



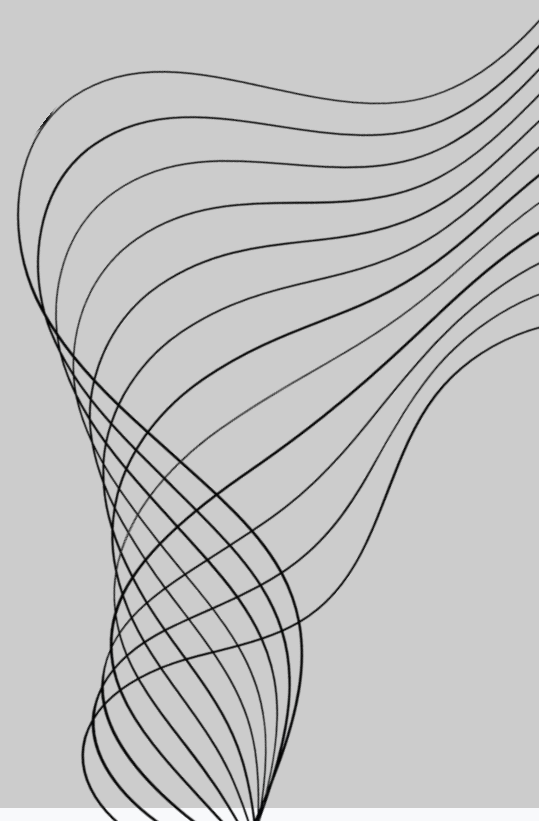
Employee Churn Prediction

*Predicting what causes employees to resign
and strategies on how to retain them*

BANA 273: Machine Learning Analytics

Team 12A

Anna Haroutounian
Viraj Vijaywargiya
Danqi Zheng
Yuh-Shin Yen
Zhiwei Lu



Overview



01

INTRODUCTION

02

DATA DESCRIPTION

03

MODELING & DATA ANALYSIS

04

INSIGHTS INTO EMPLOYEE CHURN

05

STRATEGIES AND RECOMMENDATIONS

06

CONCLUSIONS

Introduction



Our business case is to leverage employee data to identify key factors that influence employee churn, to predict at-risk employees and to develop actionable strategies to enhance retention and prevent churn.



Results: The organization can retain top talent and improve overall employee satisfaction.

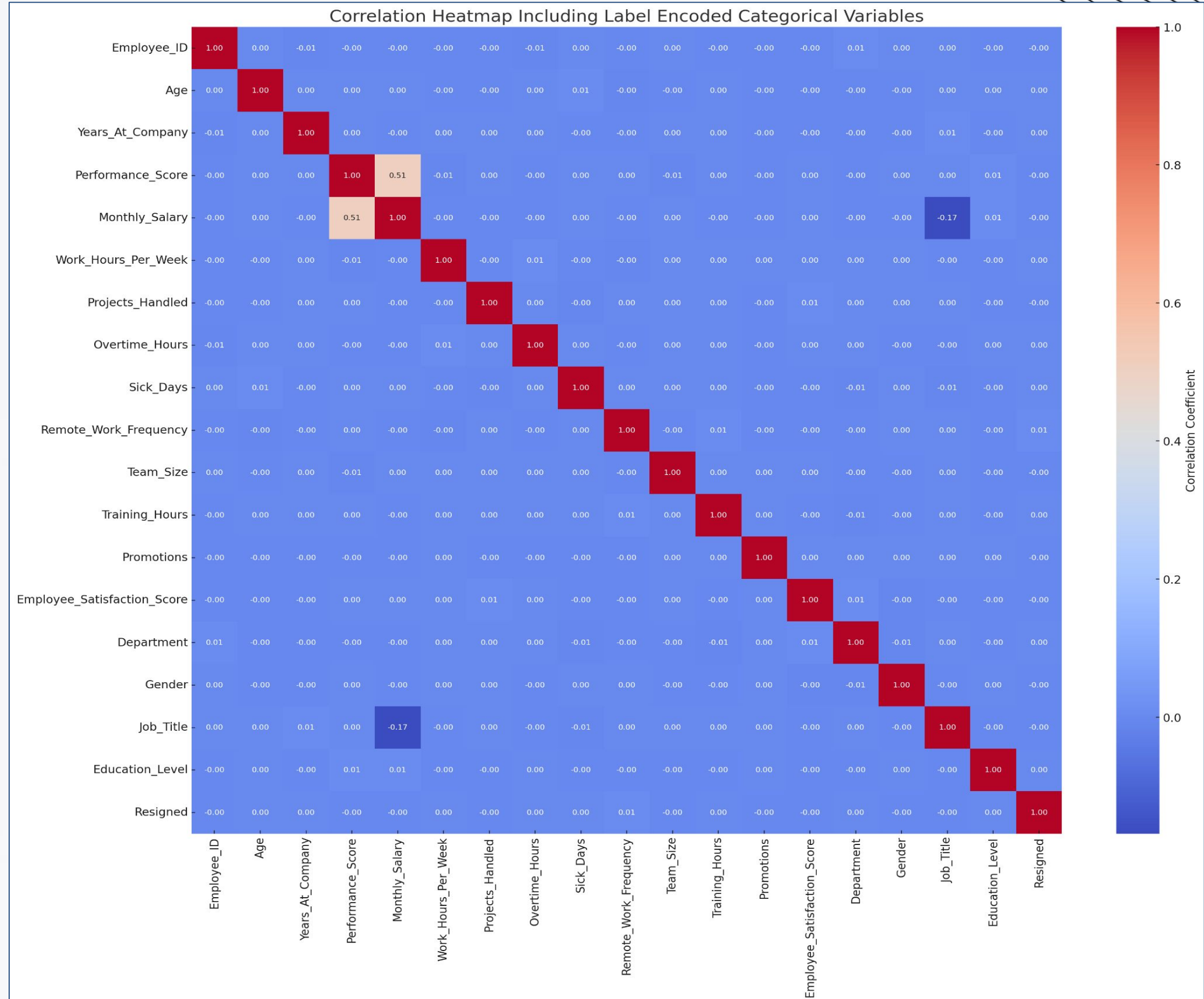
Our Dataset

Employee	Department	Gender	Age	Job_Title	Hire_Date	Years_At_Company	Education_Level	Performance_Score	Monthly_Salary	Work_Hours_Per_Week	Projects_Handled	Overtime_Hours	Sick_Days	Remote_Work_Frequency	Team_Size	Training_Hours	Promotions	Employee_Satisfaction_Score	Resigned
1	IT	Male	55	Specialist	1/19/2022	2	High School	5	6750	33	32	22	2	0	14	66	0	2.63	FALSE
2	Finance	Male	29	Developer	4/18/2024	0	High School	5	7500	34	34	13	14	100	12	61	2	1.72	FALSE
3	Finance	Male	55	Specialist	10/26/2015	8	High School	3	5850	37	27	6	3	50	10	1	0	3.17	FALSE
4	Customer Support	Female	48	Analyst	10/22/2016	7	Bachelor	2	4800	52	10	28	12	100	10	0	1	1.86	FALSE
5	Engineering	Female	36	Analyst	7/23/2021	3	Bachelor	2	4800	38	11	29	13	100	15	9	1	1.25	FALSE
6	IT	Male	43	Manager	8/14/2016	8	High School	3	7800	46	31	8	0	100	15	95	0	2.77	FALSE
7	IT	Male	37	Technician	8/28/2023	1	Bachelor	5	5250	55	20	29	2	0	16	27	0	4.46	FALSE
8	Engineering	Female	55	Engineer	10/27/2014	9	Bachelor	2	7200	42	46	7	8	100	7	64	0	2.09	FALSE
9	Marketing	Female	55	Technician	6/29/2023	1	High School	2	4200	51	23	21	14	0	1	0	1	1.44	FALSE
10	Engineering	Female	45	Consultant	12/23/2016	7	Bachelor	1	6050	41	33	2	6	75	4	53	2	2.93	FALSE
11	Customer Support	Male	52	Engineer	11/26/2019	4	Bachelor	3	7800	38	1	5	0	25	4	90	1	2.34	FALSE
12	Customer Support	Male	27	Technician	2/19/2015	9	Bachelor	5	5250	39	13	2	13	25	5	88	0	1.96	FALSE
13	HR	Male	51	Technician	7/4/2019	5	Bachelor	4	4900	31	11	16	1	75	7	17	2	2.13	FALSE
14	Engineering	Male	27	Analyst	10/14/2014	9	Bachelor	4	5600	33	30	5	13	25	4	26	1	1.46	FALSE
15	Finance	Male	46	Analyst	3/11/2023	1	Master	1	4400	33	49	13	1	100	13	37	1	2.22	FALSE
16	Customer Support	Male	26	Developer	4/19/2023	1	High School	2	6000	33	41	9	13	50	3	46	1	2.42	FALSE
17	Operations	Male	29	Engineer	9/21/2019	4	Bachelor	2	7200	43	26	29	12	0	6	76	2	3.15	FALSE
18	Sales	Other	28	Developer	11/8/2022	1	High School	3	6500	47	36	6	12	75	12	54	0	1.19	FALSE
19	Customer Support	Other	56	Developer	10/1/2015	8	Bachelor	1	5500	57	9	24	1	75	1	36	0	1.12	FALSE
20	Finance	Male	23	Technician	5/8/2015	9	High School	2	4200	52	0	9	1	0	5	66	0	3.91	FALSE
21	Operations	Female	33	Manager	12/19/2022	1	Bachelor	1	6600	33	28	13	13	50	15	53	2	3.7	FALSE
22	Sales	Male	59	Manager	5/25/2017	7	Master	5	9000	30	22	9	2	25	4	20	0	4.66	FALSE
23	Finance	Male	26	Specialist	7/23/2016	8	High School	4	6300	48	40	9	13	0	13	71	2	3.68	FALSE
24	Sales	Male	58	Consultant	1/21/2018	6	Master	2	6600	55	2	28	8	25	6	78	2	4.26	FALSE
25	HR	Female	38	Analyst	6/23/2018	6	Bachelor	2	4800	50	21	24	11	25	7	49	1	4.43	FALSE
26	Customer Support	Male	38	Specialist	5/3/2022	2	Bachelor	5	6750	59	24	0	13	25	3	9	2	2.12	FALSE
27	Finance	Male	45	Engineer	7/12/2021	3	Bachelor	2	7200	59	24	23	6	0	7	26	1	3.16	FALSE
28	Legal	Male	43	Manager	11/22/2014	9	Master	3	7800	44	42	3	10	75	4	45	2	1.35	TRUE
29	Engineering	Male	53	Specialist	4/17/2022	2	Bachelor	2	5400	31	34	14	6	50	6	42	1	4.52	FALSE
30	Finance	Male	52	Analyst	6/21/2023	1	High School	3	5200	45	25	29	9	25	2	37	2	4.95	FALSE

