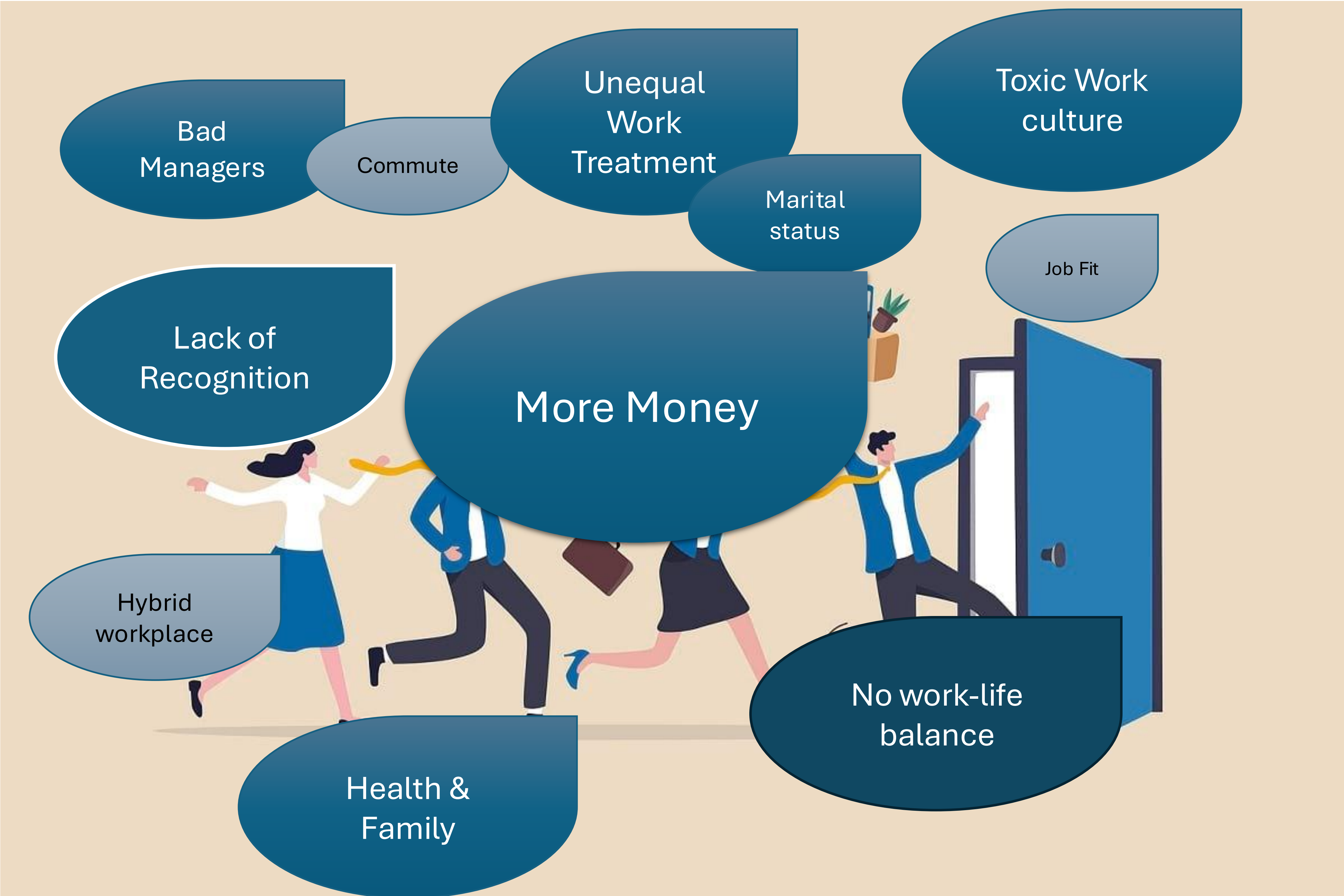
Employee Attrition Prediction

By Group - **M4**

- Akanksha Manohar
- Meenakshi R Nair
- Neha Tiwari
- Niketha Venkatesh
- Shriya Karthikeyan


Bad Managers
Commute
Unequal Work Treatment
Toxic Work culture
Marital status
Job Fit
Lack of Recognition
More Money
Hybrid workplace
No work-life balance
Health & Family

Business problem?

Employee Attrition

especially unexpectedly.

Why is this a problem?

Leads to higher recruitment and training costs.

Loss of experienced talent can affect productivity, morale, and customer service.

Warning Sign of Bigger Issues

Business Goal: Predict which employees are likely to leave so HR can proactively retain talent.

Why is This Dataset Interesting for us?

- Rich Dataset
- Real-world context
- challenges such as:
  - Employee Retention
  - Talent Development
  - Performance Prediction
  - Organizational Planning
- It simulates the type of data companies collect, making it highly relevant for practical business insights.

Data Overview

HR dataset → Uncover key drivers of employee behavior and organizational outcomes.

- Dataset Structure: 1470 observations (rows), 35 features (variables)
- The primary data mining problem is to **predict employee attrition**, i.e.,determine whether an employee is likely to leave the company based on historical data.
- **Target Variable** : Attrition – Yes or No (**Binary Classification**)

➢ Key drivers → 12 features

  - ➢ Demographics – Age, Gender, Marital status
  - ➢ Compensation and Benefits: Monthly Income, Stock Option Level
  - ➢ Work – Life Factors: Work-Life Balance, Total Working Years, Over Time
  - ➢ Satisfaction and Engagement: Job Satisfaction, Environment Satisfaction ,Job Involvement.
  - ➢ Commute and Location: Distance From Home

Sample Data

Age	Attrition	Business _Travel	Daily_Rate	Department	Distance_From_ Home	Education	Education _Field	Employee Count	Employee Number	Environment Satisfaction	Gender	HourlyRate	JobInvolvement	JobLevel	JobRole	JobSatisfaction	MaritalStatus	MonthlyIncome
41	Yes	avel_Rare	1102	Sales	1	2	ife Science	1	1	2	Female	94	3	2	Sales Executive	4	Single	5993
49	No	vel_Freque	279	rch & Develo	8	1	ife Science	1	2	3	Male	61	2	2	Research Scientist	2	Married	5130
37	Yes	avel_Rare	1373	rch & Develo	2	2	Other	1	4	4	Male	92	2	1	aboratory Technicia	3	Single	2090
33	No	vel_Freque	1392	rch & Develo	3	4	ife Science	1	5	4	Female	56	3	1	Research Scientist	3	Married	2909
27	No	avel_Rare	591	rch & Develo	2	1	Medical	1	7	1	Male	40	3	1	aboratory Technicia	2	Married	3468
32	No	vel_Freque	1005	rch & Develo	2	2	ife Science	1	8	4	Male	79	3	1	aboratory Technicia	4	Single	3068
59	No	avel_Rare	1324	rch & Develo	3	3	Medical	1	10	3	Female	81	4	1	aboratory Technicia	1	Married	2670
30	No	avel_Rare	1358	rch & Develo	24	1	ife Science	1	11	4	Male	67	3	1	aboratory Technicia	3	Divorced	2693
38	No	vel_Freque	216	rch & Develo	23	3	ife Science	1	12	4	Male	44	2	3	anufacturing Directo	3	Single	9526
36	No	avel_Rare	1299	rch & Develo	27	3	Medical	1	13	3	Male	94	3	2	althcare Representat	3	Married	5237
35	No	avel_Rare	809	rch & Develo	16	3	Medical	1	14	1	Male	84	4	1	aboratory Technicia	2	Married	2426
29	No	avel_Rare	153	rch & Develo	15	2	ife Science	1	15	4	Female	49	2	2	aboratory Technicia	3	Single	4193
31	No	avel_Rare	670	rch & Develo	26	1	ife Science	1	16	1	Male	31	3	1	Research Scientist	3	Divorced	2911
34	No	avel_Rare	1346	rch & Develo	19	2	Medical	1	18	2	Male	93	3	1	aboratory Technicia	4	Divorced	2661
28	Yes	avel_Rare	103	rch & Develo	24	3	ife Science	1	19	3	Male	50	2	1	aboratory Technicia	3	Single	2028
29	No	avel_Rare	1389	rch & Develo	21	4	ife Science	1	20	2	Female	51	4	3	anufacturing Directo	1	Divorced	9980
32	No	avel_Rare	334	rch & Develo	5	2	ife Science	1	21	1	Male	80	4	1	Research Scientist	2	Divorced	3298
22	No	Non-Trave	1123	rch & Develo	16	2	Medical	1	22	4	Male	96	4	1	aboratory Technicia	4	Divorced	2935
53	No	avel_Rare	1219	Sales	2	4	ife Science	1	23	1	Female	78	2	4	Manager	4	Married	15427
38	No	avel_Rare	371	rch & Develo	2	3	ife Science	1	24	4	Male	45	3	1	Research Scientist	4	Single	3944

Is it supervised or unsupervised?

