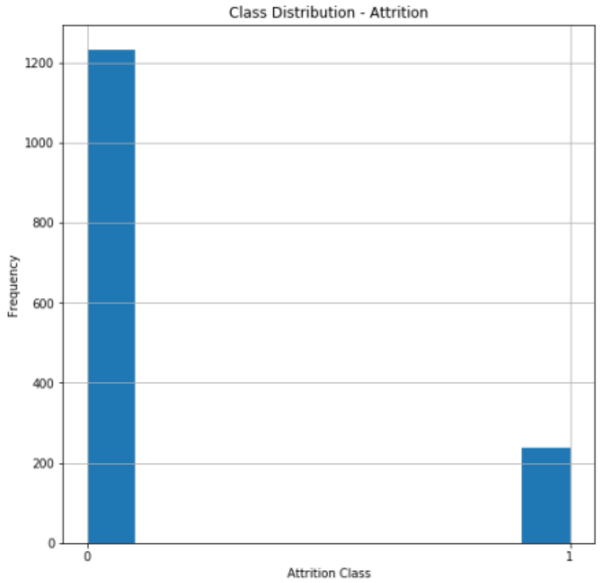
FEATURES

The X-variables in the dataset are made up of:

- numerical variables: measures of time in the workforce, wages earned, and some demographics
- nominal categorical variables: objective classifications by job, education, and demographic
- ordinal categorical variables: subjective rankings of job-related engagement, satisfaction, and performance

DATA PREPARATION

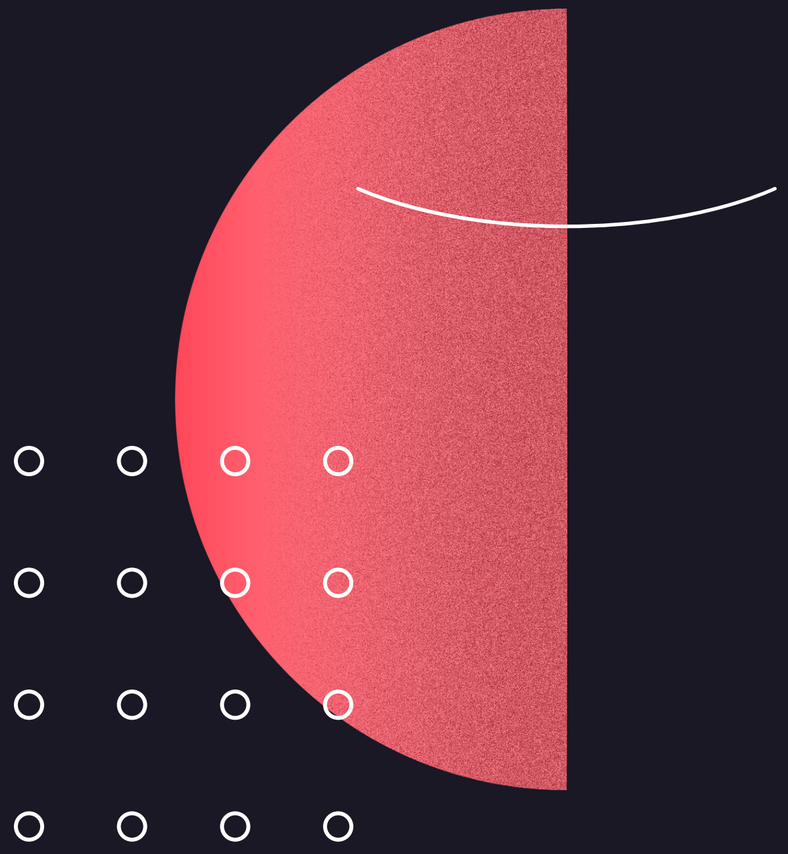
```
# Ordinal Variable: BusinessTravel
unique_travel = X['BusinessTravel'].unique()
print(unique_travel)
size_mapping = {'Travel_Frequently': 3,
                'Travel_Rarely': 2,
                'Non-Travel': 1}
X['BusinessTravel'] = X['BusinessTravel'].map(size_mapping)

# Binary Variable: Gender
X["Gender"] = X["Gender"].replace(["Male"], 1)
X["Gender"] = X["Gender"].replace(["Female"], 0)
unique_gender = X['Gender'].unique()
print(unique_gender)

# Binary Variable: Over18
X["Over18"] = X["Over18"].replace(["Y"], 1)
X["Over18"] = X["Over18"].replace(["N"], 0)
unique_over18 = ibm_hr["Over18"].unique()
print(unique_over18) #can remove!

# Binary Variable: OverTime
X["OverTime"] = X["OverTime"].replace(["Yes"], 1)
X["OverTime"] = X["OverTime"].replace(["No"], 0)
unique_overtime = X["OverTime"].unique()
print(unique_overtime)
```



Our approach to preparing the data involved a few key steps:

1) Transform text variables to numeric variables.
2) Remove uninformative attributes.
3) Reduce multicollinearity.
4) Balance the data.
5) Conduct feature engineering.

DATA ENGINEERING



We add 5 new estimated features based on existing features:

- Annual Salary = 12 * Monthly Income
- Annual Compensation = 12 * Monthly Rate
- Bonus = Annual Compensation - Annual Salary
- Fidelity = Number of Company Worked / Total Working Years
- Performance of Job Satisfaction = Average of (Job Satisfaction + Job Involvement)

MODEL

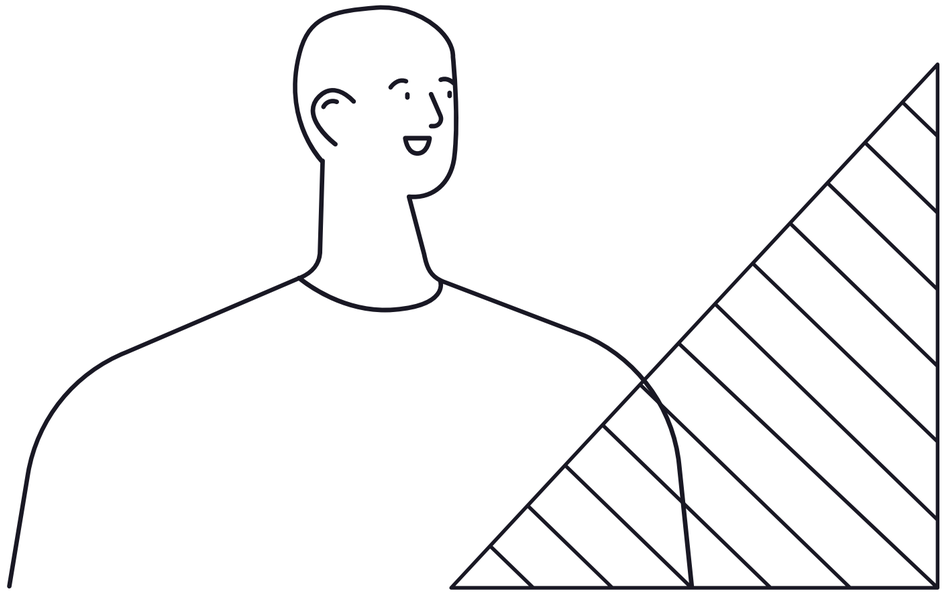
Model: Decision Tree Classifier

Advantages:

- Consider all possible outcomes of a decision
- Assign specific values to each problem
- Easy to use
- Easy to understand for non-professionals
- The deeper branch can identify if an employee will leave the company

Disadvantage:

- Overfitting

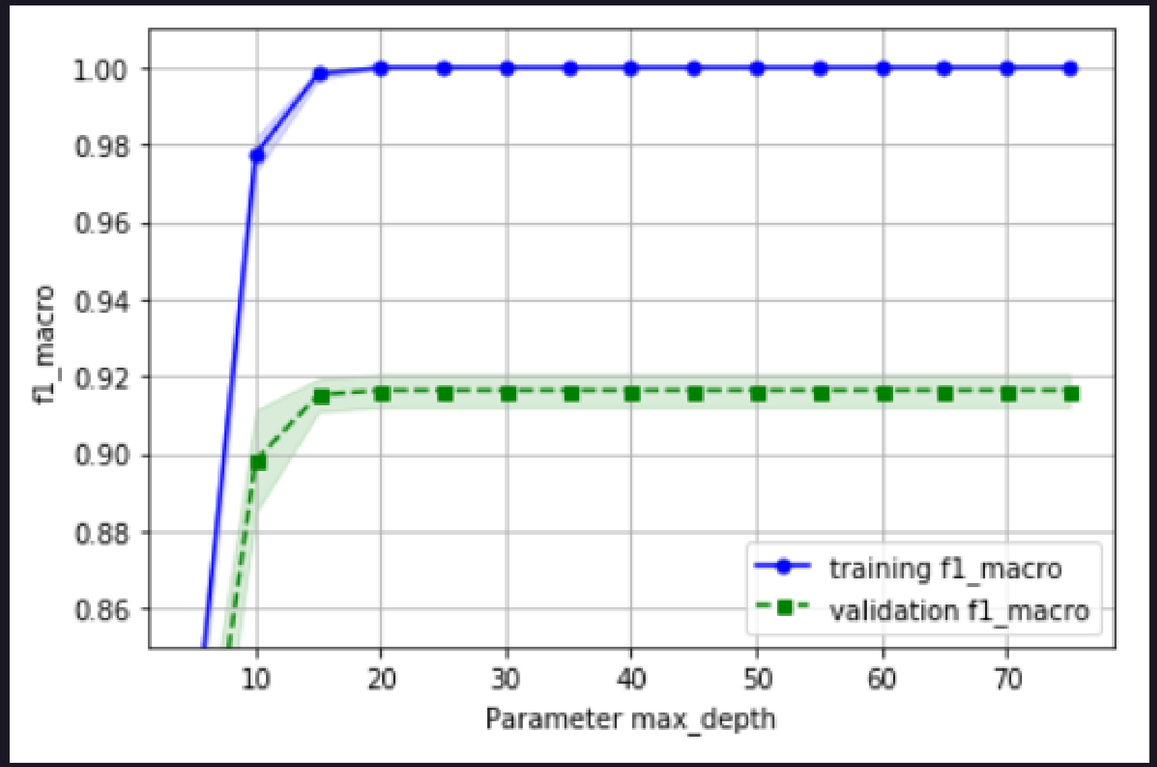


EVALUATION

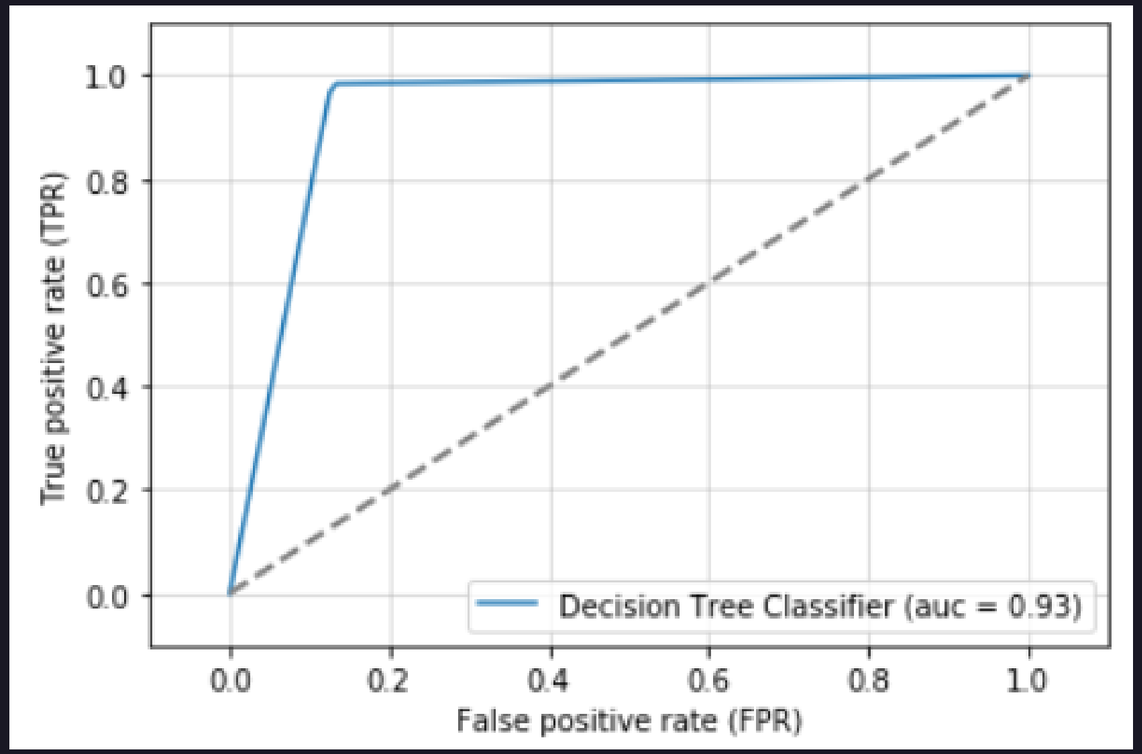
Non-nested f1_macro: 93.4%

Nested f1_macro: 92.07%

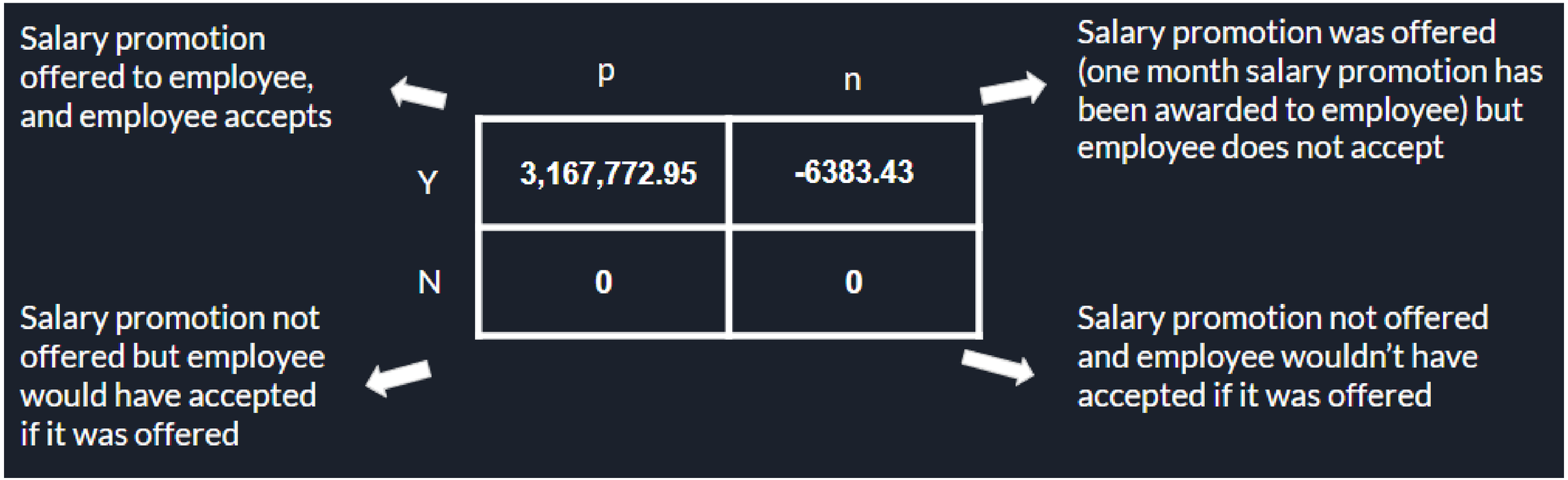
FITTING GRAPH