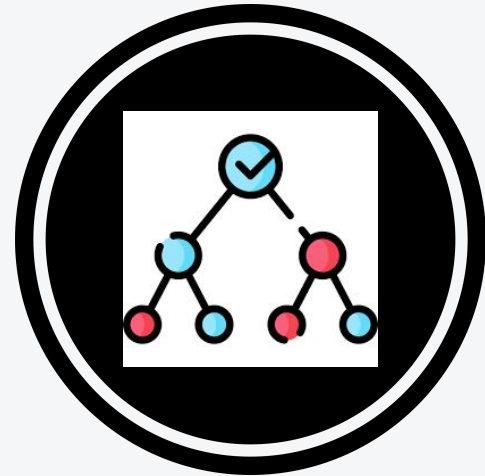
- Kaggle Dataset of an organization's employees from the past 10 years:**
- 100,000 rows of data capturing key aspects of employee performance, productivity, and demographics in a corporate environment
 - Target Variable: “Resigned”

Correlation Heatmap (After Transformation)

- This heatmap shows the correlation between all the features in the dataset (The categorical features have been transformed using Label Encoder)
- Most of the feature pairs show low or near-zero correlations, indicating weak linear relationships
- There is no strong correlation between other features and the "Resigned" status, which might indicate that the decision to resign is not linearly related to the other variables
- Regression models not the best choice as,
 - The target variable, "Resigned," is binary (True/False)
 - Non-linear relationships between "Resigned" and many features



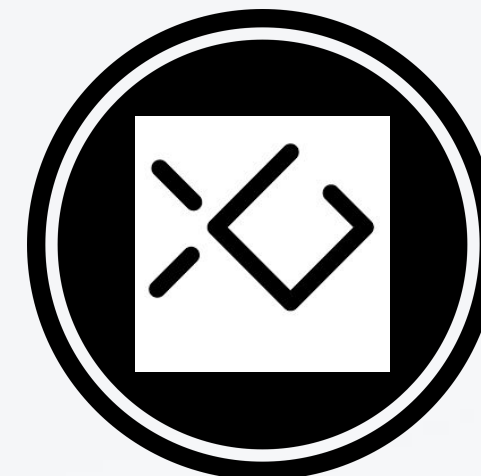
Classifiers Chosen



Decision Tree

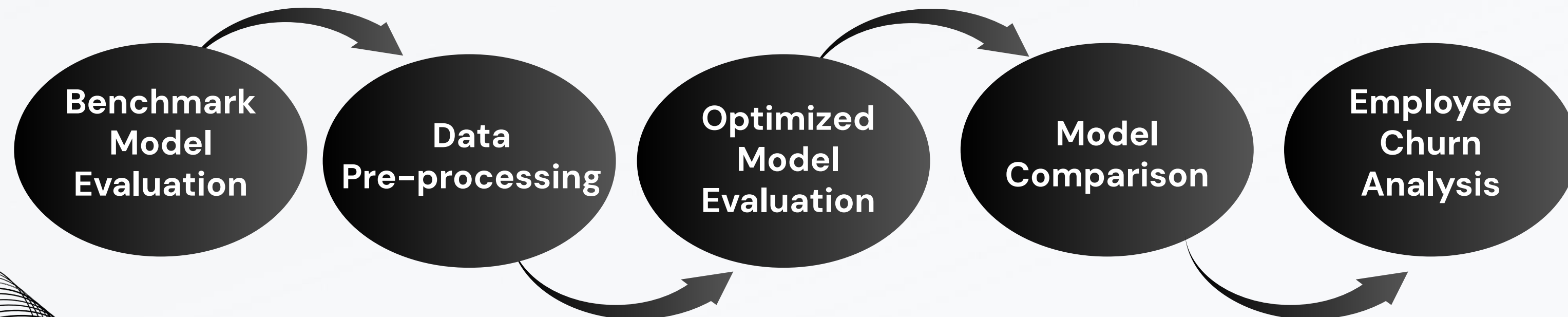


Random Forest



XGBoost

Analysis Approach



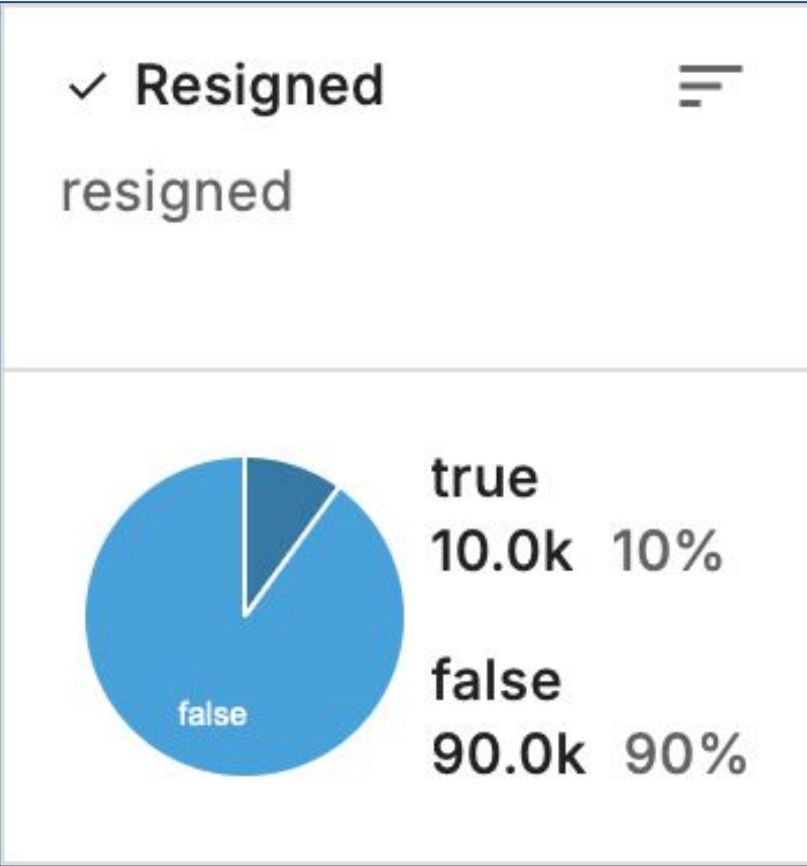
Benchmark Class Distribution

The dataset is highly imbalanced with respect to the target variable "Resigned"

- **Class 0 (Not Resigned):** 89.99% of the instances
- **Class 1 (Resigned):** 10.01% of the instances

Accuracy is not a reliable metric for comparing machine learning models

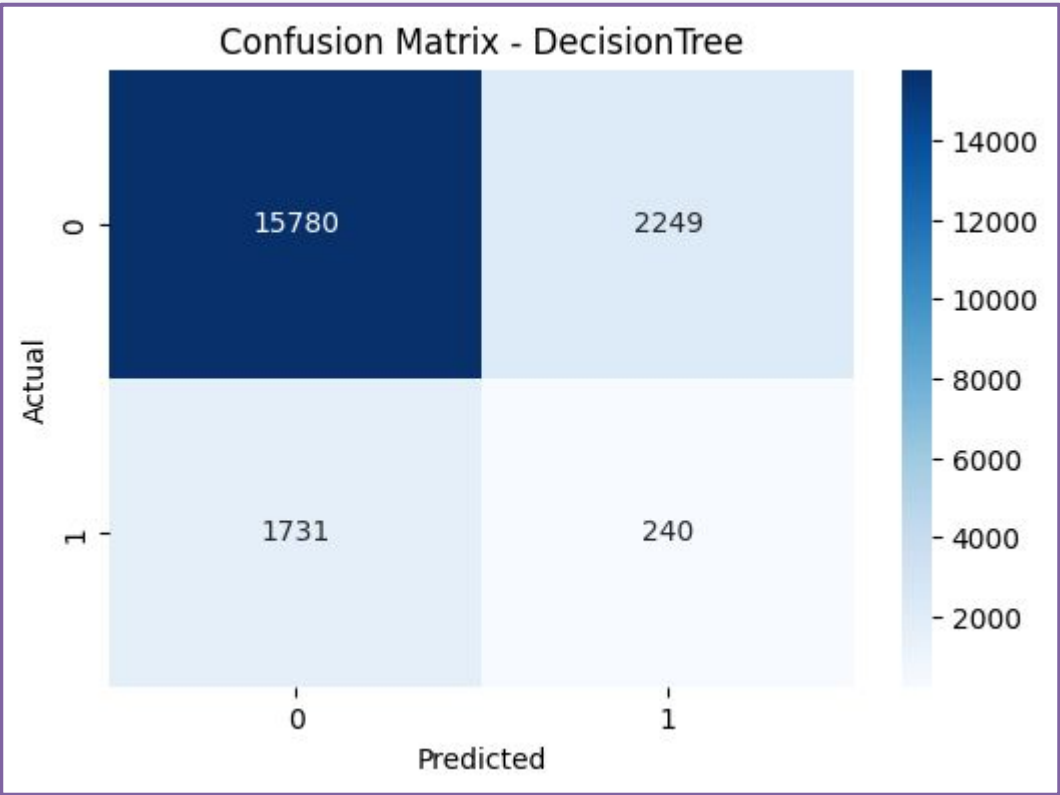
- A model could achieve nearly perfect accuracy by always predicting Class 0 (the majority class), but it would completely fail to identify instances of Class 1 (the minority class)
- Evaluate Stratified Accuracy and Confusion Matrix is more accurate