This is a **supervised classification** problem because:

- We have a labeled data with defined target variable(Attrition: Yes or No).

- Using data mining ,we aim to train the model to learn patterns from historical data and **predict** new cases.

Data Preprocessing

- **Handling Missing Values: using Simple Imputer**

  - Numerical columns → Mean

  - Categorical columns → Mode

- **Detecting Outliers** : 3 **features**

  - Years_Since_Last_Promotion

  - Training_Times_LastYear

  - Performance_Rating

- **Encoding Categorical Variables: using Label encoder**

  – Attrition (Yes = 1 , No = 0)

- **Feature Scaling : using Minmax scaler**

  - MinMax Scaling → Min-Max Scaling is applied to the **numerical features** to normalize them, ensuring that all features are on a similar scale.

- **Balancing dataset : using SMOTE Analysis:**

  - 1237 employees did not leave the organization while → 84% of cases

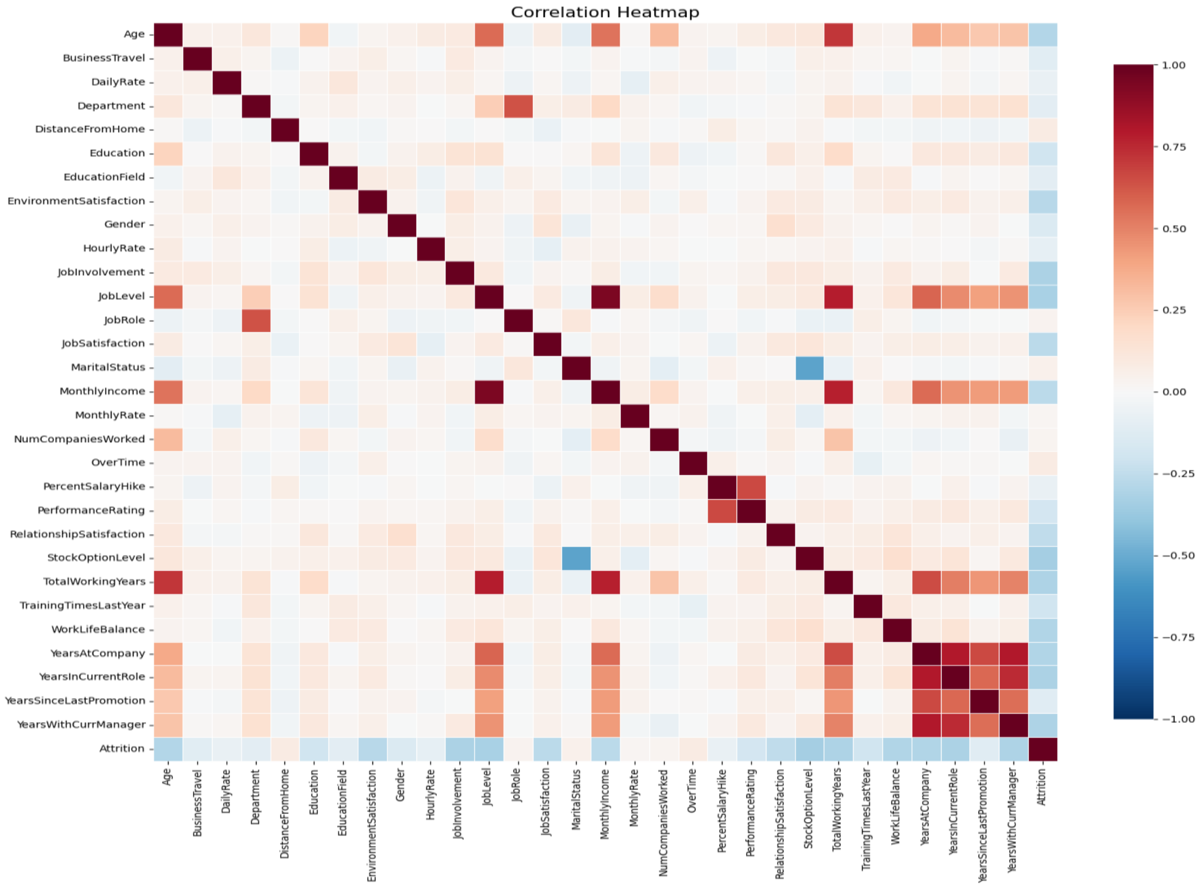
  - 237 did leave the organization →16% of cases

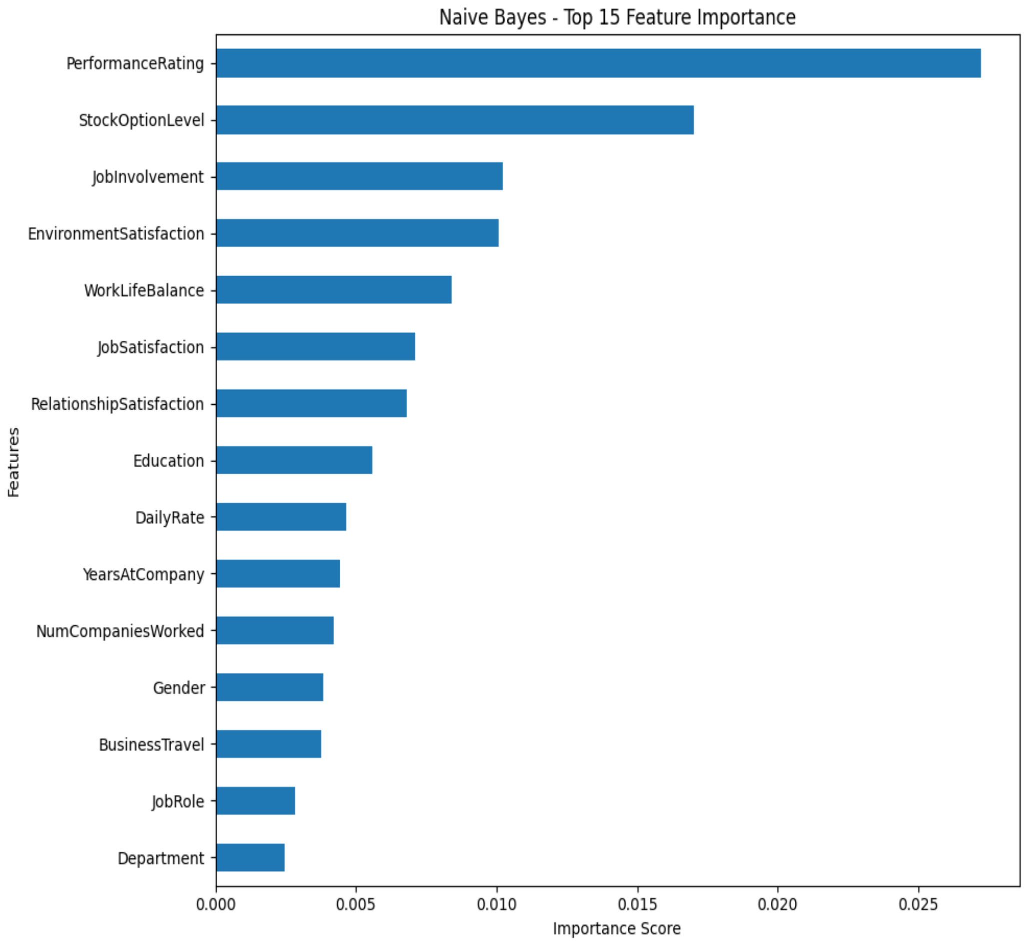
  - **Result** → Dataset is imbalanced

  - **Solution** → used SMOTE to balance class distribution.