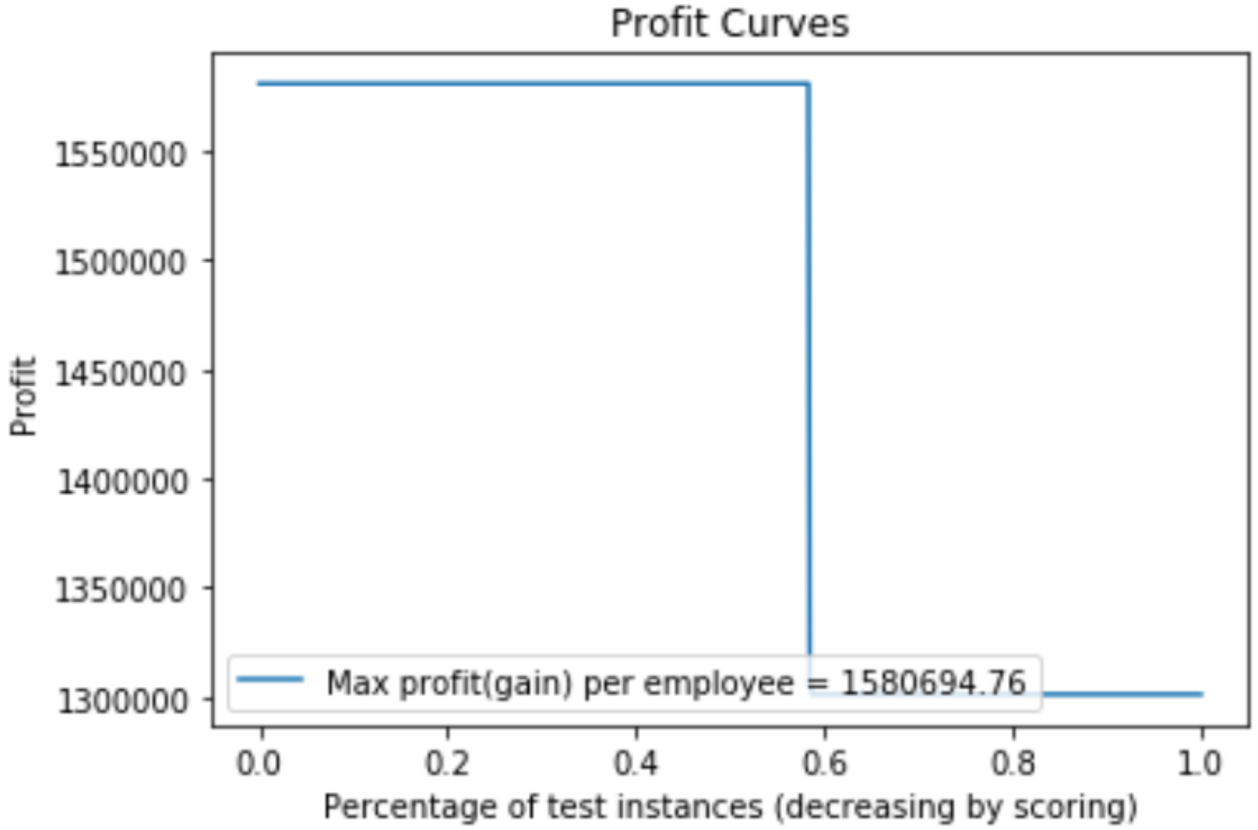
ROC CURVE

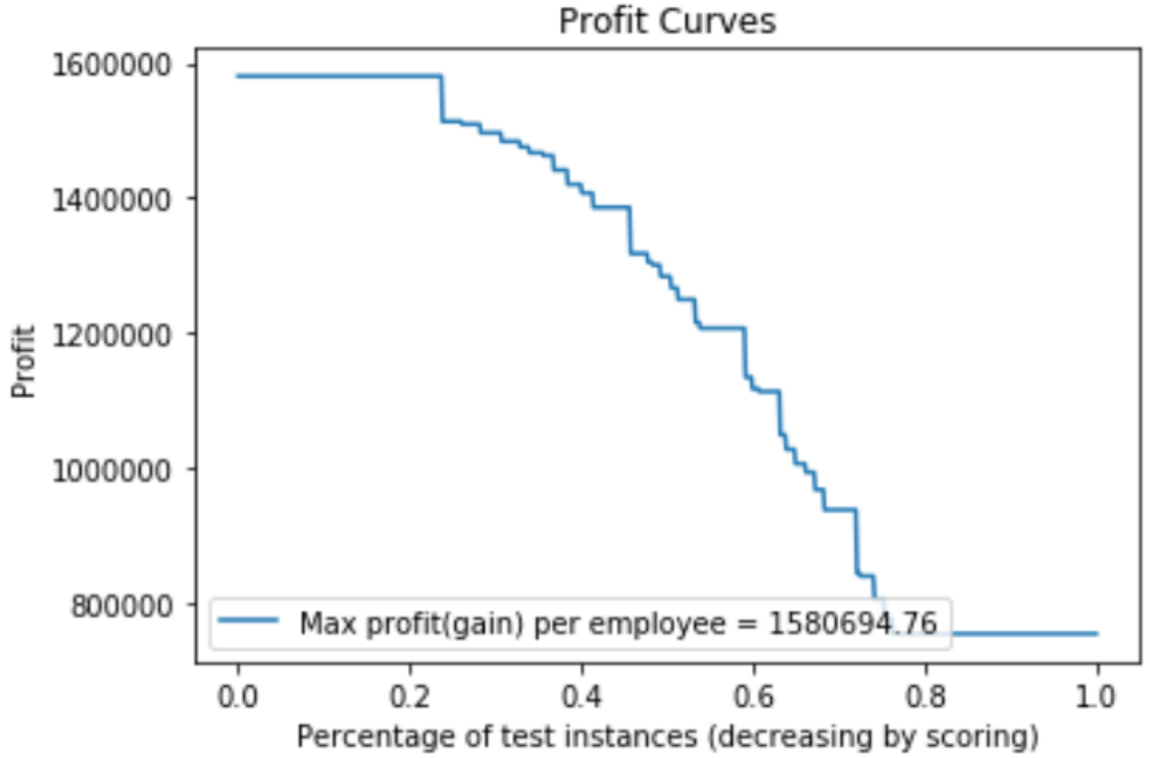


COST BENEFIT MATRIX

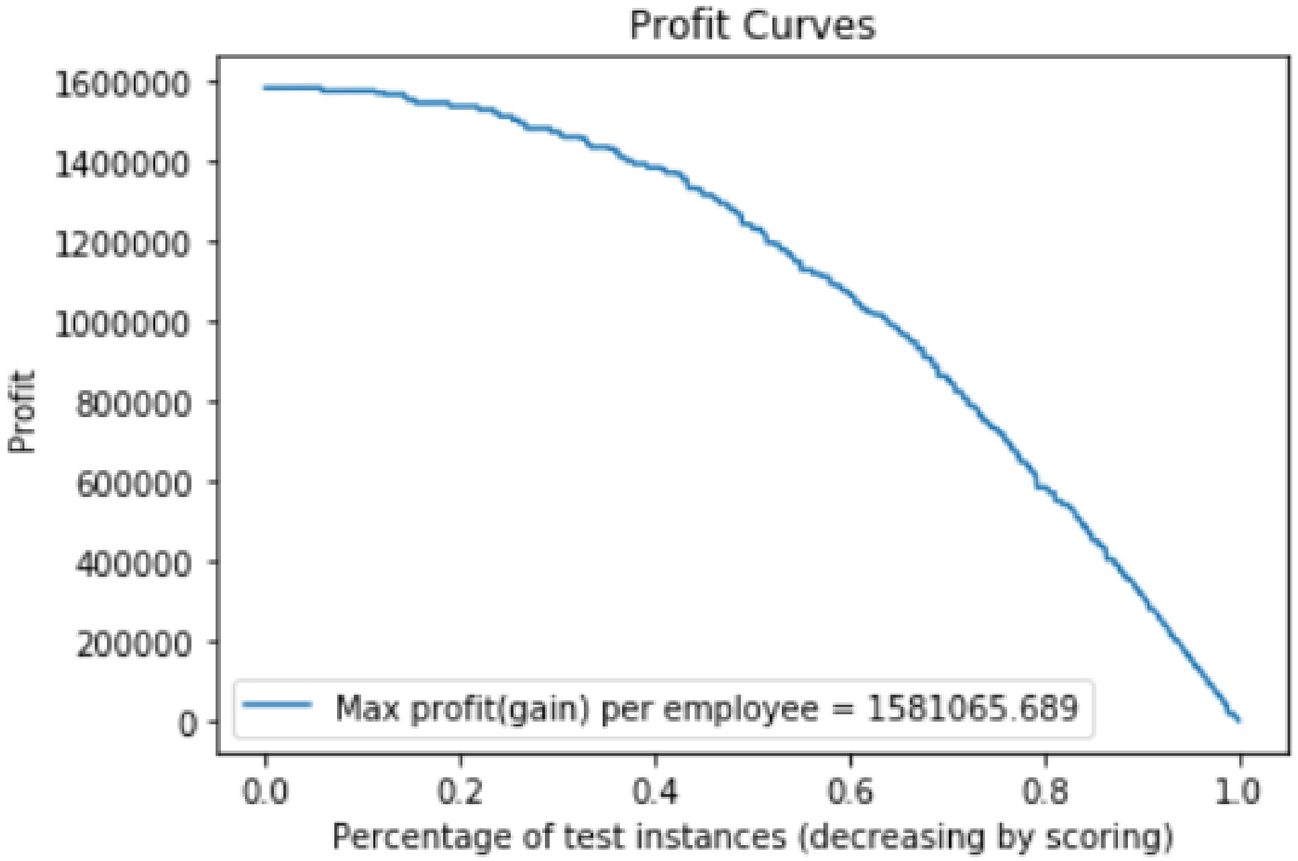


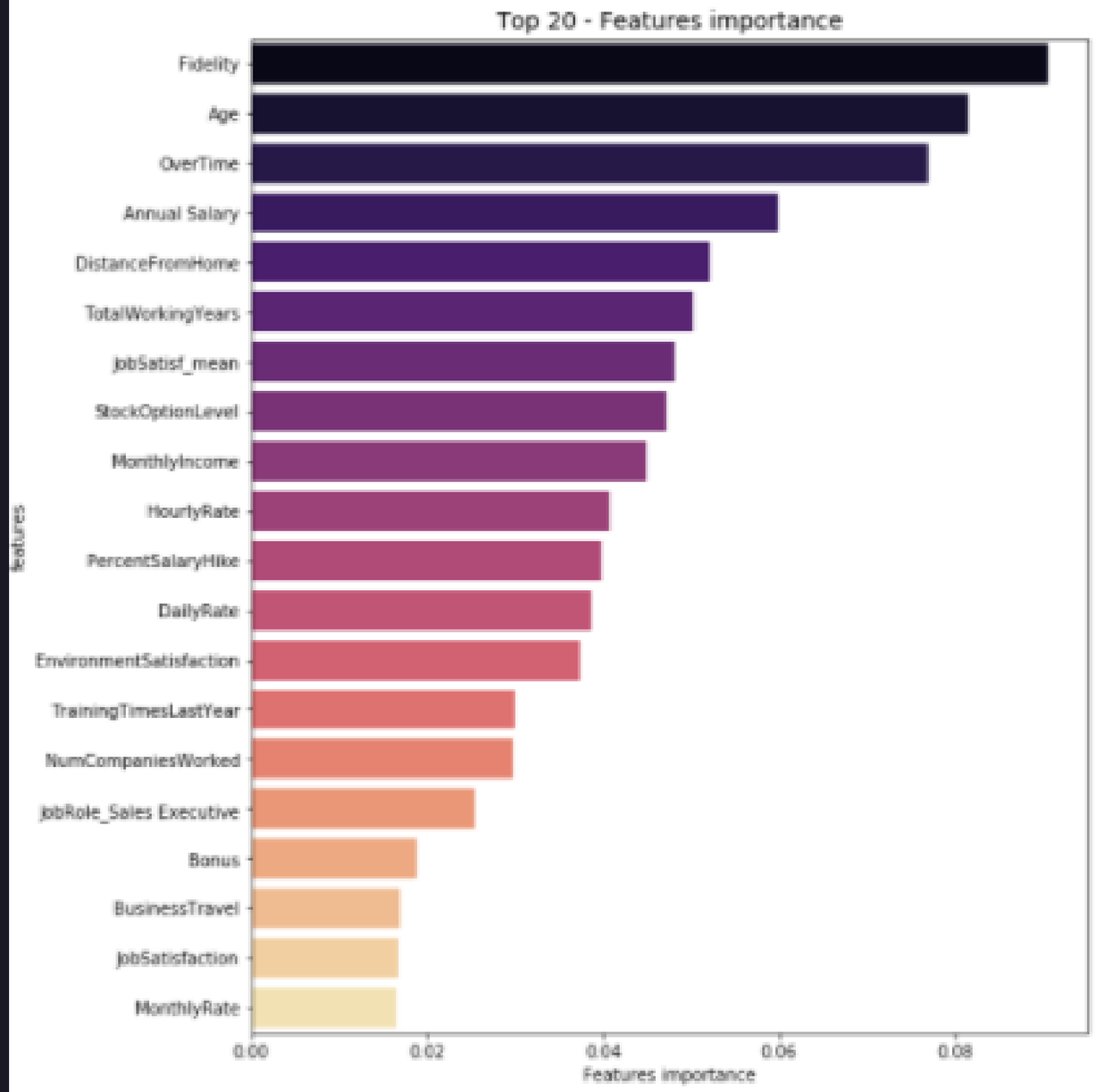
PROFIT CURVE - DECISION TREE





PROFIT CURVE - LOGISTIC REGRESSION





FEATURE IMPORTANCE

In order to know which features influence attrition more compared to others features. We used the (feature_importances) to identify the Top 20 essential features.