index	accuracy
DecisionTree	0.801
RandomForest	0.90145
XGBoost	0.90085

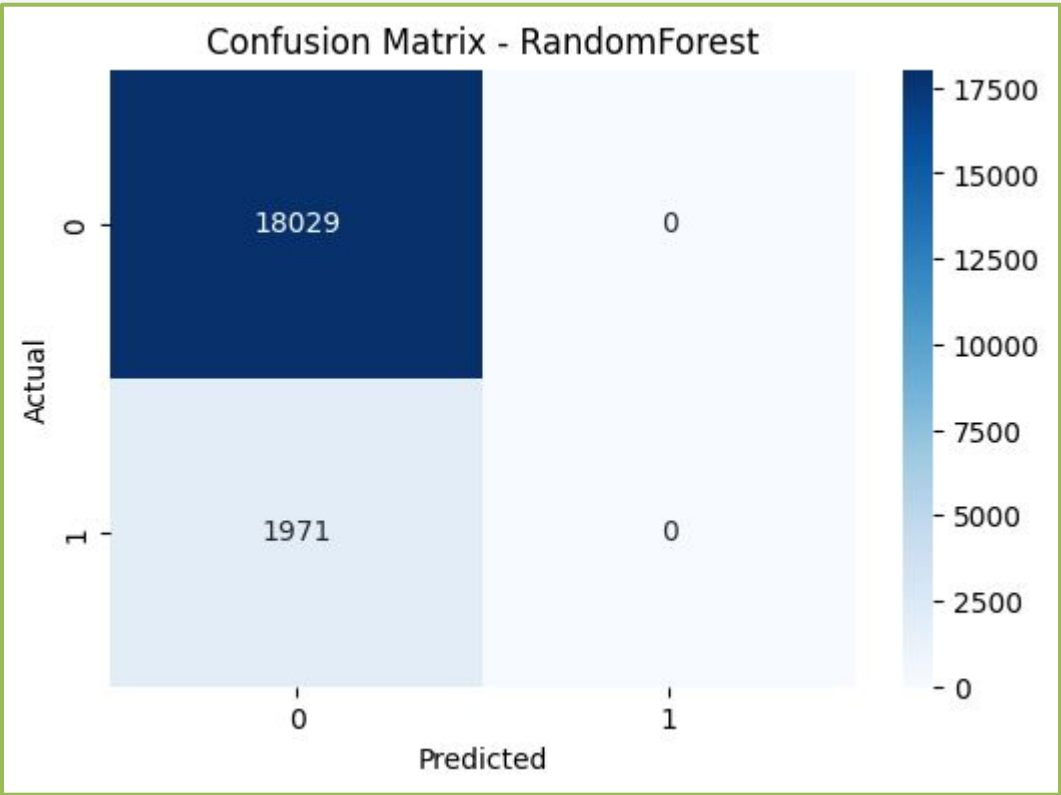
THE ACCURACY IS ALREADY
SO GOOD? NO!

Benchmark Model Evaluation: Conf. Matrix & Stratified Accuracy



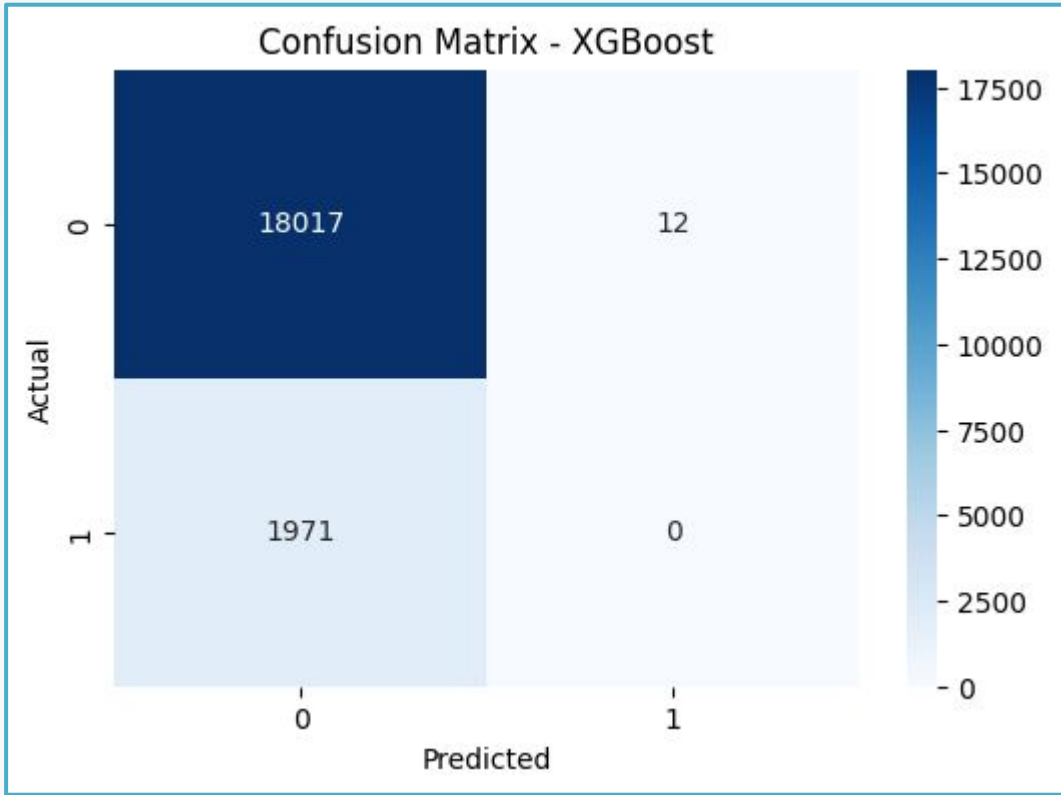
Class 0 Accuracy: 87.53%
(reasonable performance for the majority class).

Class 1 Accuracy: 12.18%
(poor performance for the minority class).



Class 0 Accuracy: 100%
(perfect classification for the majority class).

Class 1 Accuracy: 0%
(fails entirely to predict any minority class instances).



Class 0 Accuracy: 99.93%
(near-perfect classification for the majority class).

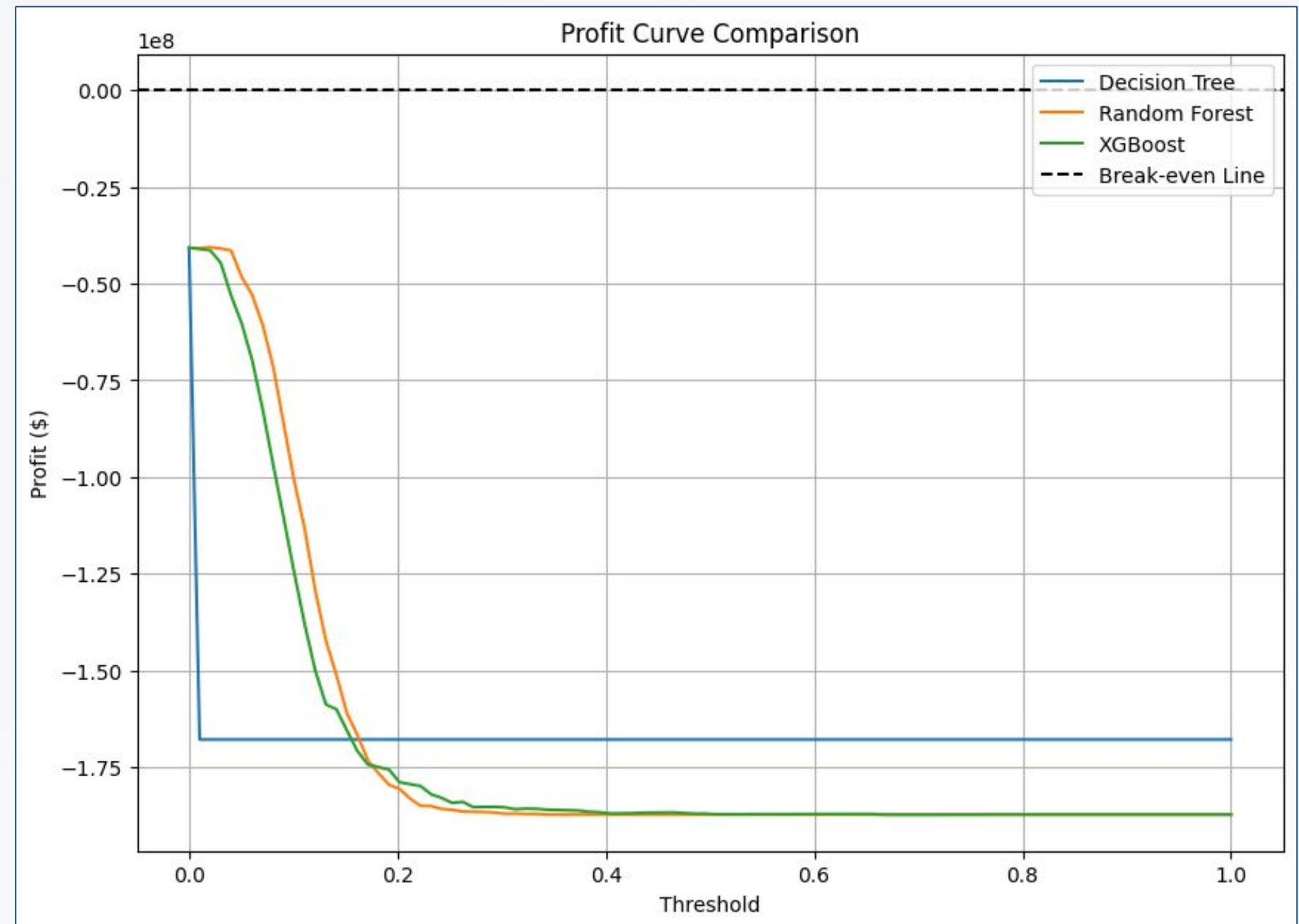
Class 1 Accuracy: 0%
(fails completely for the minority class).