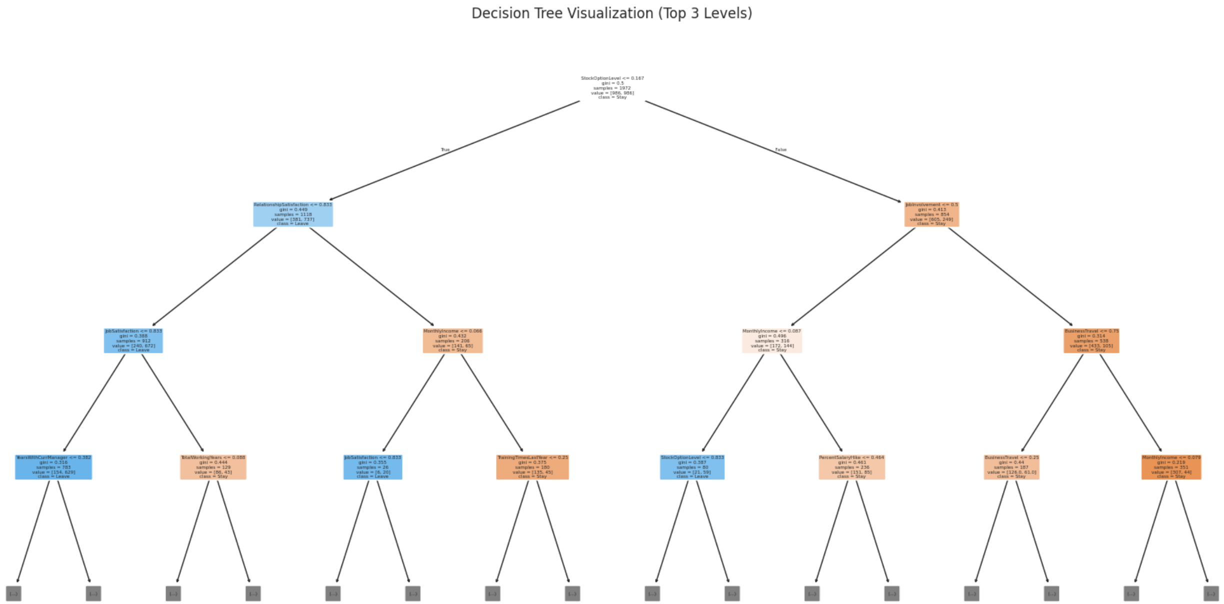
Naïve Bayes

- Top features: Performance rating, stock option level, job involvement, environment satisfaction, work life balance
- Accuracy =0.7341, AUC=0.8555