KEY TAKEAWAYS FOR HR

	Attrition	No Attrition
Fidelity	Lower	Higher
Annual Salary	Very low and mid-high	Mid and Very high
Age	27- 32	33 - 40
Performance of Job Satisfaction	Low and mid	High
Distance From Home	Far and near	Mid and near
Stock Option Level (4 is the highest)	1 and 3	2 and 4

MODEL DEPLOYMENT

Ways that HR can utilize the model:

- Most predictive attributes/key takeaways
- Actionable, intentional follow-up after Employee Pulse Surveys and Performance Evaluations
- Proactively plan for turnover

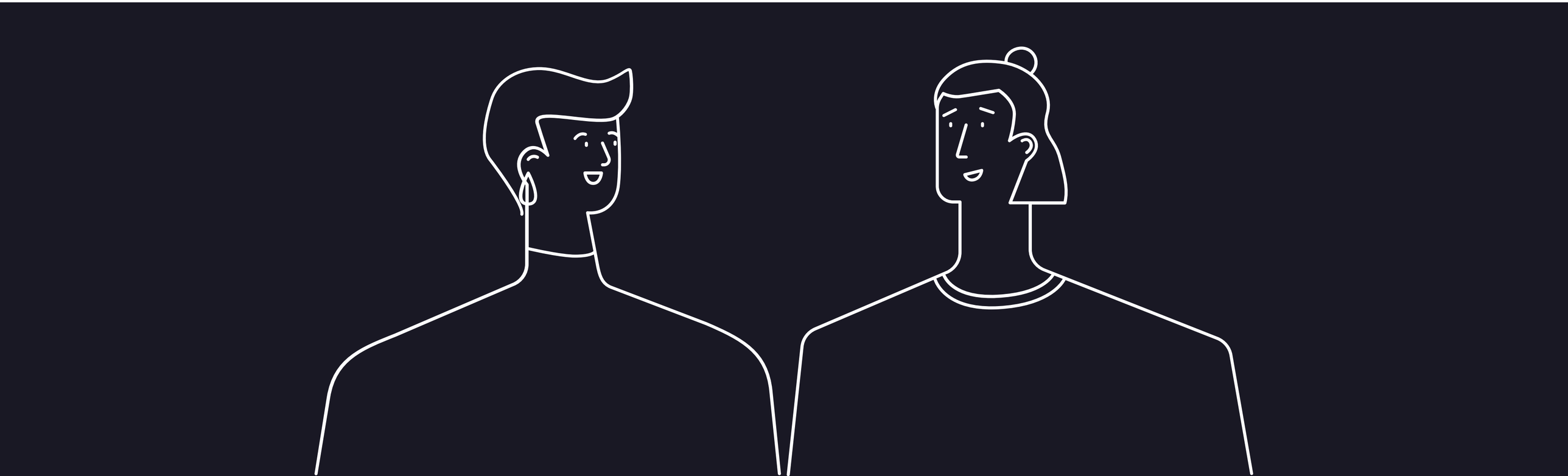
A word of caution:

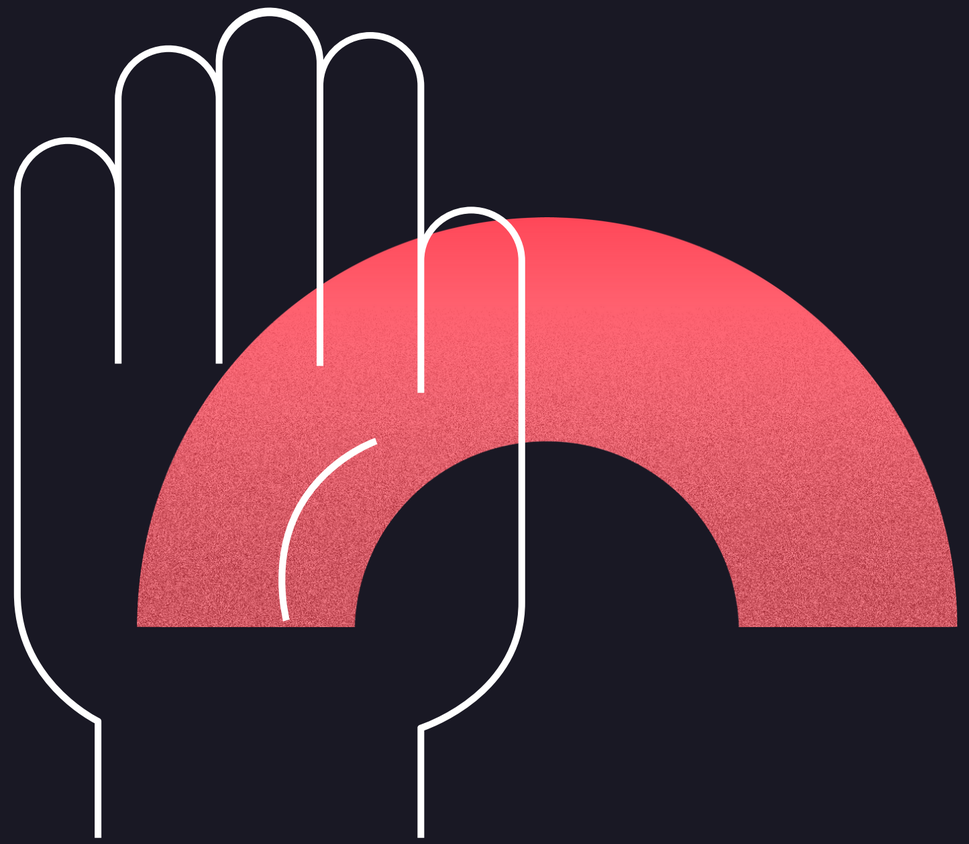
HR folks need to be aware that although the model is informative, it is not 100% accurate; it is important that those taking action based on model results are mindful to not treat employees differently based on what the model predicts about their future choices.

ETHICS

When it comes to building an ethical model, we want to be particularly mindful of variables that could be used in a discriminatory fashion. In the case of the data provided to us, variables that we may consider removing in favor of a more ethical model are:

- Age
- Gender
- TotalWorkingYears (as a proxy for Age)