Benchmark Model Evaluation: Cost/Benefit Analysis

- **Estimated Cost of Employee Turnover:**
\$95,000
- **Estimated Cost of Employee Retention:**
\$11,400
- **Benefit of Retaining an At-Risk Employee:**
 $\$95,000 - \$11,400 = \$83,600$

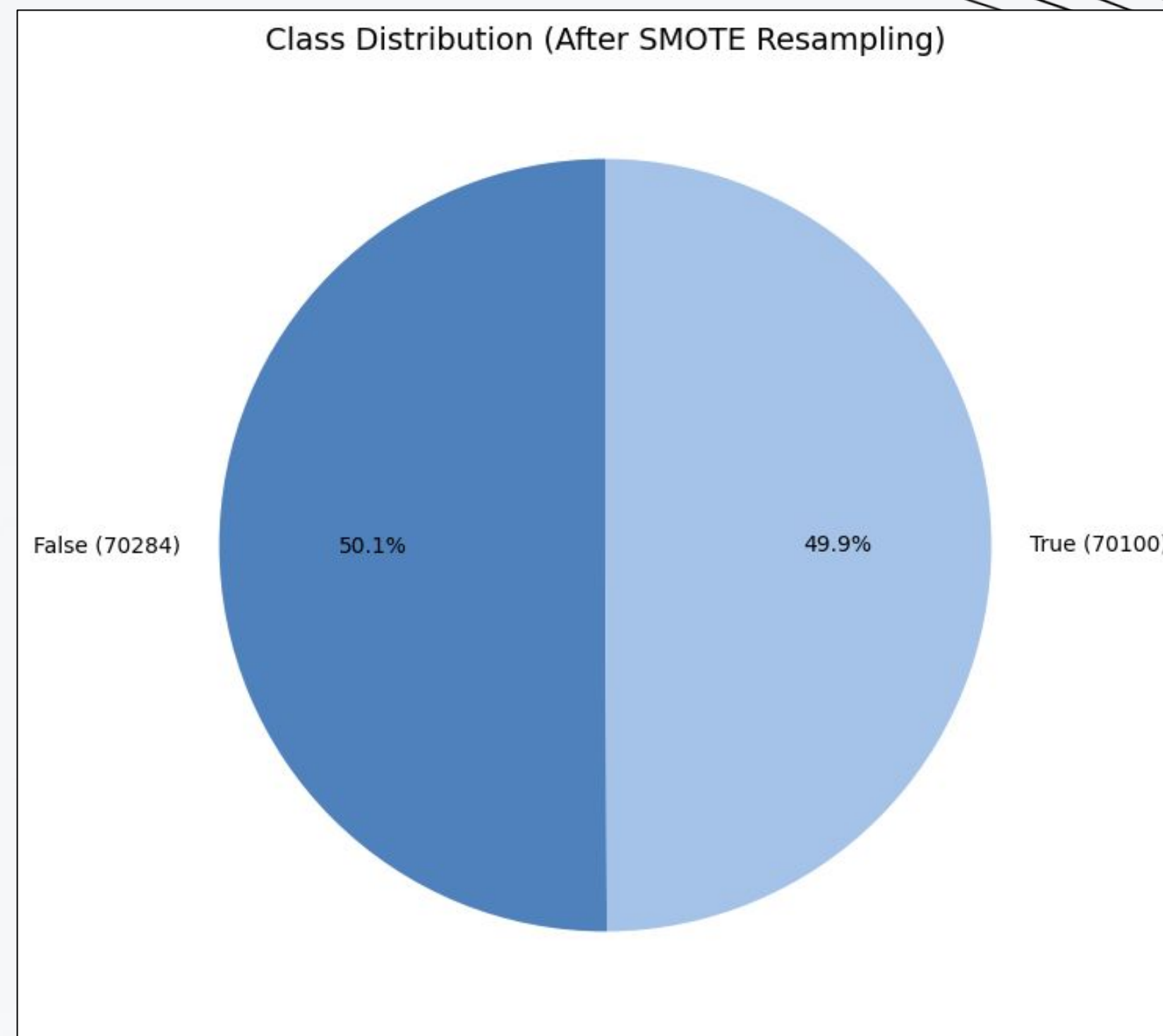
TP = 83600
FP = -11400
FN = -95000
TN = 0

Expected Value (Decision Tree): -167842200
Expected Value (Random Forest): -187245000
Expected Value (XGBoost): -187180400



Pre-processing, Balancing, and Hyper-parameter Tuning

- **Dropped ineffective features:**
Employee_ID, Hire_Date
- **Transformed** the following categorical features using Label Encoder:
Department, Gender, Job_Title, Education_Level
- **Resampled/Balanced** the data using *SMOTE*
- Performed **Hyper-parameter Tuning & Grid Search** to optimize the models

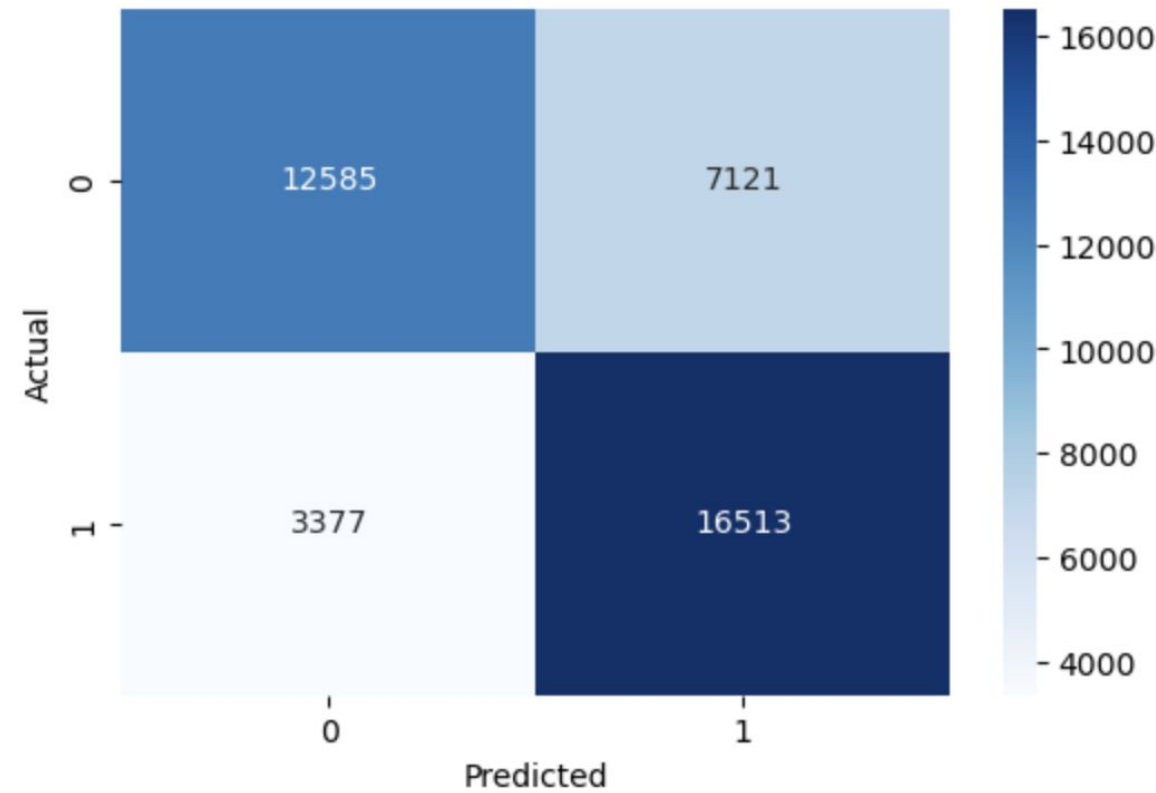


Best Hyperparameters:

```
Decision Tree: {'criterion': 'gini', 'max_depth': 20, 'min_samples_leaf': 10, 'min_samples_split': 2}  
Random Forest: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}  
XGBoost: {'learning_rate': 0.2, 'max_depth': 10, 'n_estimators': 300, 'subsample': 0.8}
```