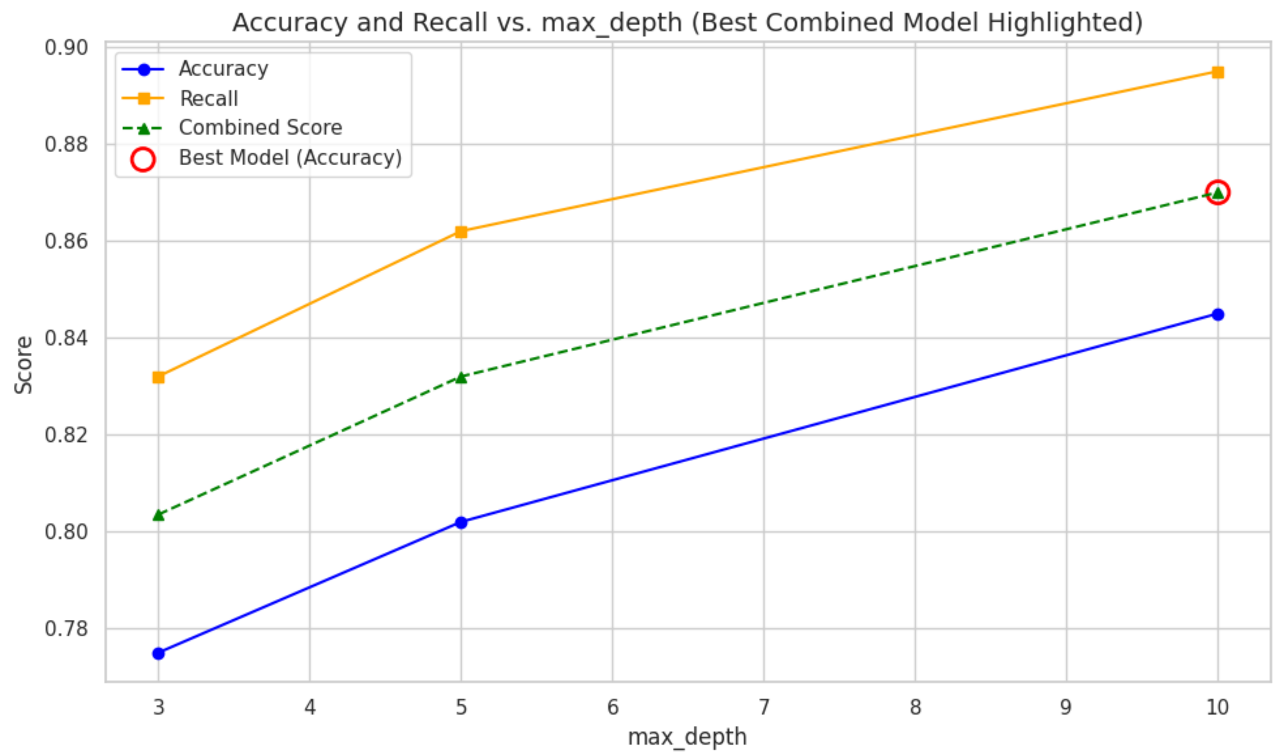


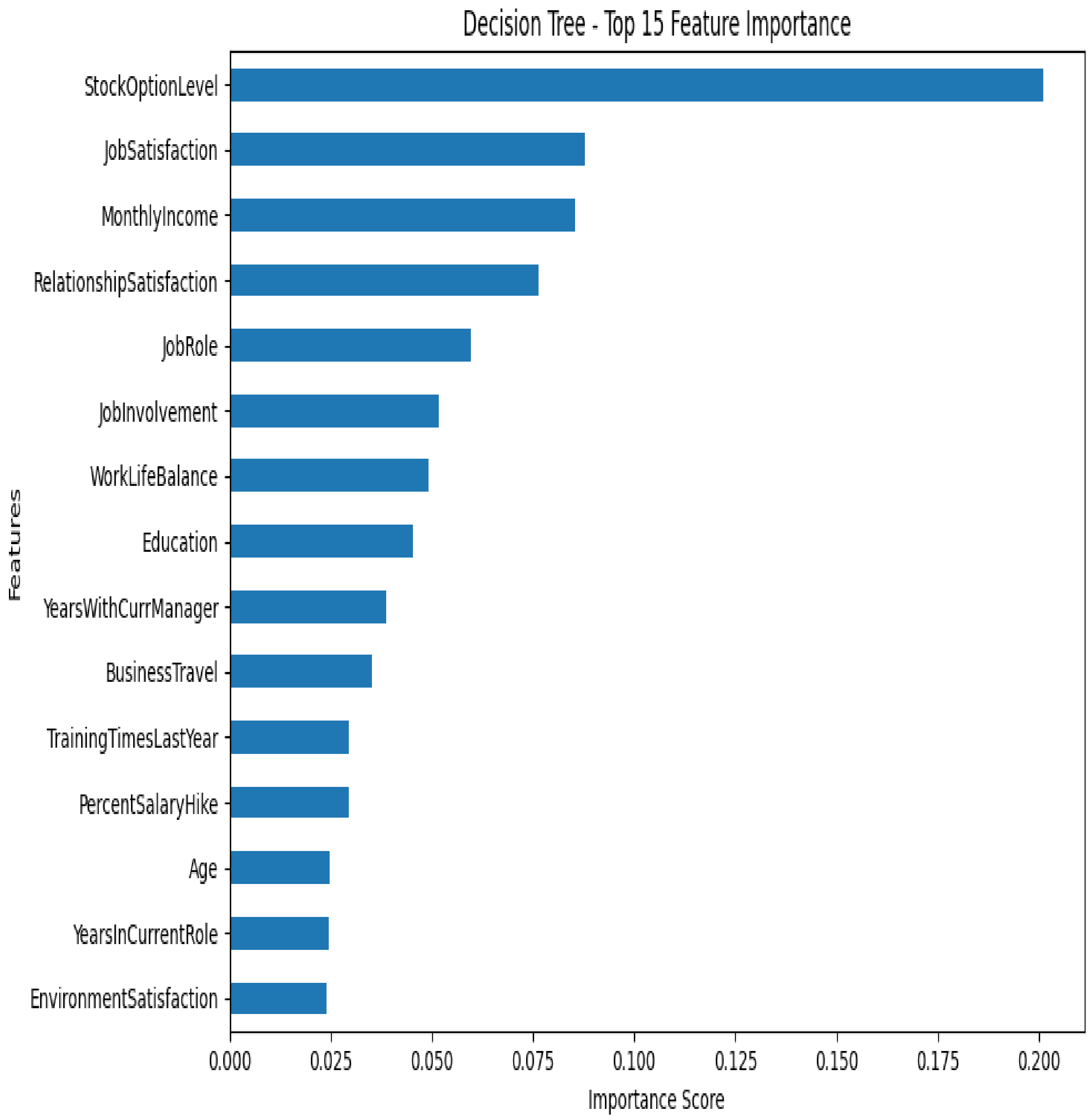
Decision Tree



- Tree depths we tried = 3, 5, 10
- Top features – Stock Option Level, Job Satisfaction, Monthly Income, Relationship satisfaction and Job role
- Found the optimal tree depth to be at 10 with

  AUC =0.81 and Recall = 0.86


Accuracy and Recall vs. max_depth (Best Combined Model Highlighted)
Accuracy
Recall
Combined Score
Best Model (Accuracy)
Score
max_depth
0.90
0.88
0.86
0.84
0.82
0.80
0.78
3
4
5
6
7
8
9
10


Decision Tree - Top 15 Feature Importance
StockOptionLevel
JobSatisfaction
MonthlyIncome
RelationshipSatisfaction
JobRole
JobInvolvement
WorkLifeBalance
Education
YearsWithCurrManager
BusinessTravel
TrainingTimesLastYear
PercentSalaryHike
Age
YearsInCurrentRole
EnvironmentSatisfaction
Features
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importance Score

Logistic Regression:

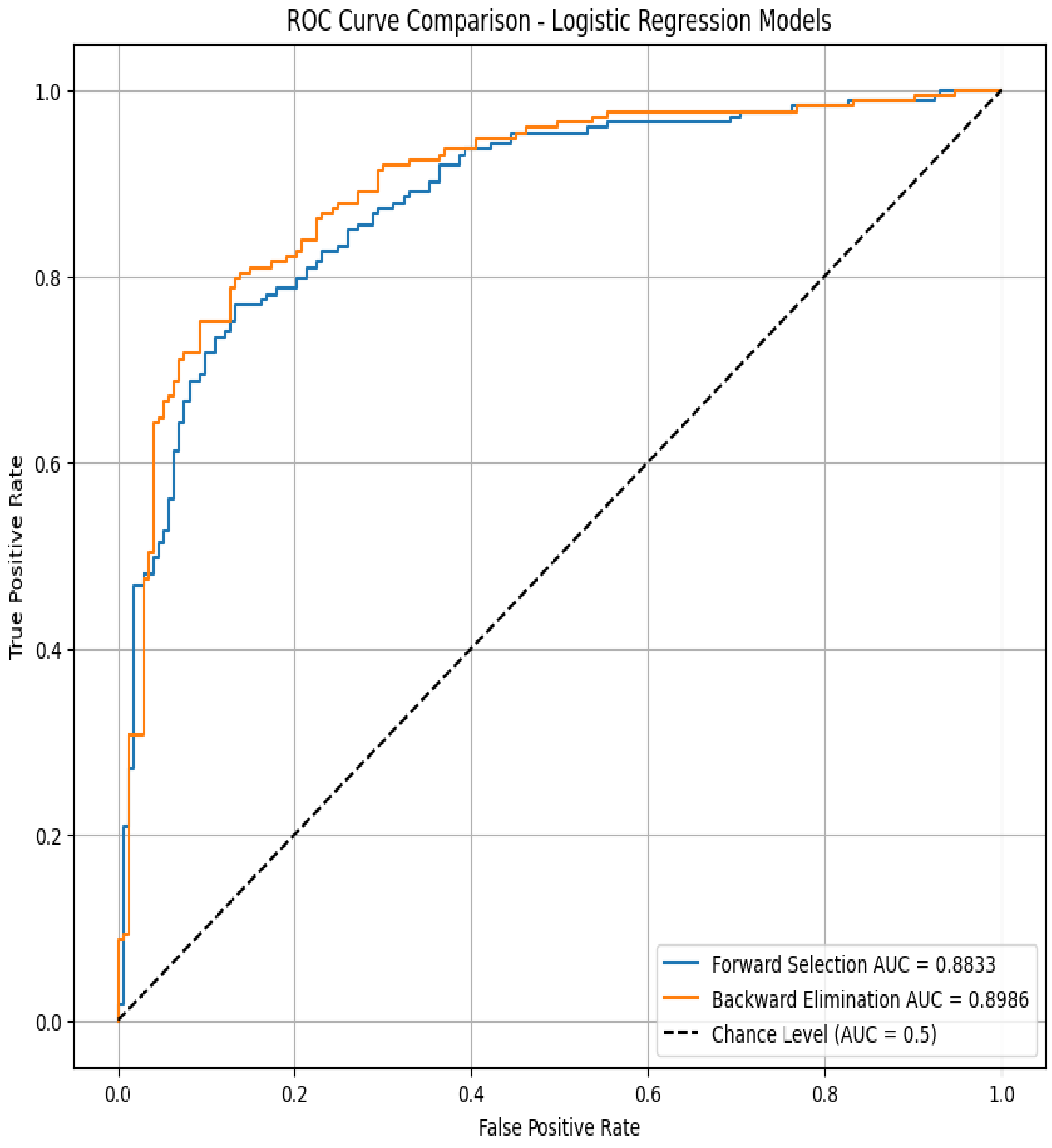
Forward Feature Selection: Accuracy=0.7976, AUC=0.8683, Recall= 0.8016