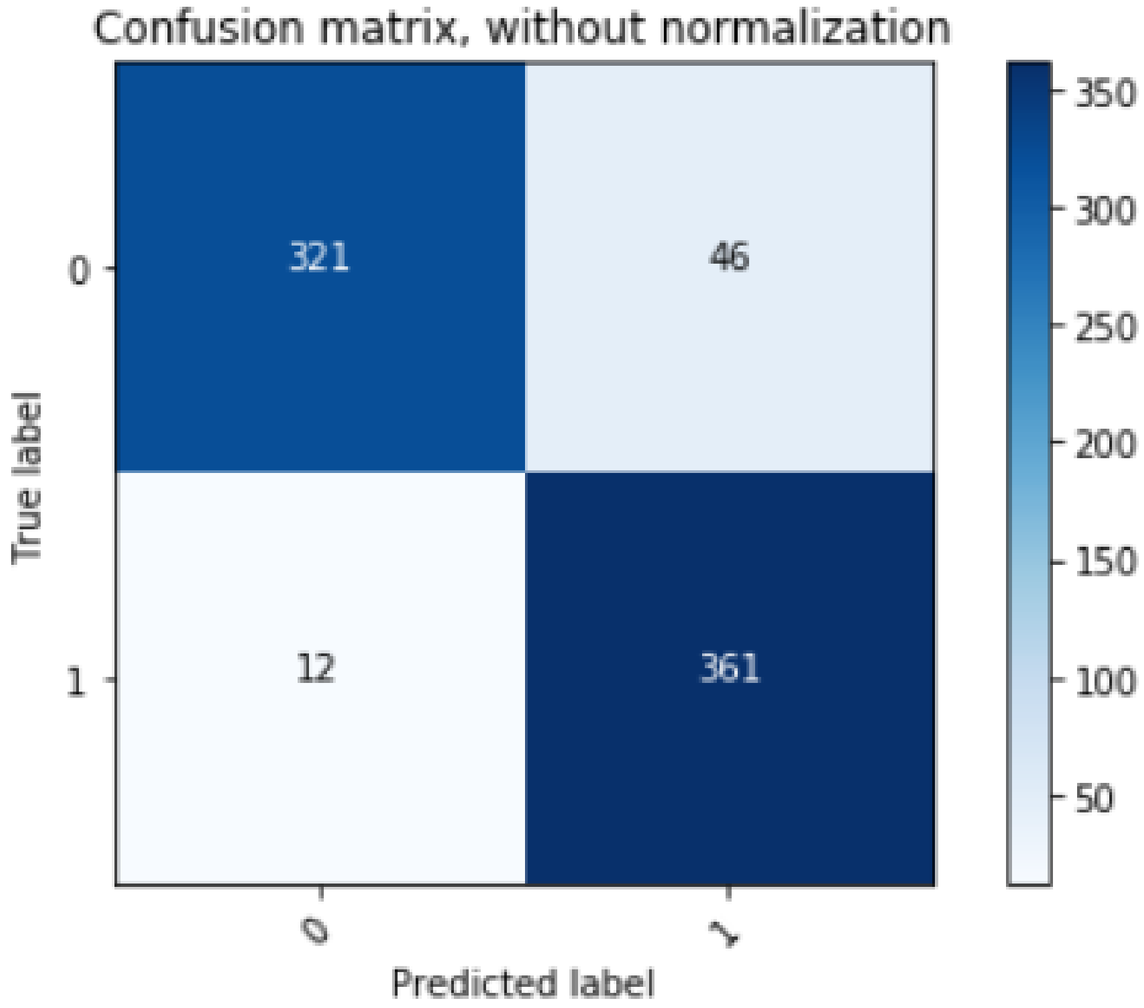




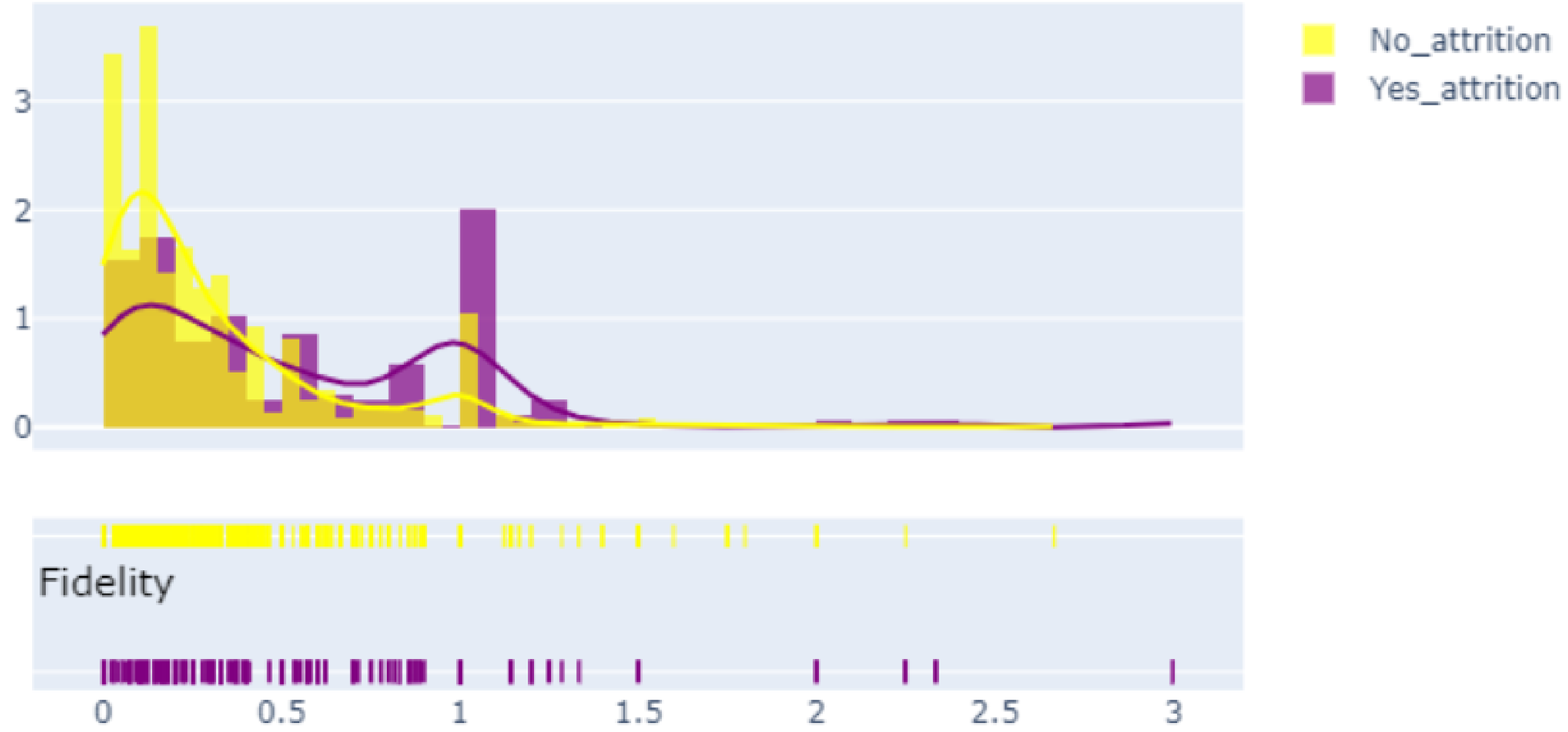
THANK YOU

APPENDIX

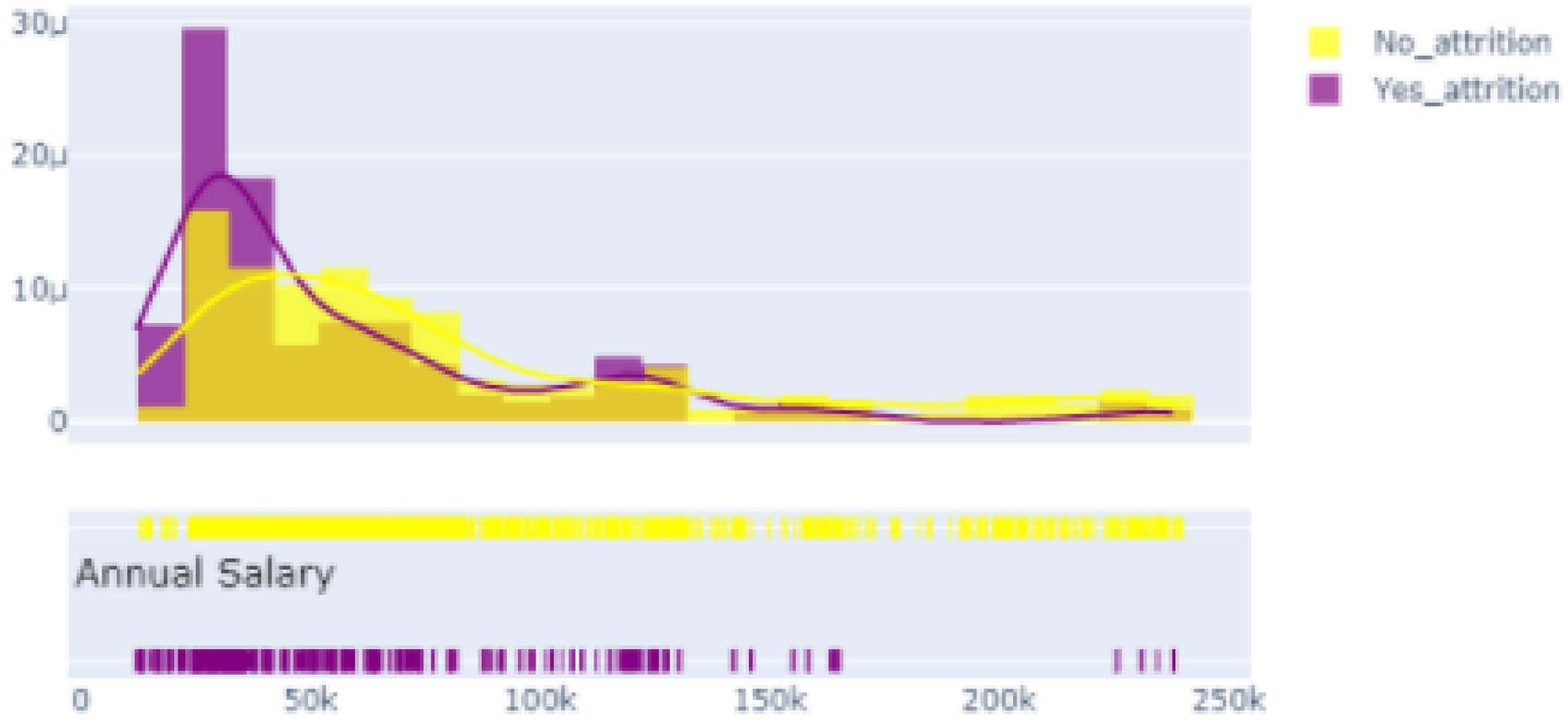
CONFUSION MATRIX



