CIS 508 - Final Project

Predicting Employee Resignations to Enhance Retention and Productivity

Machine Learning for Workforce Optimization

Team 401 - Divyansh Shrivastava, Lucas Smith, Kavya Murugan, Sami Fahim, Sravani Bolla

TABLE OF CONTENTS

Business Understanding

Business Problem:

High employee resignation rates are impacting organizational productivity and increasing costs due to frequent hiring and training.

Goal:

Predict employee resignations and identify key factors influencing them.

Propose actionable strategies to enhance retention and productivity.

Value:

Help HR teams make data-driven decisions to reduce employee turnover.

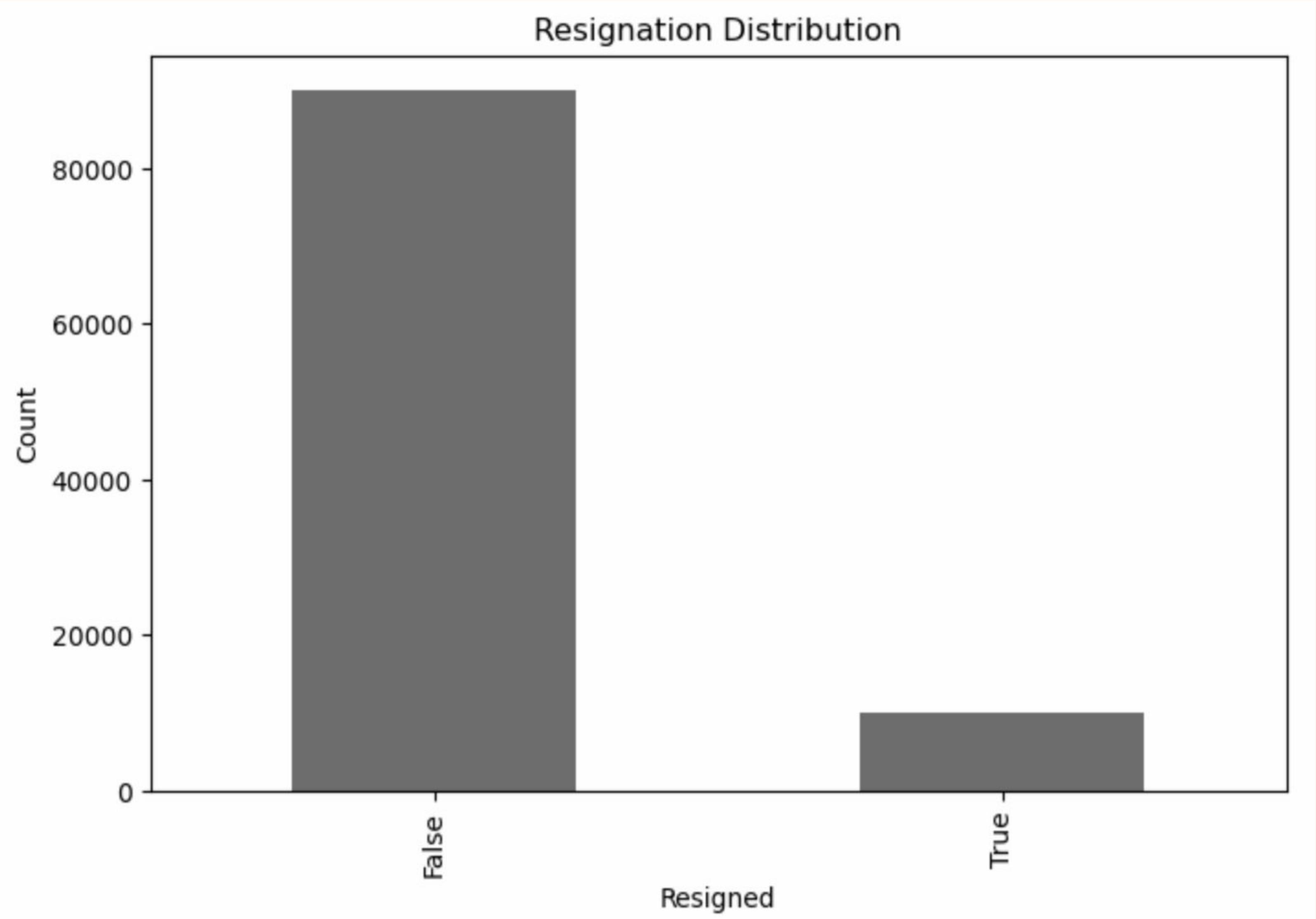
Data Understanding

Dataset Overview

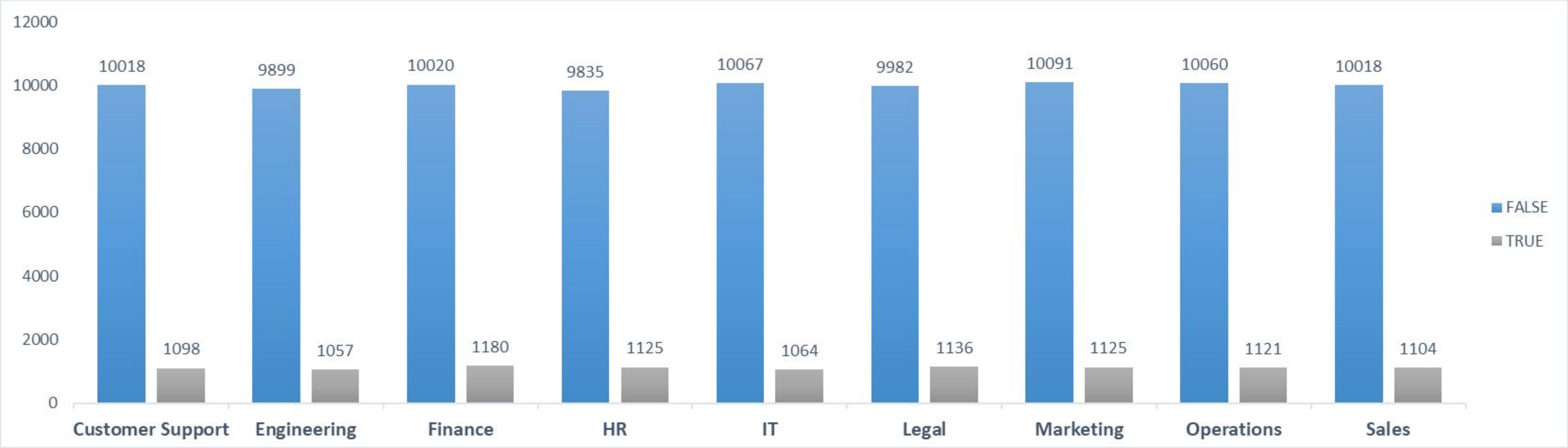
Source: [Kaggle - Employee Performance and Productivity Data.]

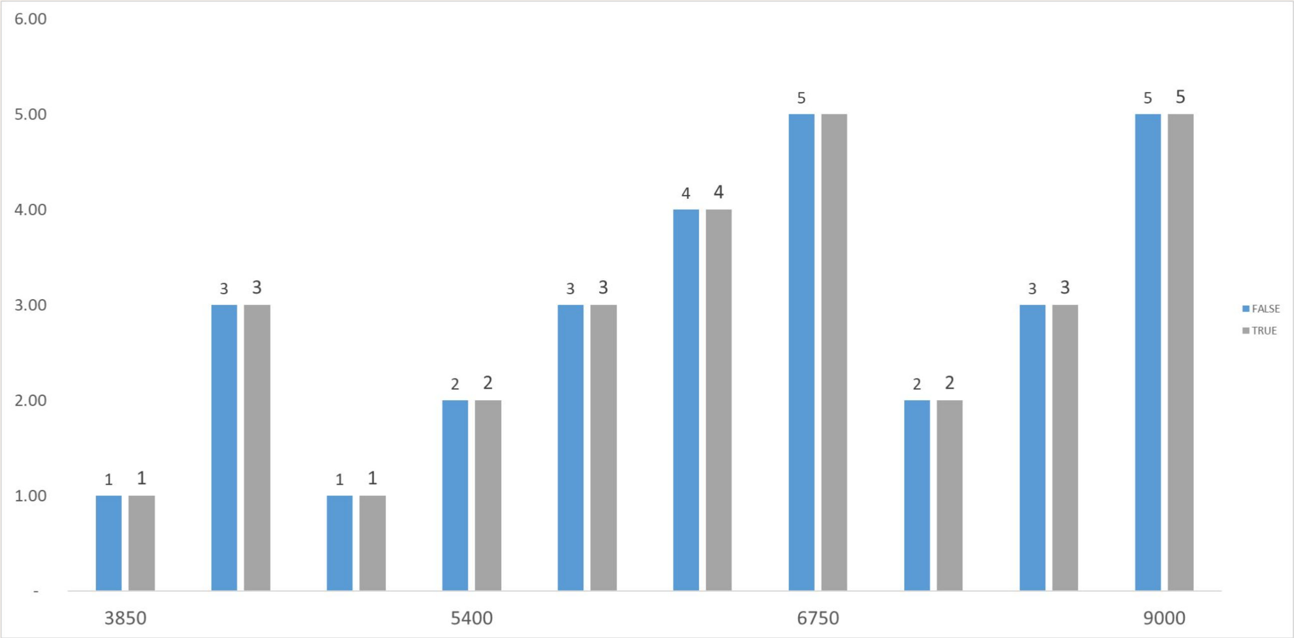
Target Variable: Resignation Status (True/False).

```
Data after handling datetime columns:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 22 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   Employee_ID                  100000 non-null  int64
 1   Department                   100000 non-null  object
 2   Gender                       100000 non-null  object
 3   Age                          100000 non-null  int64
 4   Job_Title                    100000 non-null  object
 5   Years_At_Company             100000 non-null  int64
 6   Education_Level              100000 non-null  object
 7   Performance_Score            100000 non-null  int64
 8   Monthly_Salary               100000 non-null  int64
 9   Work_Hours_Per_Week          100000 non-null  int64
 10  Projects_Handled             100000 non-null  int64
 11  Overtime_Hours               100000 non-null  int64
 12  Sick_Days                    100000 non-null  int64
 13  Remote_Work_Frequency        100000 non-null  int64
 14  Team_Size                    100000 non-null  int64
 15  Training_Hours               100000 non-null  int64
 16  Promotions                   100000 non-null  int64
 17  Employee_Satisfaction_Score  100000 non-null  float64
 18  Resigned                     100000 non-null  bool
 19  Hire_Date_year               100000 non-null  int32
 20  Hire_Date_month              100000 non-null  int32
 21  Hire_Date_day                100000 non-null  int32
dtypes: bool(1), float64(1), int32(3), int64(13), object(4)
memory usage: 15.0+ MB
None
```