Optimized Model Evaluation: Conf. Matrix & Stratified Accuracy

Confusion Matrix - Decision Tree



Model: Decision Tree
Class 0 Accuracy: 0.6386
Class 1 Accuracy: 0.8302

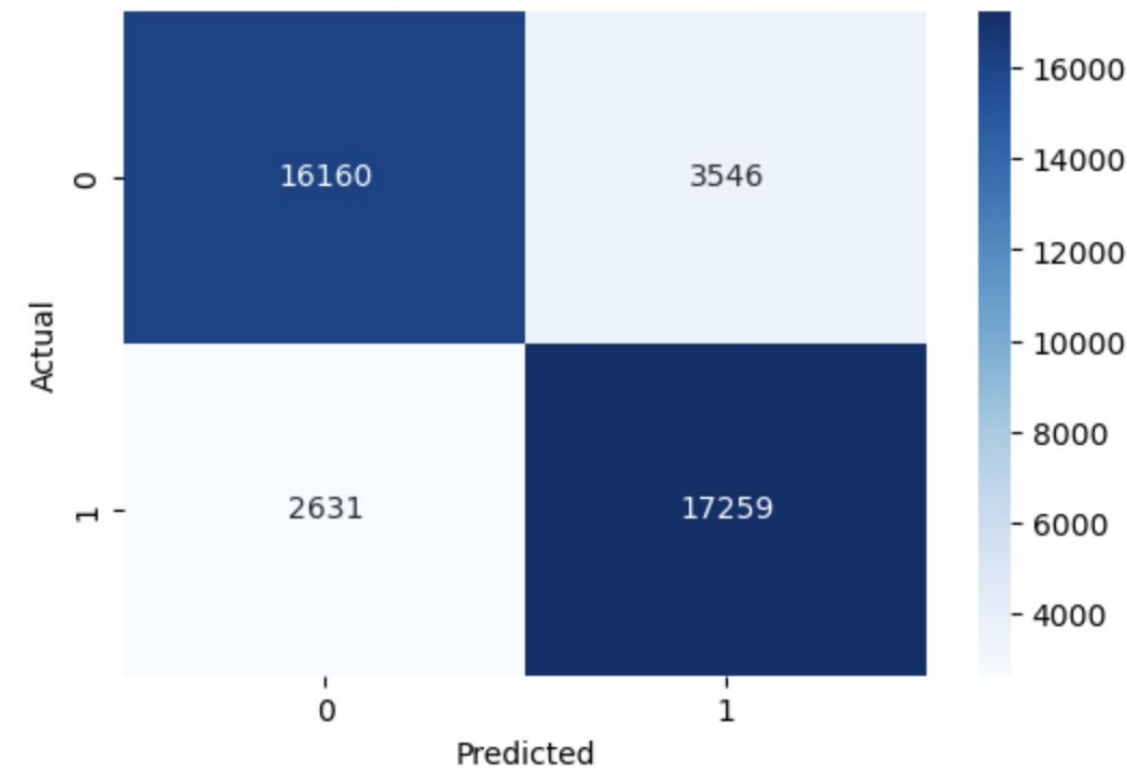
Decision Tree Classifier

- Achieved **Class 1 Accuracy**: ~83.02% (significant improvement)
- **Class 0 Accuracy**: ~63.86% (trade-off due to improved minority detection)

Random Forest Classifier

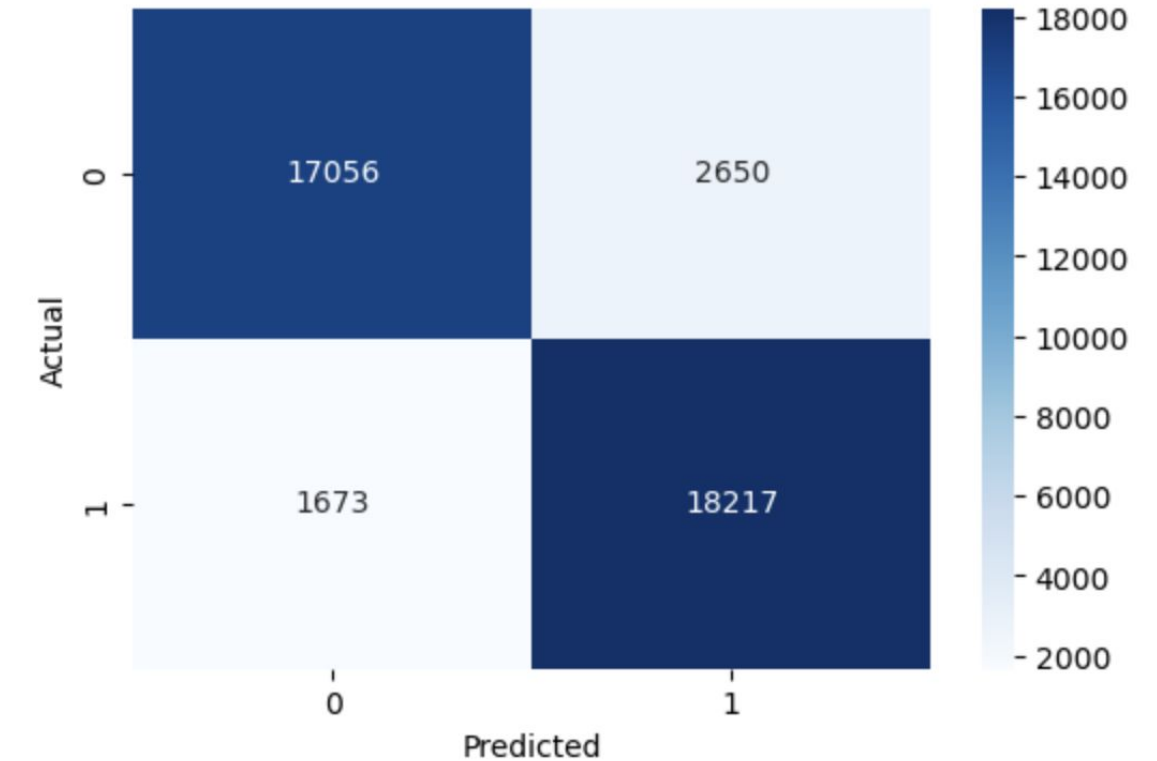
- Improved **Class 1 Accuracy**: ~86.77%
- Balanced performance with **Class 0 Accuracy**: ~82.01%

Confusion Matrix - Random Forest



Model: Random Forest
Class 0 Accuracy: 0.8201
Class 1 Accuracy: 0.8677

Confusion Matrix - XGBoost



Model: XGBoost
Class 0 Accuracy: 0.8655
Class 1 Accuracy: 0.9159

XGBoost Classifier

- Best performance among the models
- **Class 1 Accuracy**: ~91.59%
 - **Class 0 Accuracy**: ~86.55%

Optimized Model Evaluation: Cost/Benefit Analysis

- **Estimated Cost of Employee Turnover:**
\$95,000
- **Estimated Cost of Employee Retention:**
\$11,400
- **Benefit of Retaining an At-Risk Employee:**
 $\$95,000 - \$11,400 = \$83,600$