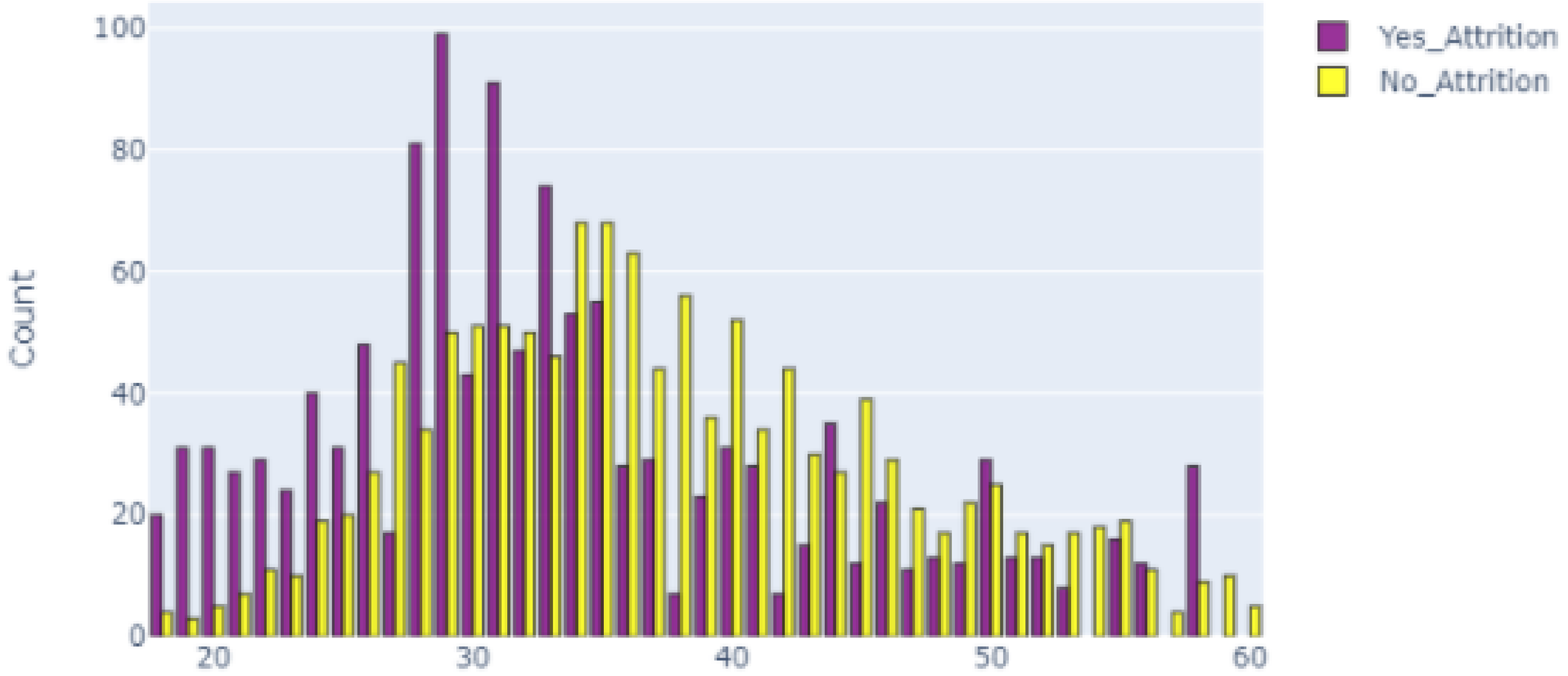
DISTRIBUTION OF FIDELITY (BASED ON WEIGHT)



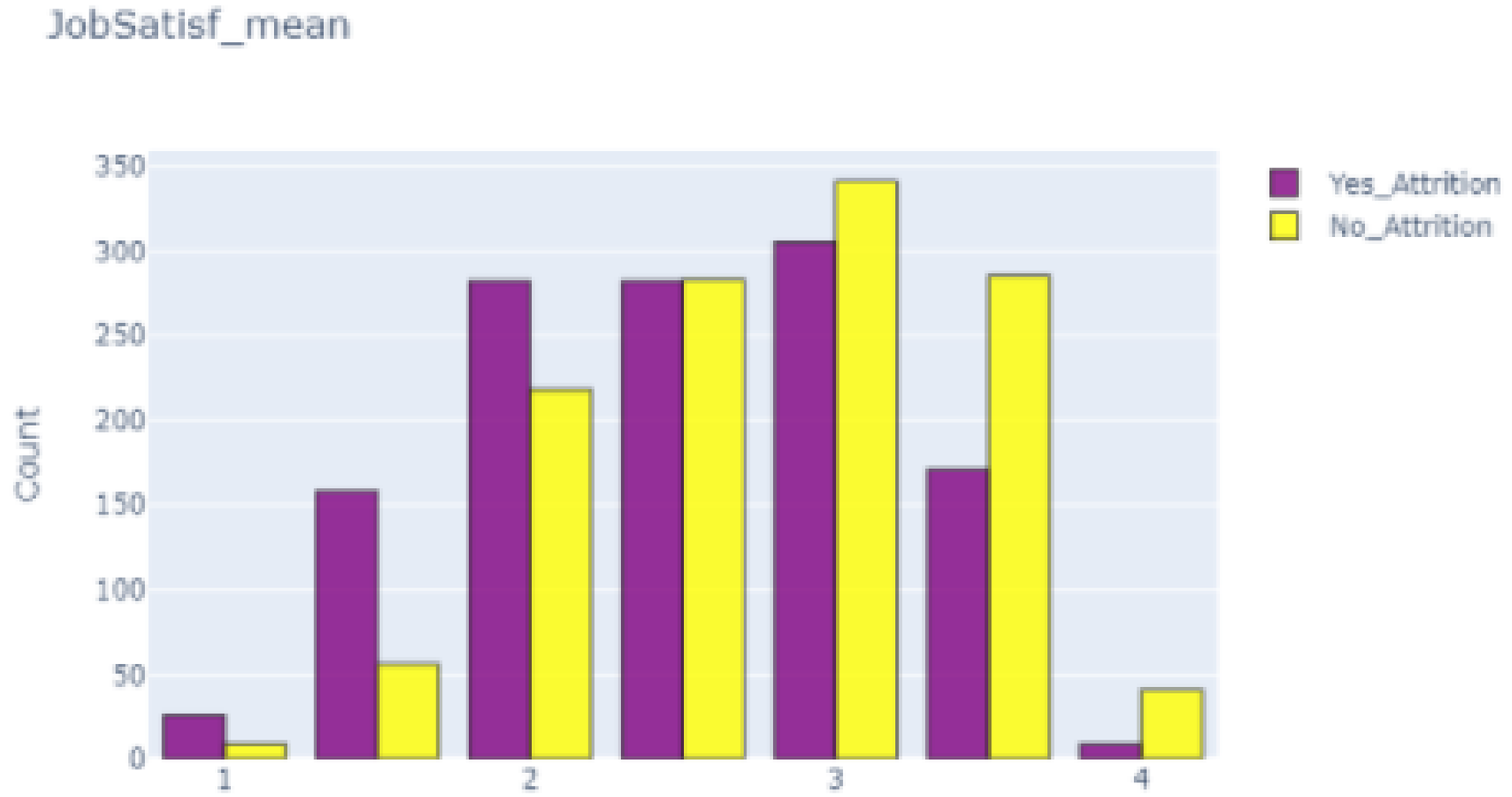
DISTRIBUTION OF ANNUAL SALARY (BASED ON WEIGHT)