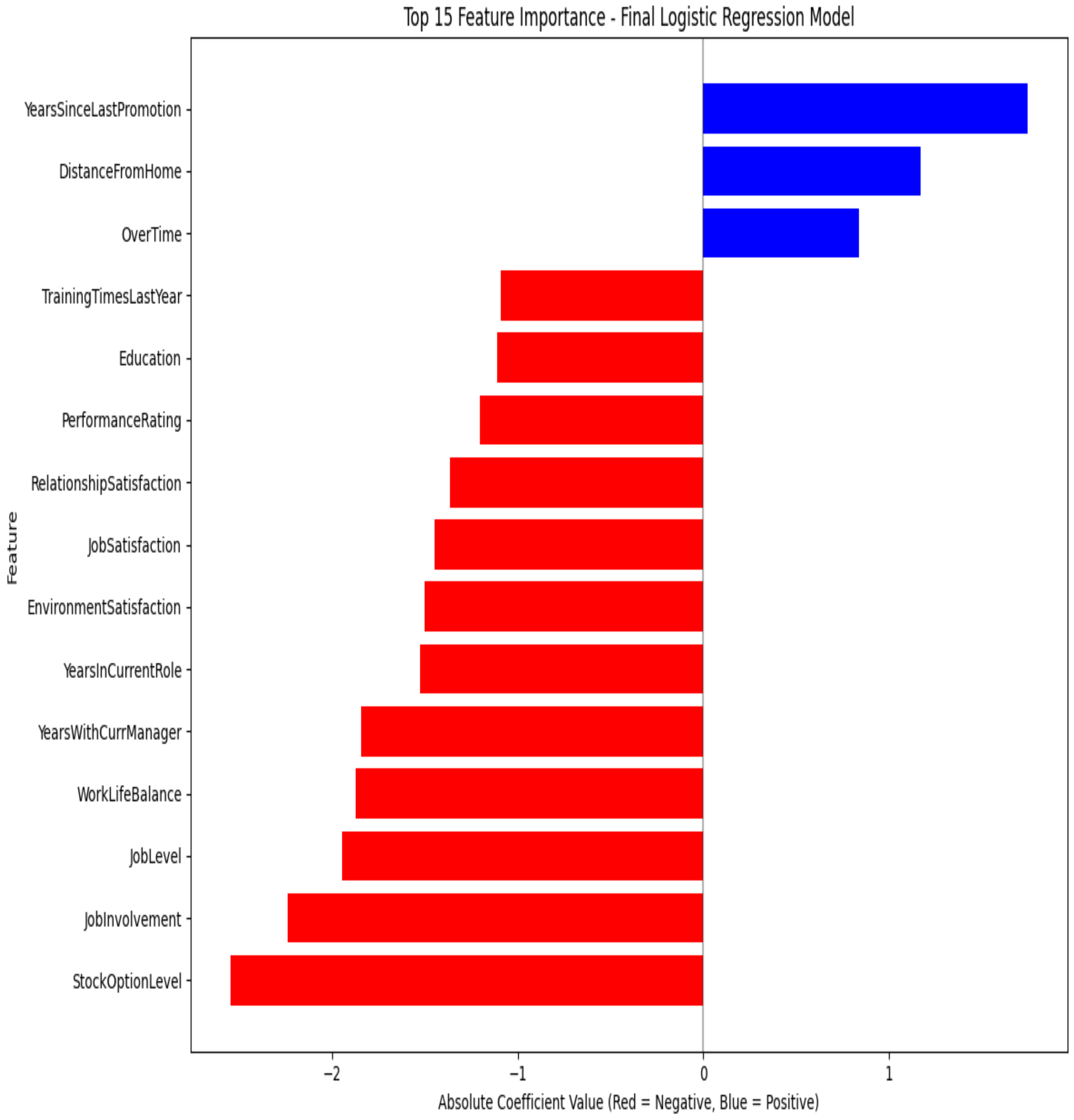
Backward Feature Selection: Accuracy=0.8057, AUC=0.8822, Recall = 0.8016

- Found Backward feature selection model to be the best one

Insights from the top features:

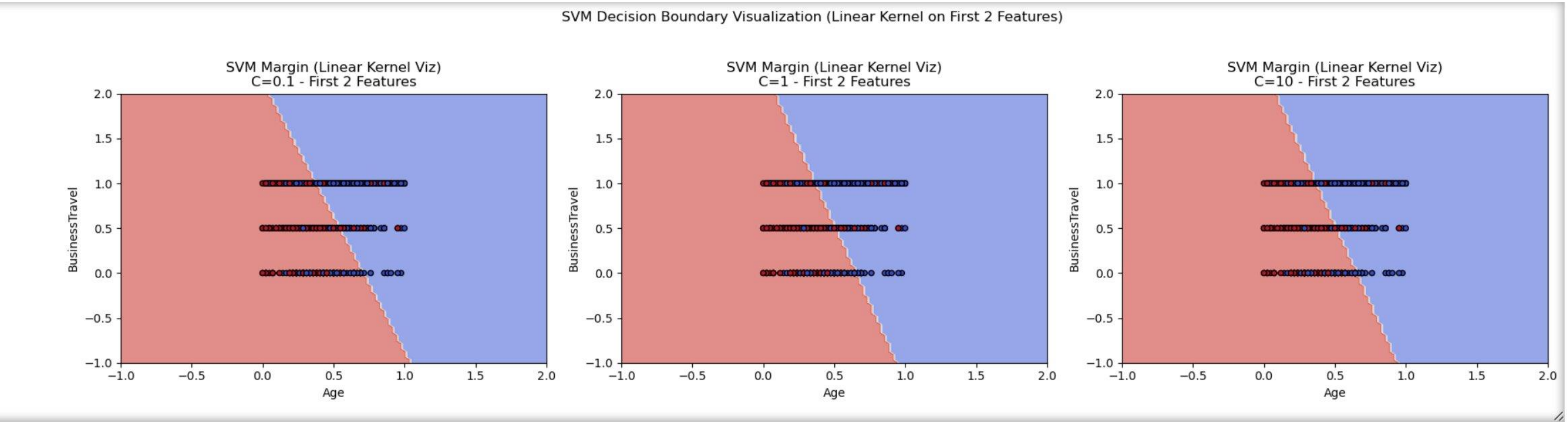
- Higher stock options make employees 2.5 times less likely to leave
- Employees who are more involved in their jobs are 2.2 times less likely to leave
- Higher job levels (more senior positions) reduce departure risk by about 2 times
- Working overtime increases the chance of leaving by 0.8 times
- Living farther from work increases departure risk by 1.2 times
- More years since last promotion significantly increases departure likelihood by 1.7 times

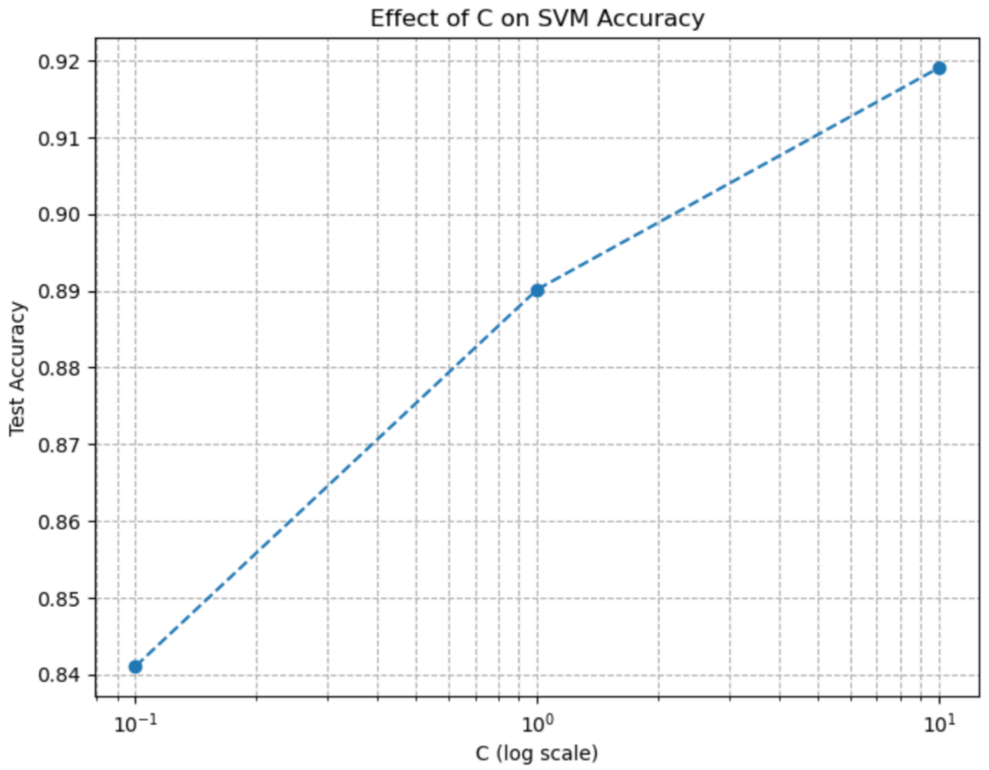

ROC Curve Comparison - Logistic Regression Models
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Forward Selection AUC = 0.8833
Backward Elimination AUC = 0.8986
Chance Level (AUC = 0.5)