TP = 83600

FP = -11400

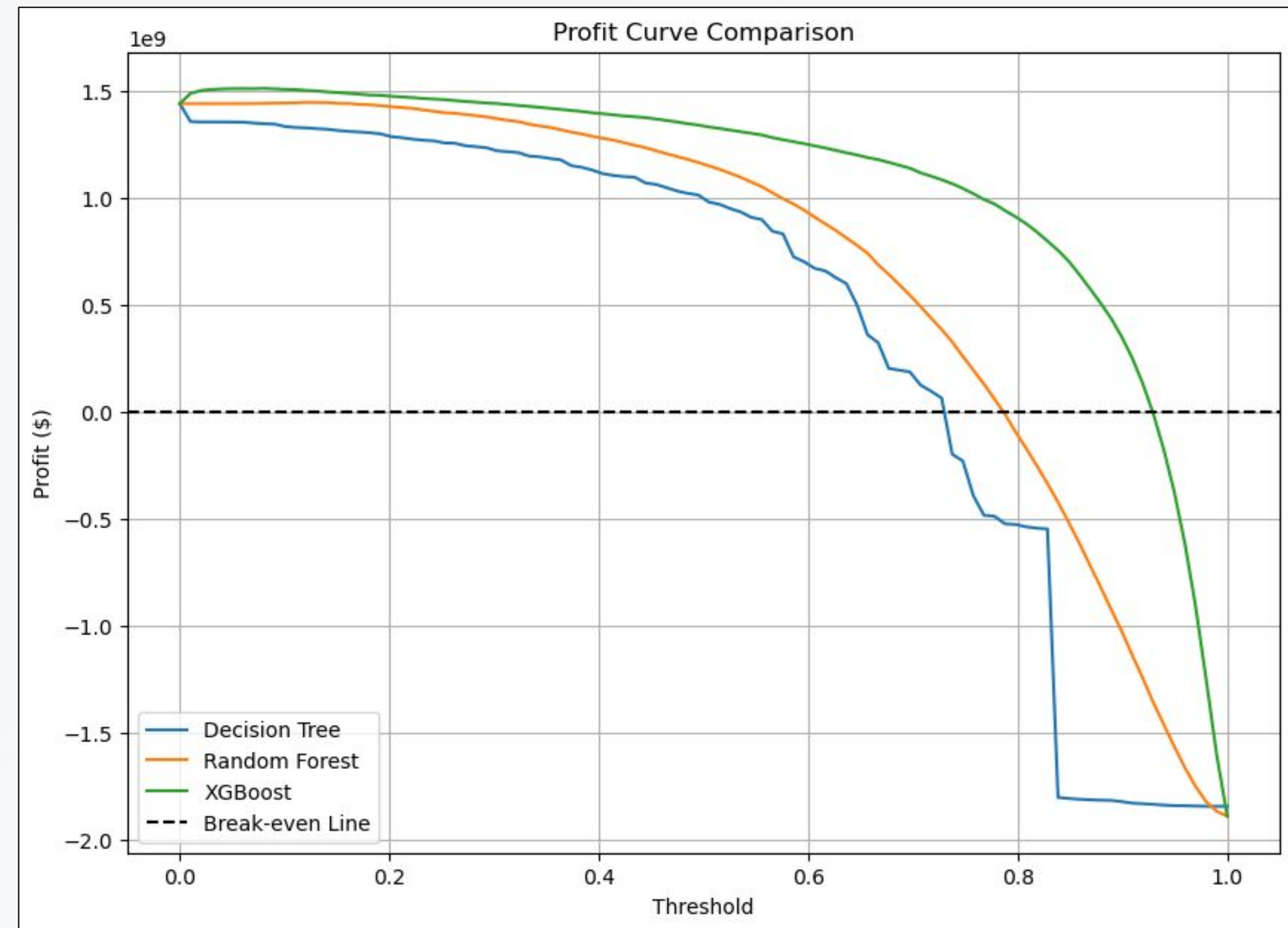
FN = -95000

TN = 0

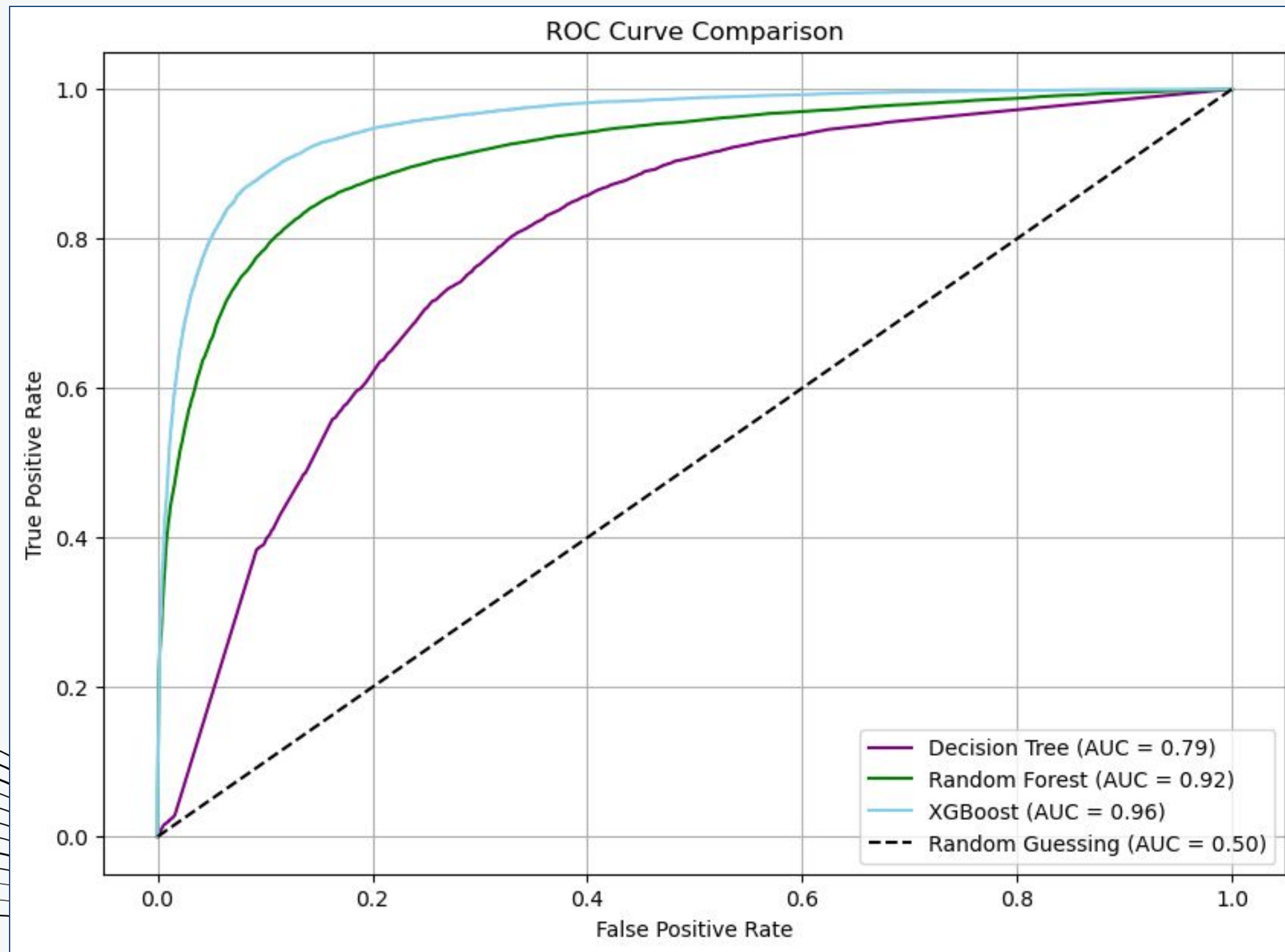
Expected Value (Decision Tree): 978,492,400

Expected Value (Random Forest): 1,152,483,000

Expected Value (XGBoost): 1,333,796,200



Model Comparison



Model Performance:

- XGBoost achieves the highest AUC of 0.96, indicating superior classification performance
- Random Forest follows with an AUC of 0.92, showing strong performance but slightly less accurate than XGBoost
- Decision Tree has the lowest AUC of 0.79, suggesting lower discriminatory power compared to the other models

Key Insights:

- XGBoost proves to be the best model to predict employee churn in this case

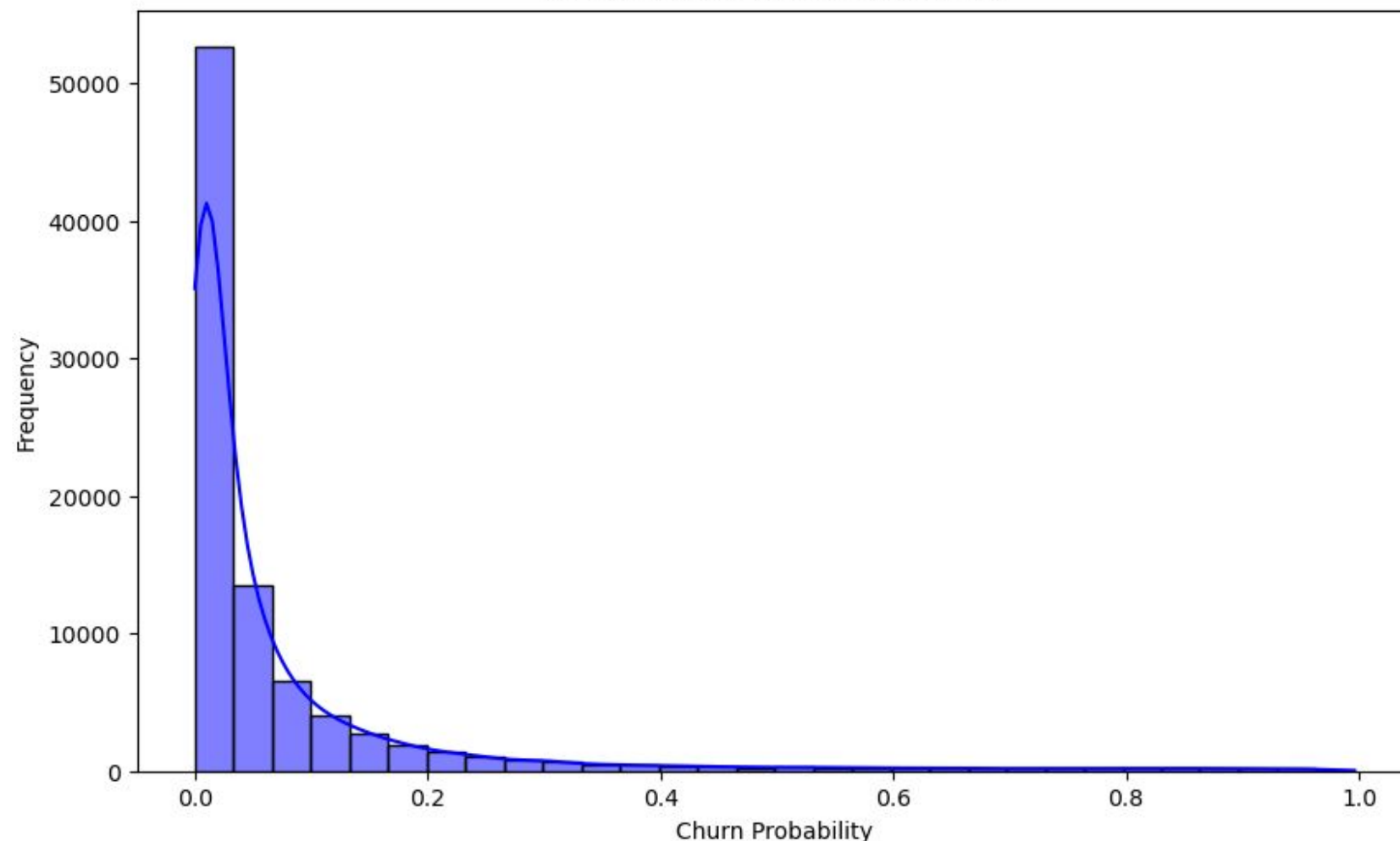
Employee Churn Analysis

- **Gender** is the most influential factor in predicting churn
- **Promotions** and **education level** significantly impact **churn likelihood**
- **Department** and **years** at the company are **strong predictors**