Data Understanding





Profit Matrix

True Positive

Correctly predicting resignation and intervening.

- Cost: $10,000
- Benefit: $50,000
- **Net Savings: $40,000**

False Positive

Predicting an employee will resign when they were not going to.

- Cost: $10,000
- **Net Loss: $10,000**

False Negative

Failing to predict an employee is going to resign.

- Cost: $50,000
- **Net Loss: $50,000**

True Negative

Correctly predicting an employee will stay

- **Net Savings of 0**

Example Scenario Assumptions

Employee Count: 1,000 employees
Annual Turnover Rate: 15% (150 resignations per year)
Average Turnover Cost per Employee: $50,000 (includes recruiting, training, lost productivity, etc.)
Retention Intervention Cost: $10,000 per employee (e.g., salary adjustment, training, incentives).

Model Accuracy Assumptions:

True Positive Rate (Recall): 80%
True Negative Rate (Specificity): 90%
False Positive Rate: 10%
False Negative Rate: 20%

Overall Cost-Benefit Calculation

True Positives (120 employees): Savings = 120 × $40,000 = $4,800,000.	False Positives (100 employees): Cost = 100 × $10,000 = $1,000,000.
True Negatives (850 employees): Savings = 850 × $0 = $0.	False Negatives (30 employees): Cost = 30 × $50,000 = $1,500,000.

Net Savings & Insights

Total Savings = $4,800,000
Total Costs = $1,000,000 + $1,500,000 = $2,500,000

Net Savings = $2,300,000

Insights

True Positives drive the majority of the savings by avoiding turnover costs

False Positives are less expensive than False Negatives, as retention efforts are cheaper than turnover costs.

Improving the recall is key to maximizing savings

Data Preparation

STEPS TAKEN

1. Encoding Categorical Variables

2. Handling Missing Values

3. Scaling Numerical Features

4. Outlier Detection and Removal

Split our dataset of 100,000 entries into 70% training and 30% testing data

Modeling Approach

We experimented the following Classification models:

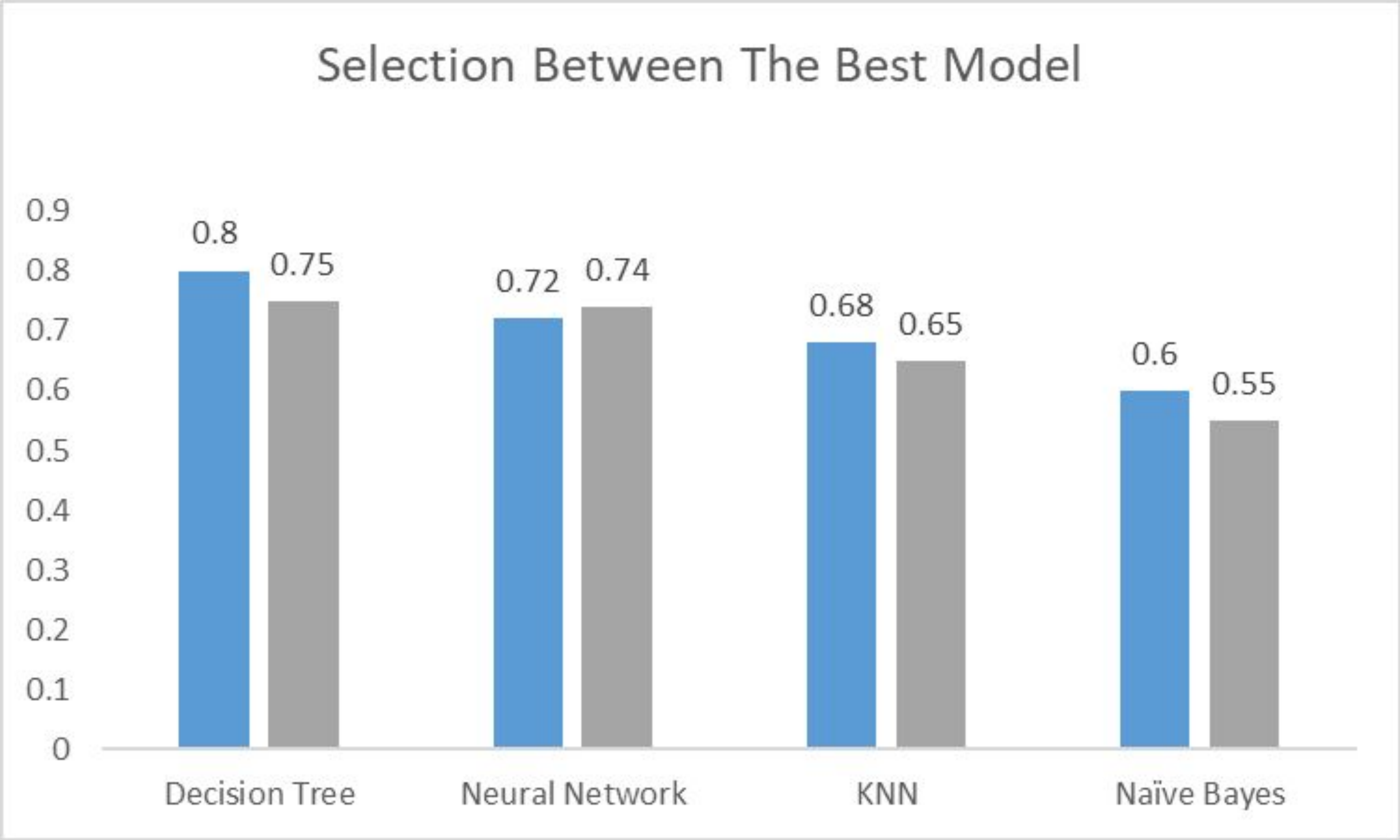
- Random Forest
- KNN
- Decision Tree
- XGBoost
- Logistic Regression
- Naive Bayes
- SVM
- Neural Network

We followed these steps:

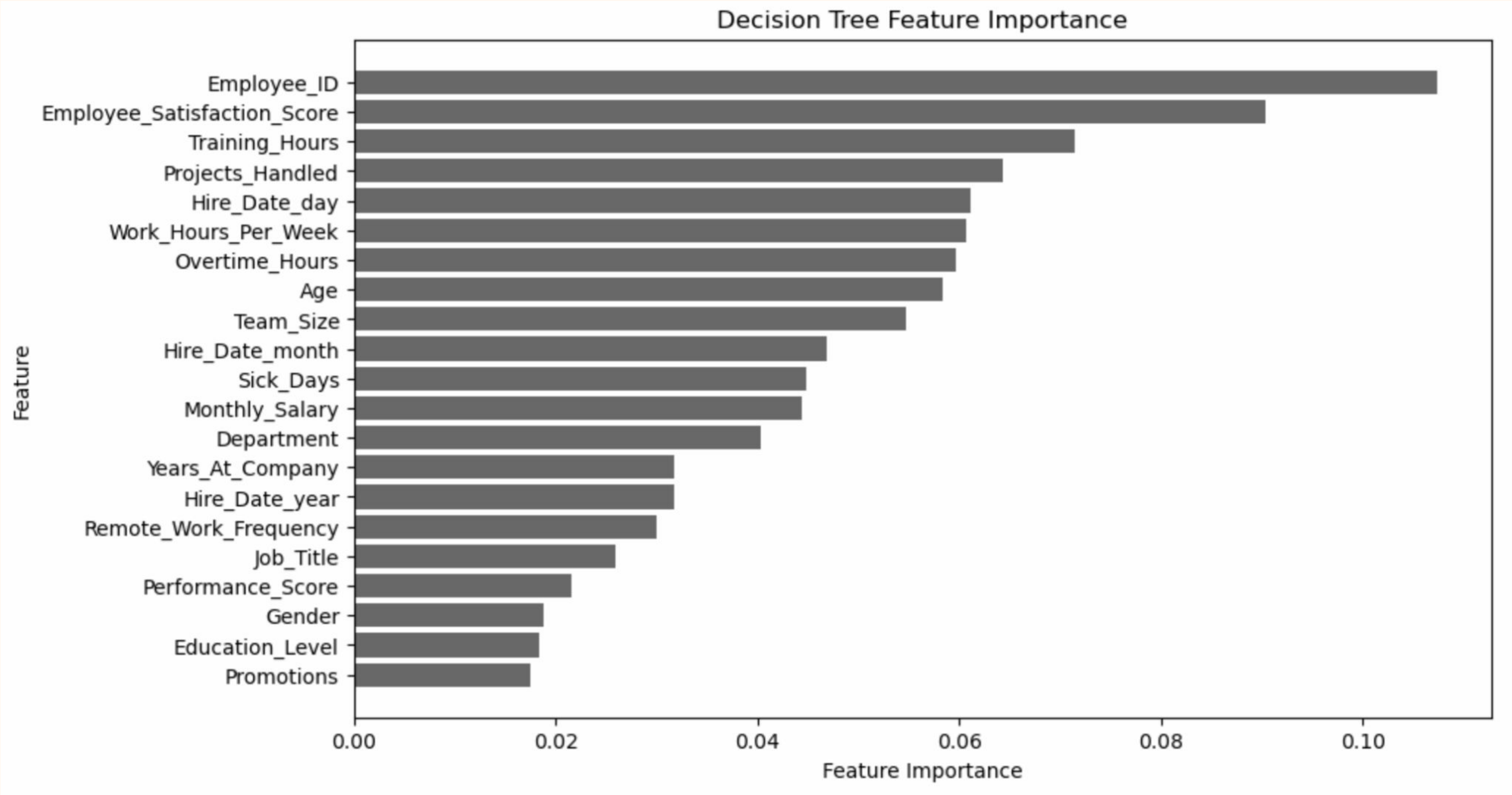
1. **Initial Modeling:** Built and evaluated 8 models.
2. **Best Model Selection:** Chose the top-performing model based on their performance.
3. **SMOTE Analysis:** Applied SMOTE to address class imbalance and refine model performance, again choosing a best model.
4. **Hyperparameter Tuning:** Optimized the top 2-3 models for the best performance.

Model Selection

	Decision Tree	Neural Network	KNN	Naïve Bayes
Recall	0.8	0.72	0.68	0.6
F1	0.75	0.74	0.65	0.55



Evaluation


Decision Tree Feature Importance
Employee_ID
Employee_Satisfaction_Score
Training_Hours
Projects_Handled
Hire_Date_day
Work_Hours_Per_Week
Overtime_Hours
Age
Team_Size
Hire_Date_month
Sick_Days
Monthly_Salary
Department
Years_At_Company
Hire_Date_year
Remote_Work_Frequency
Job_Title
Performance_Score
Gender
Education_Level
Promotions
Feature
0.00
0.02
0.04
0.06
0.08
0.10
Feature Importance

Deployment

- The best Decision Tree model was saved as best_decision_tree_model.pkl using joblib to enable seamless reuse and deployment.

- This ensures quick integration into HR systems for real-time resignation predictions, saving time and computational resources.

New Sample Data			
Employee_ID	1001	1002	1003
Department	Sales	IT	HR
Gender	Male	Female	Female
Age	30	25	40
Job_Title	ales Executiv	Software Engineer	HR Manager
Hire_Date	2020-01-15	2019-06-10	2018-09-20
Years_At_Company	4	5	6
Education_Level	Bachelor	Master	Bachelor
Performance_Score	3	4	2
Monthly_Salary	5000	7000	6500
Work_Hours_Per_Week	40	45	35
Projects_Handled	3	5	2
Overtime_Hours	5	10	3
Sick_Days	2	1	4
Remote_Work_Frequency	50	80	20
Team_Size	5	8	4
Training_Hours	20	15	25
Promotions	1	2	0
Employee_Satisfaction_Score	4	5	3

Predictions for Employees	
1001	0
1002	1
1003	1

0 - not resigned
1 - resigned

Thank You!!

Any Questions?