Top 15 Feature Importance - Final Logistic Regression Model
Feature
YearsSinceLastPromotion
DistanceFromHome
OverTime
TrainingTimesLastYear
Education
PerformanceRating
RelationshipSatisfaction
JobSatisfaction
EnvironmentSatisfaction
YearsInCurrentRole
YearsWithCurrManager
WorkLifeBalance
JobLevel
JobInvolvement
StockOptionLevel
−2
−1
0
1
Absolute Coefficient Value (Red = Negative, Blue = Positive)

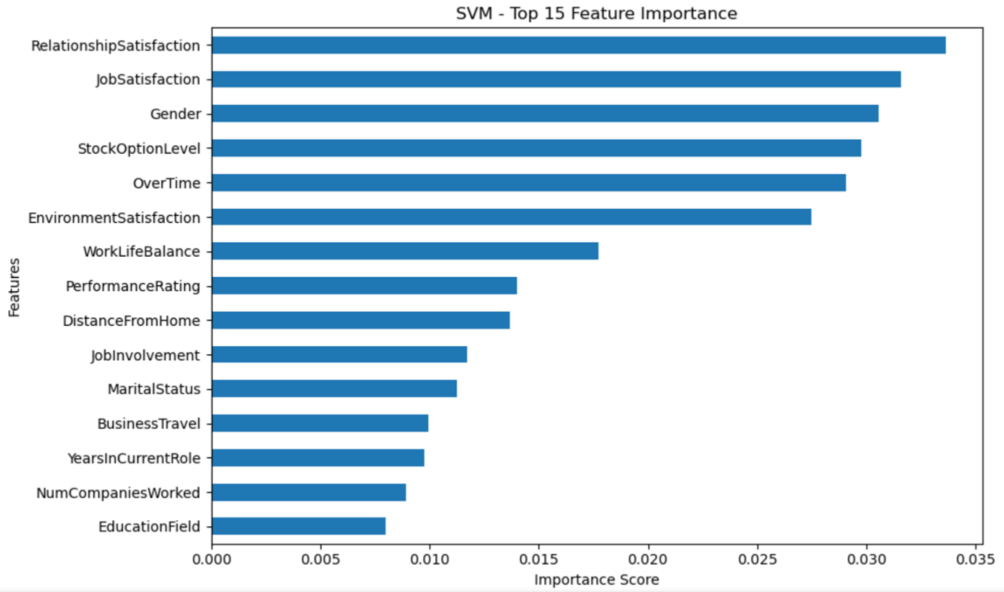
SVM(Support Vector Machine):

- **Tried different C values: 0.1, 1, 10**
- **Accuracy with different C values**: C=0.1: Accuracy=0.8410, C = 1: Accuracy = 0.8902, C = 10: Accuracy = 0.9191
- **Optimal C = 10** gave **AUC = 96.66, Accuracy = 91.91**





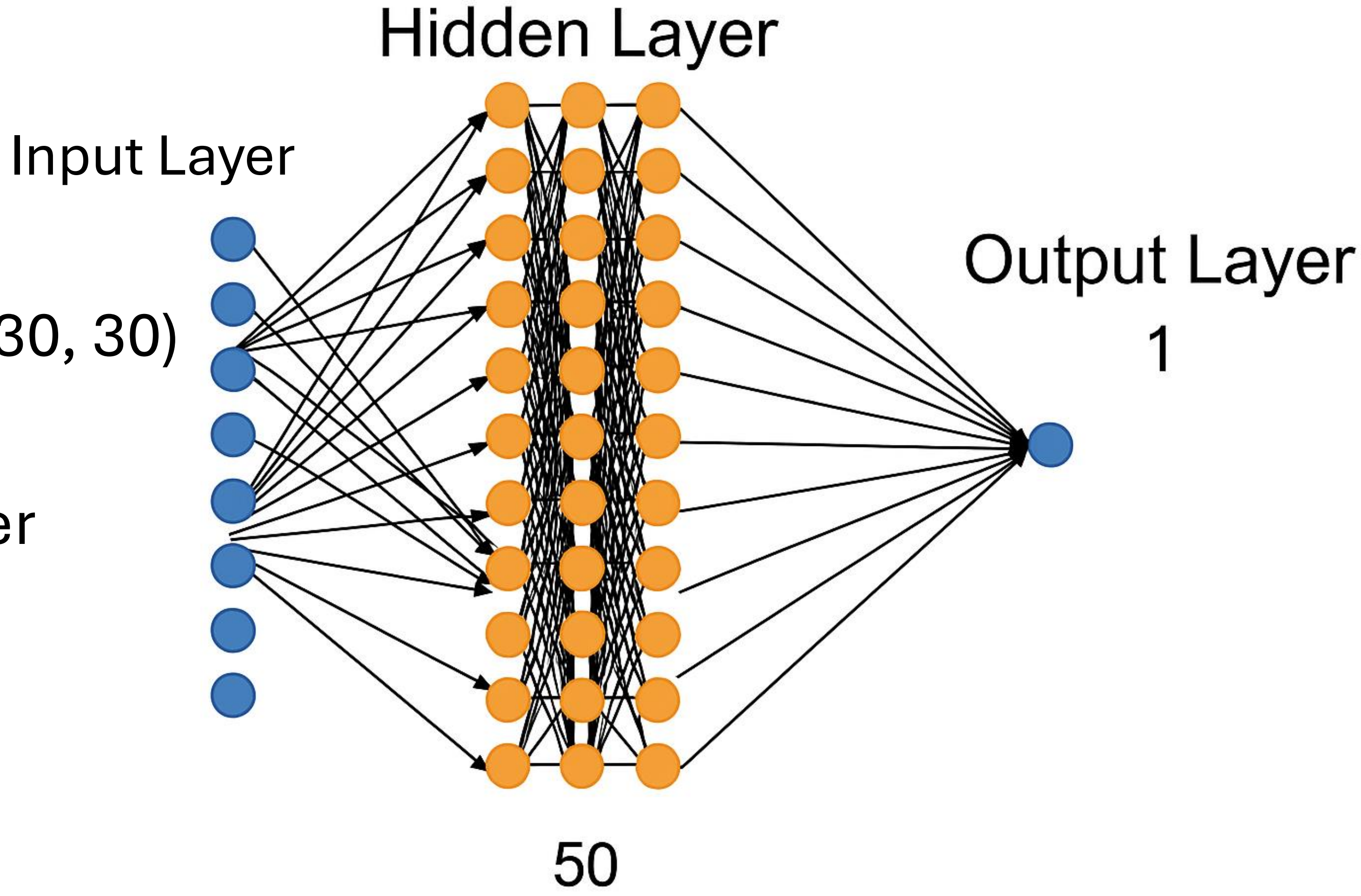
Effect of C on SVM Accuracy
Test Accuracy
C (log scale)
0.92
0.91
0.90
0.89
0.88
0.87
0.86
0.85
0.84
10^{-1}
10^{0}
10^{1}