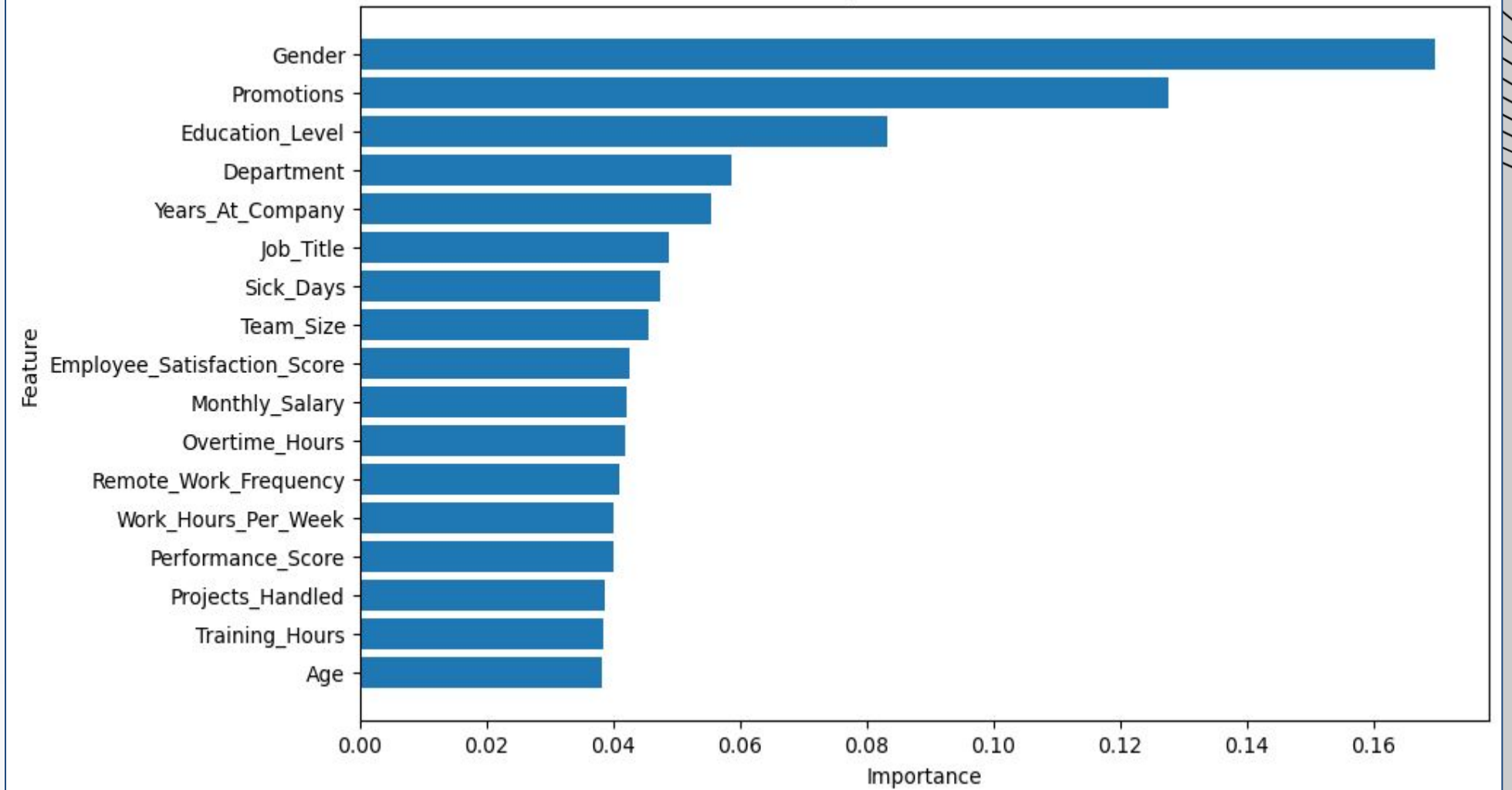
Key Insights

- Understanding and addressing the most influential factors can guide targeted retention strategies to retain top talent

Distribution of Churn Probabilities



Feature Importance from XGBoost

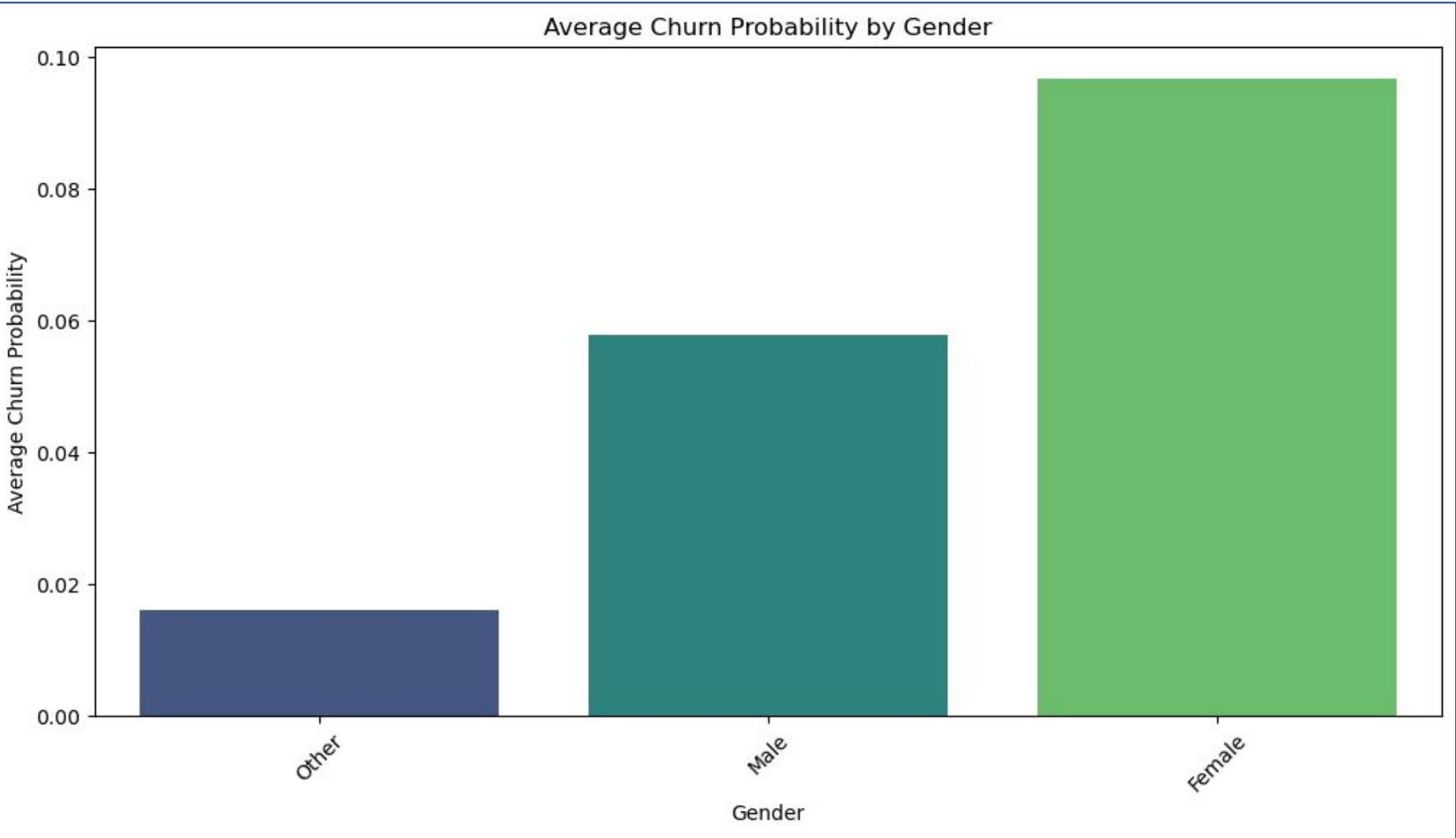


- The **majority of employees** have a **low churn probability**, mostly between **0 and 0.2**

Key Insights

- These at-risk employees can be identified and prioritized for intervention, improving the efficiency of retention efforts

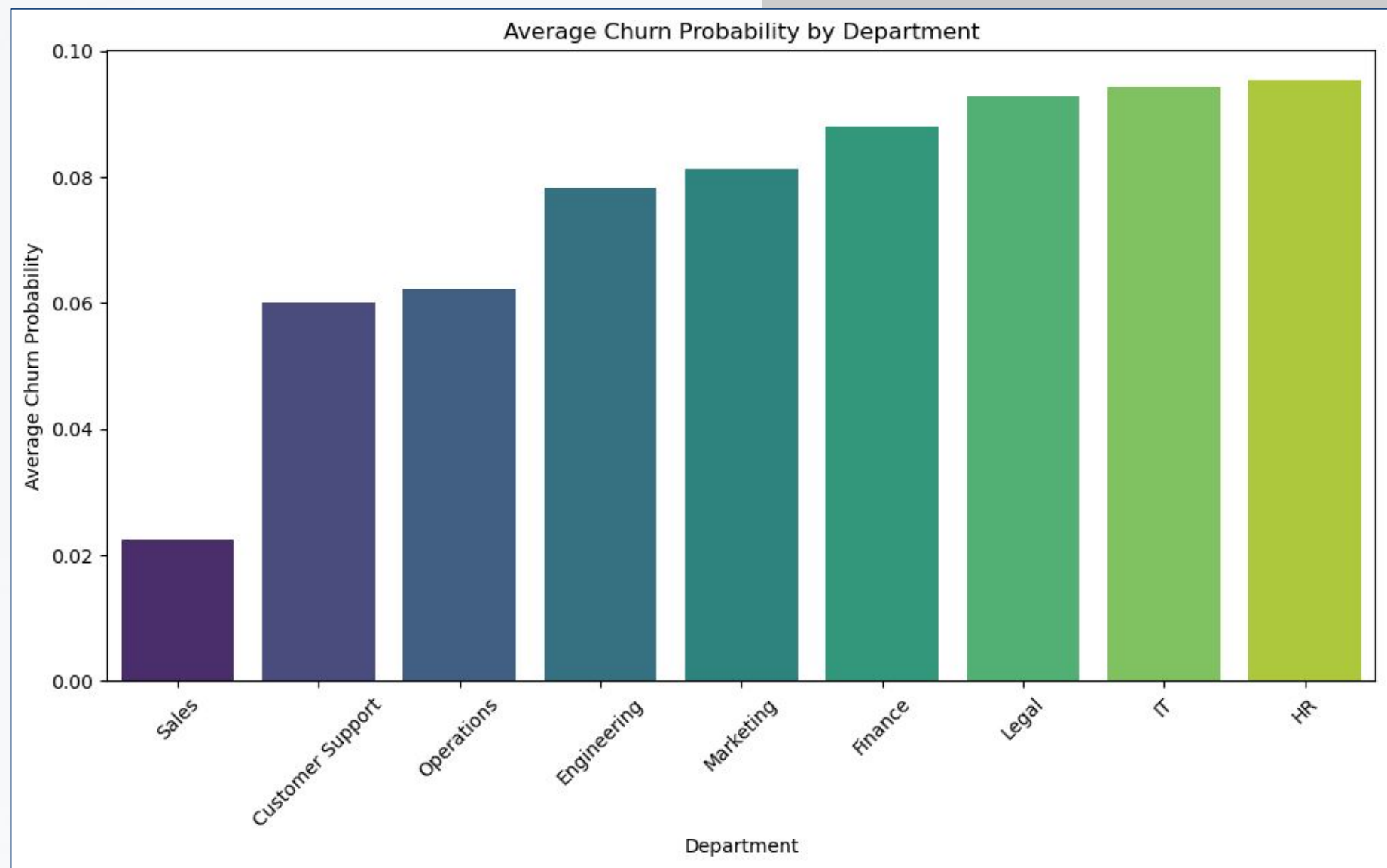
Churn Probability by Department and Gender