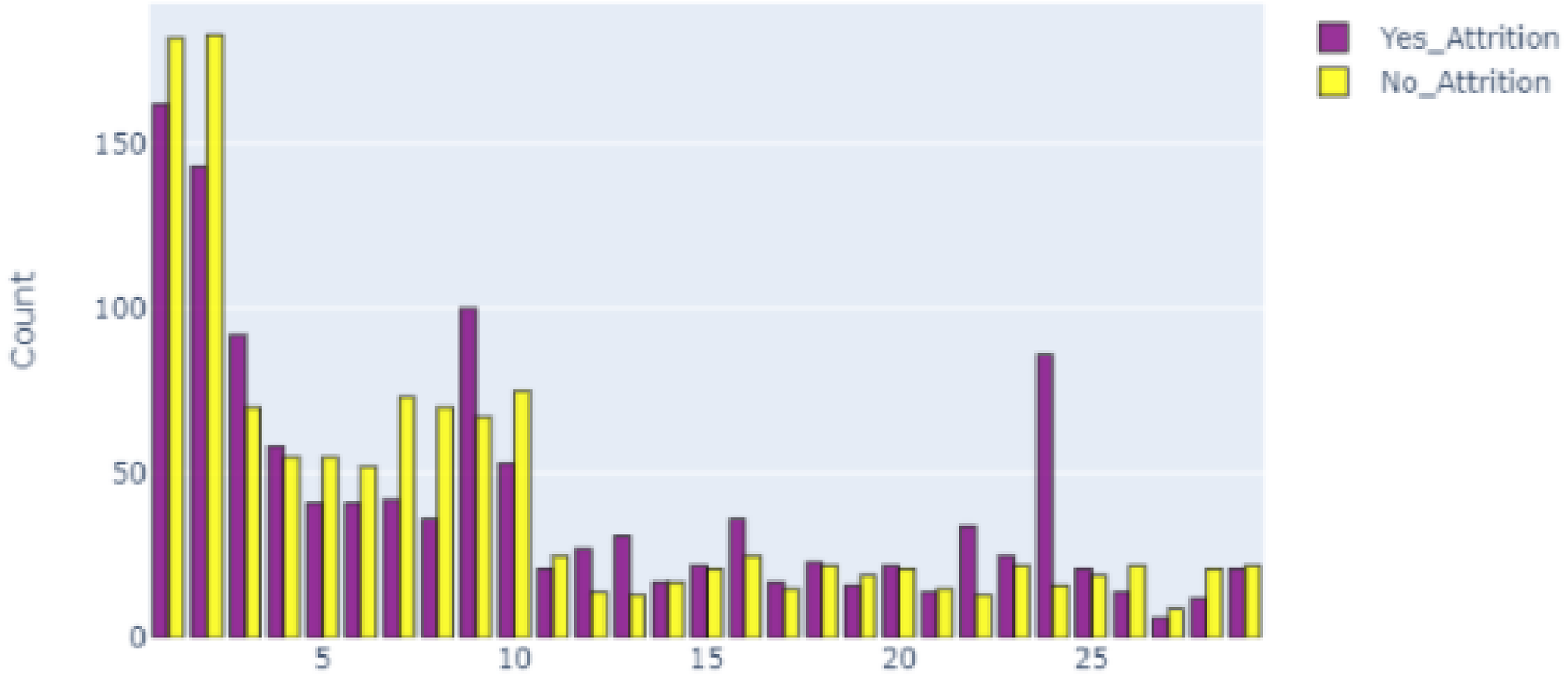
COUNT OF ATTRITION SEGMENTED BY AGE



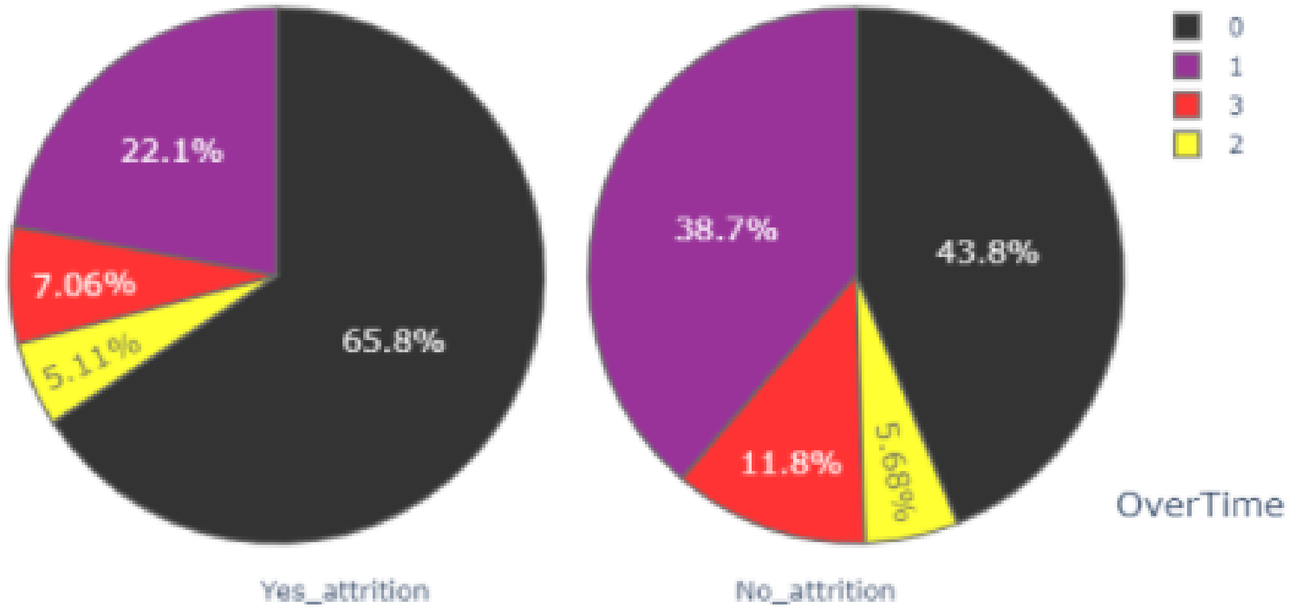
COUNT OF ATTRITION SEGMENTED BY JOB SATISFACTION AND JOB INVOLVEMENT



COUNT OF ATTRITION SEGMENTED BY DISTANCE FROM HOME



StockOptionLevel distribution in employes attrition



OverTime distribution in employes attrition