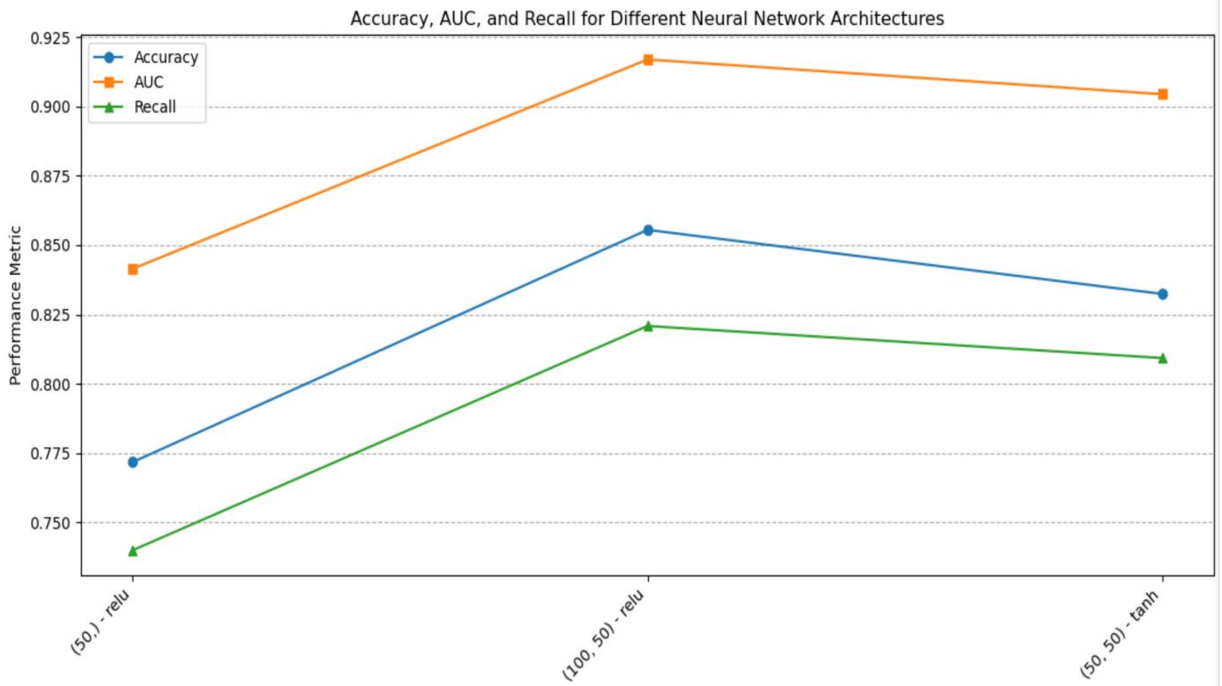


SVM - Top 15 Feature Importance
Features
Importance Score
RelationshipSatisfaction
JobSatisfaction
Gender
StockOptionLevel
OverTime
EnvironmentSatisfaction
WorkLifeBalance
PerformanceRating
DistanceFromHome
JobInvolvement
MaritalStatus
BusinessTravel
YearsInCurrentRole
NumCompaniesWorked
EducationField
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035

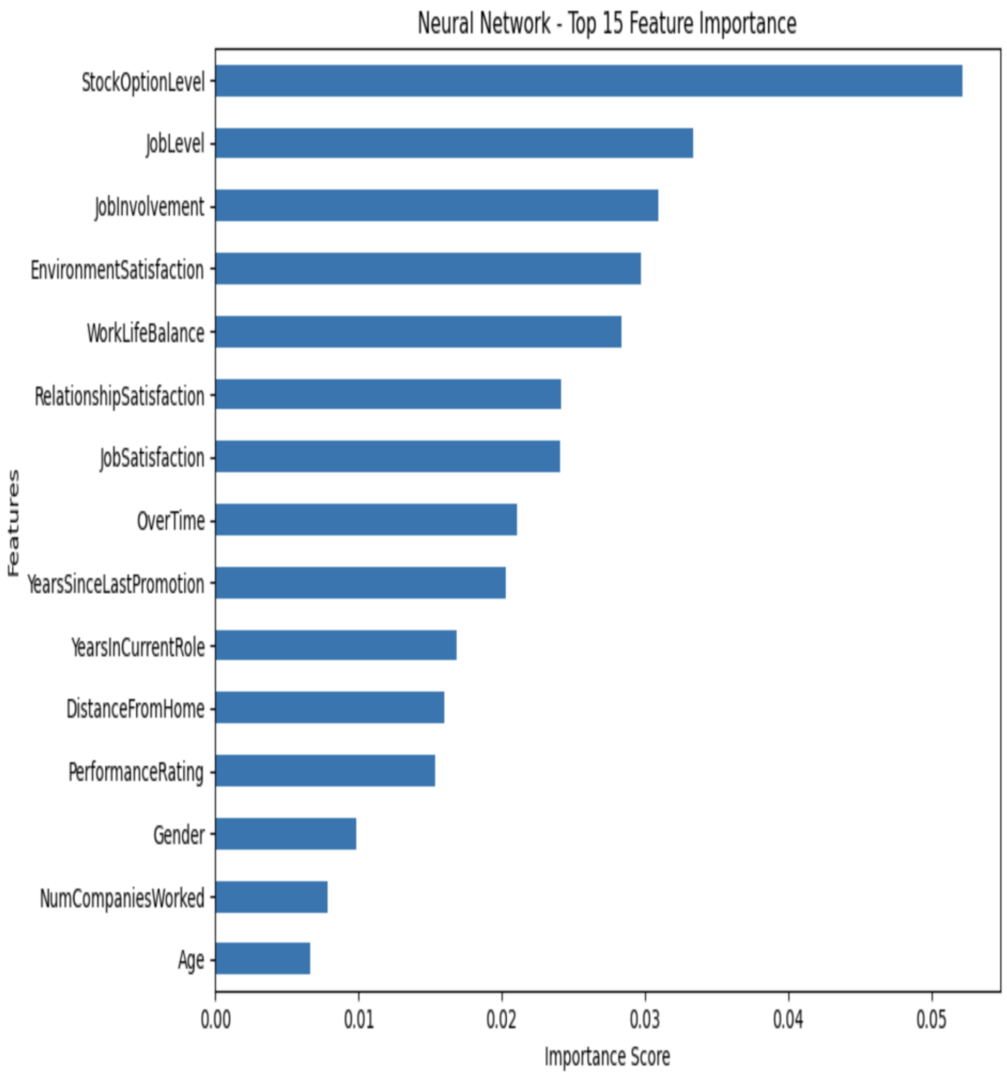
Neural Network:

- MLP CLASSIFIER
- LAYER SIZES: (30,), (50,), (30, 30)
- ALPHA: 0.0001, 0.01
- BEST: 'alpha': 0.0001, 'layer sizes': (50,)
- 'Accuracy': 0.86
- 'Precision': 0.84
- 'Recall': 0.88
- 'AUC': 0.94





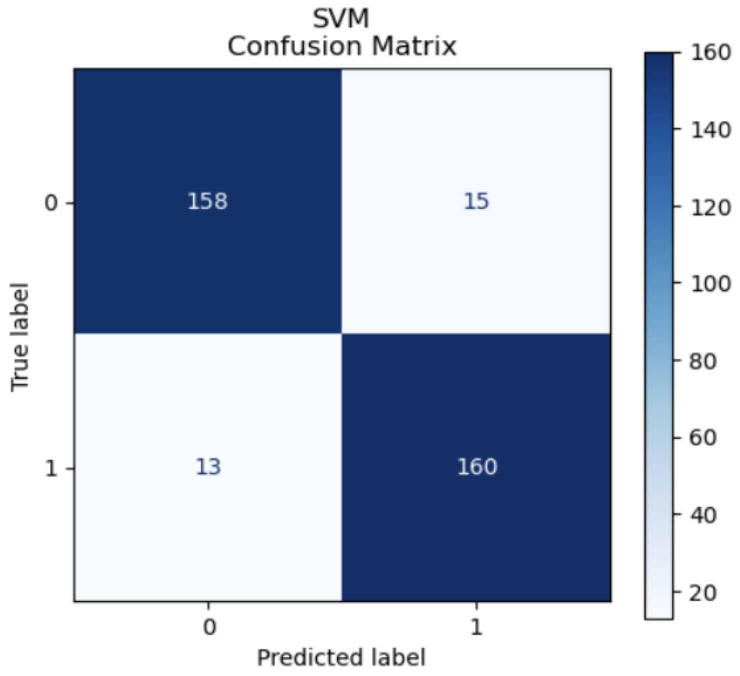
Accuracy, AUC, and Recall for Different Neural Network Architectures
Accuracy
AUC
Recall
Performance Metric
0.925
0.900
0.875
0.850
0.825
0.800
0.775
0.750
(50,) - relu
(100, 50) - relu
(50, 50) - tanh