Department	Job_Title	Education_Level	Promotions	Remote_Work_Frequency	Years_At_Company
IT	Engineer	Bachelor	0	75	3

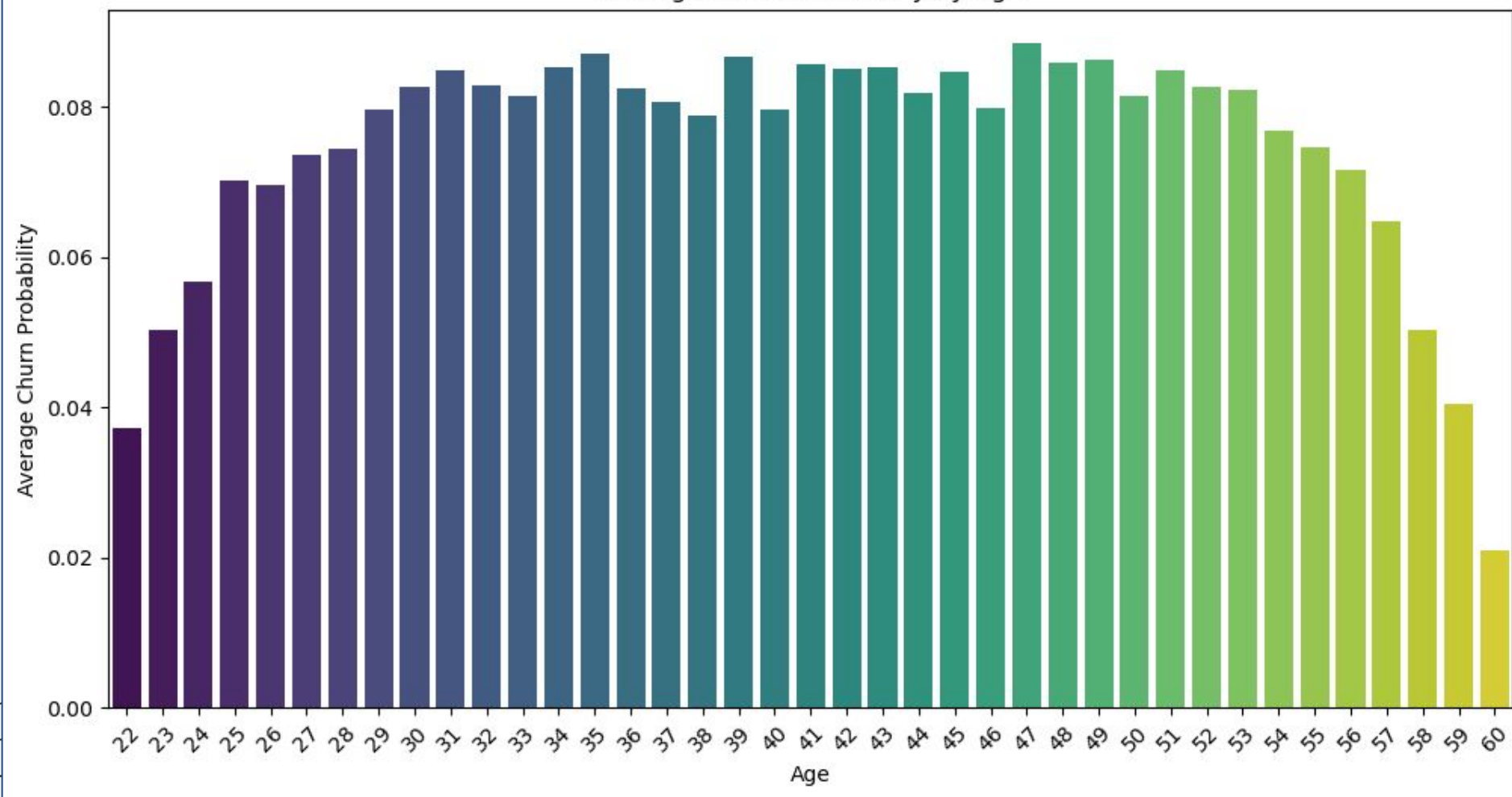
- **Female** employees exhibit the **highest churn probability**
- **Male** employees show a **moderate** churn probability
- **'Other'** category has the **lowest** risk of churning

- **HR, IT and Legal** departments show the **highest** churn probability while **Sales** has the **lowest**
- **Finance** also has a high churn probability but not as high as HR, IT and Legal



Churn Probability by Age and Years at Company

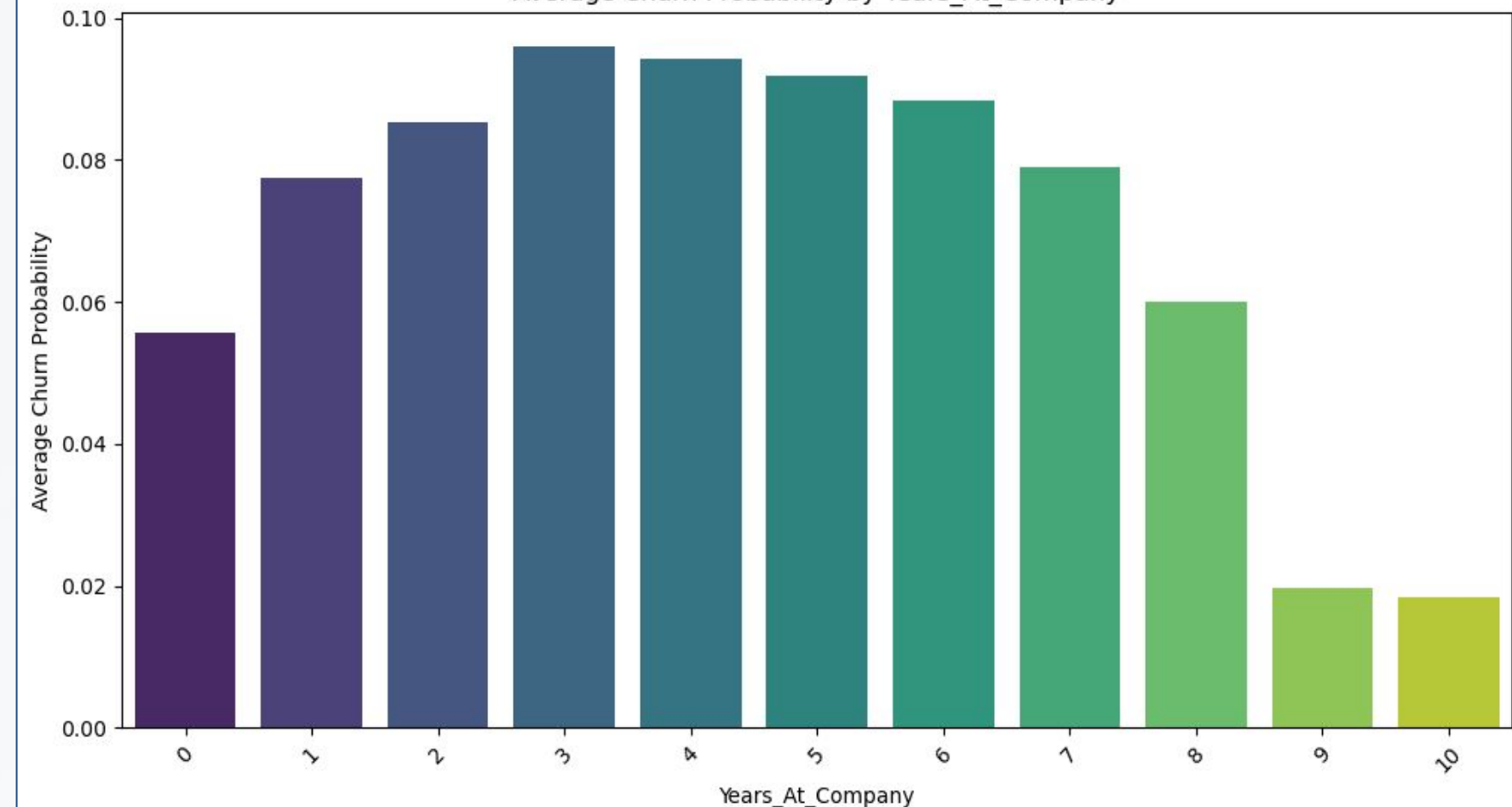
Average Churn Probability by Age



- Churn probability **trends upward** for employees **until their mid-40s**
- Younger (**below 30**) are **less likely to churn** indicating better retention or loyalty
- Churn **stabilizes** and **declines** after **age 55**

- Churn probability is **highest** for employees with **3-6 years** of tenure and are **most likely to churn** around their **3rd year**
- Churn rates decline steadily after 7 years of tenure
- Employees with 7+ years of tenure show significantly lower churn probabilities

Average Churn Probability by Years_At_Company



Strategies and Recommendations



Department-Specific Interventions

Action:

- Reduce workload for HR, IT and Legal – redistribute projects within their departments and automate repetitive tasks when available
- Address issues causing dissatisfaction, ex: unclear job expectations or lack of recognition

Retention Tools:

- Implement collaboration between functions for skill-building



Enhance Career Development Opportunities

Promotions:

- Establish criteria for promotions and paths for career progression
- Provide high-risk employees with leadership opportunities or rotational assignments
- Regular reviews to ensure high performers are recognized and rewarded

Training and Upskilling:

- Offer certifications and advanced training



Diversity/Inclusion & Work-Life Balance

Gender Specific Challenges:

- Conduct gender-specific engagement surveys to understand concerns better
- Create ERGs (Employee Resource Groups)

Flexible Work Arrangements:

- Review and optimize remote work policies to align with employee needs
 - Most low-risk employees prefer on-site

THANK YOU!