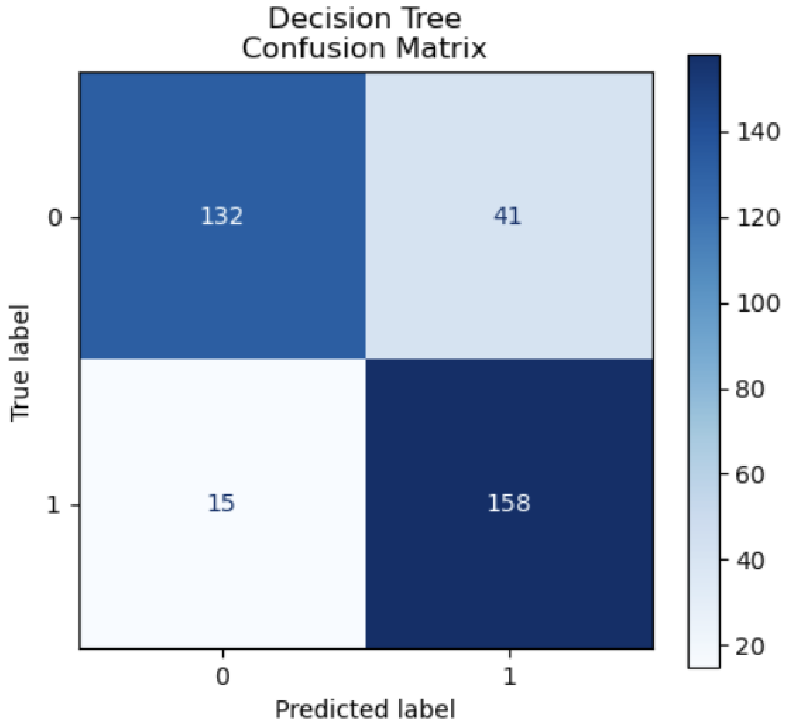


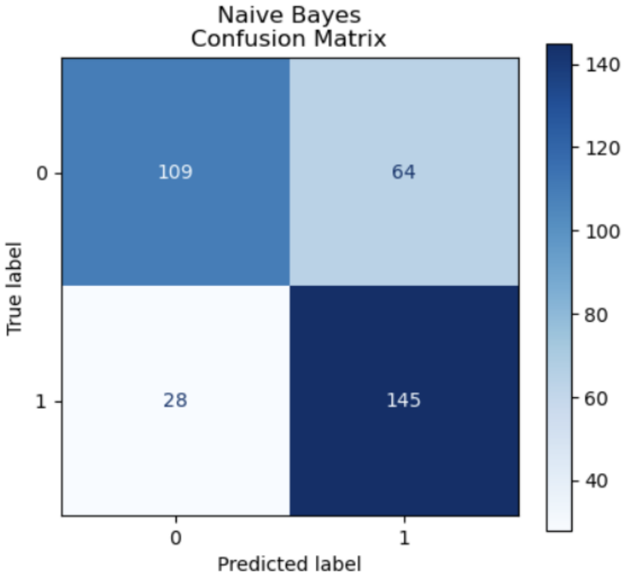
Neural Network - Top 15 Feature Importance
StockOptionLevel
JobLevel
JobInvolvement
EnvironmentSatisfaction
WorkLifeBalance
RelationshipSatisfaction
JobSatisfaction
OverTime
YearsSinceLastPromotion
YearsInCurrentRole
DistanceFromHome
PerformanceRating
Gender
NumCompaniesWorked
Age
Features
0.00
0.01
0.02
0.03
0.04
0.05
Importance Score

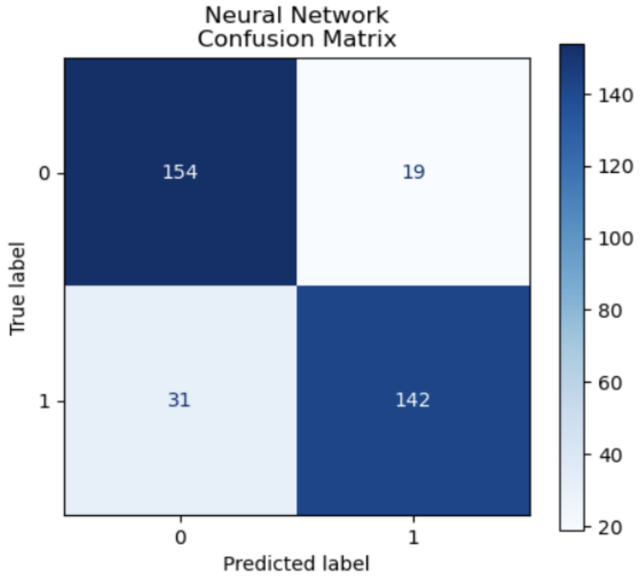
Model Evaluation

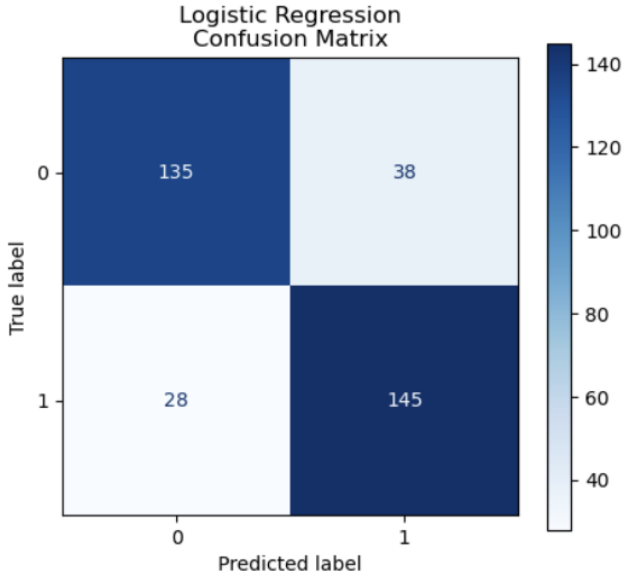
Which? Why?
What? How?





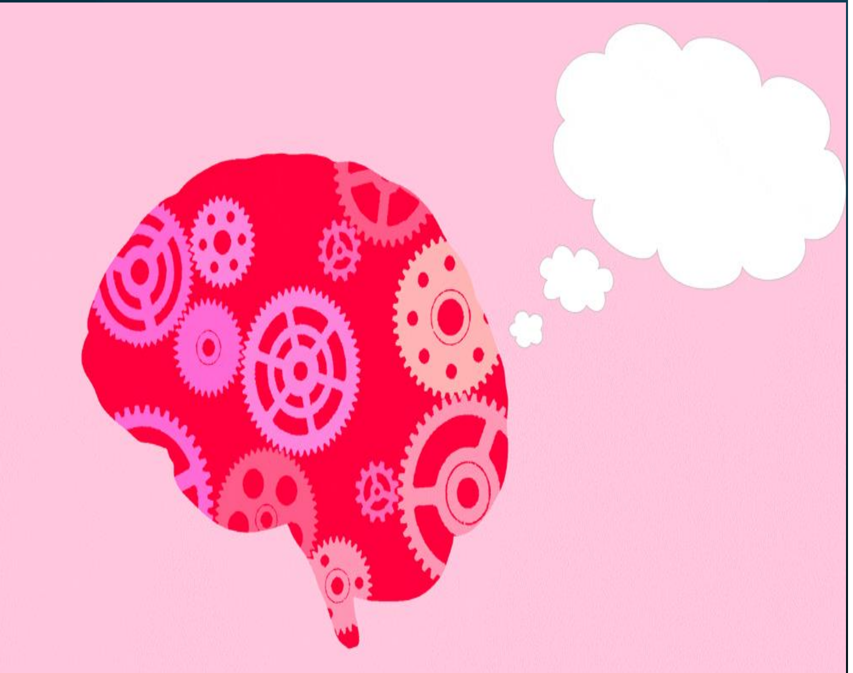




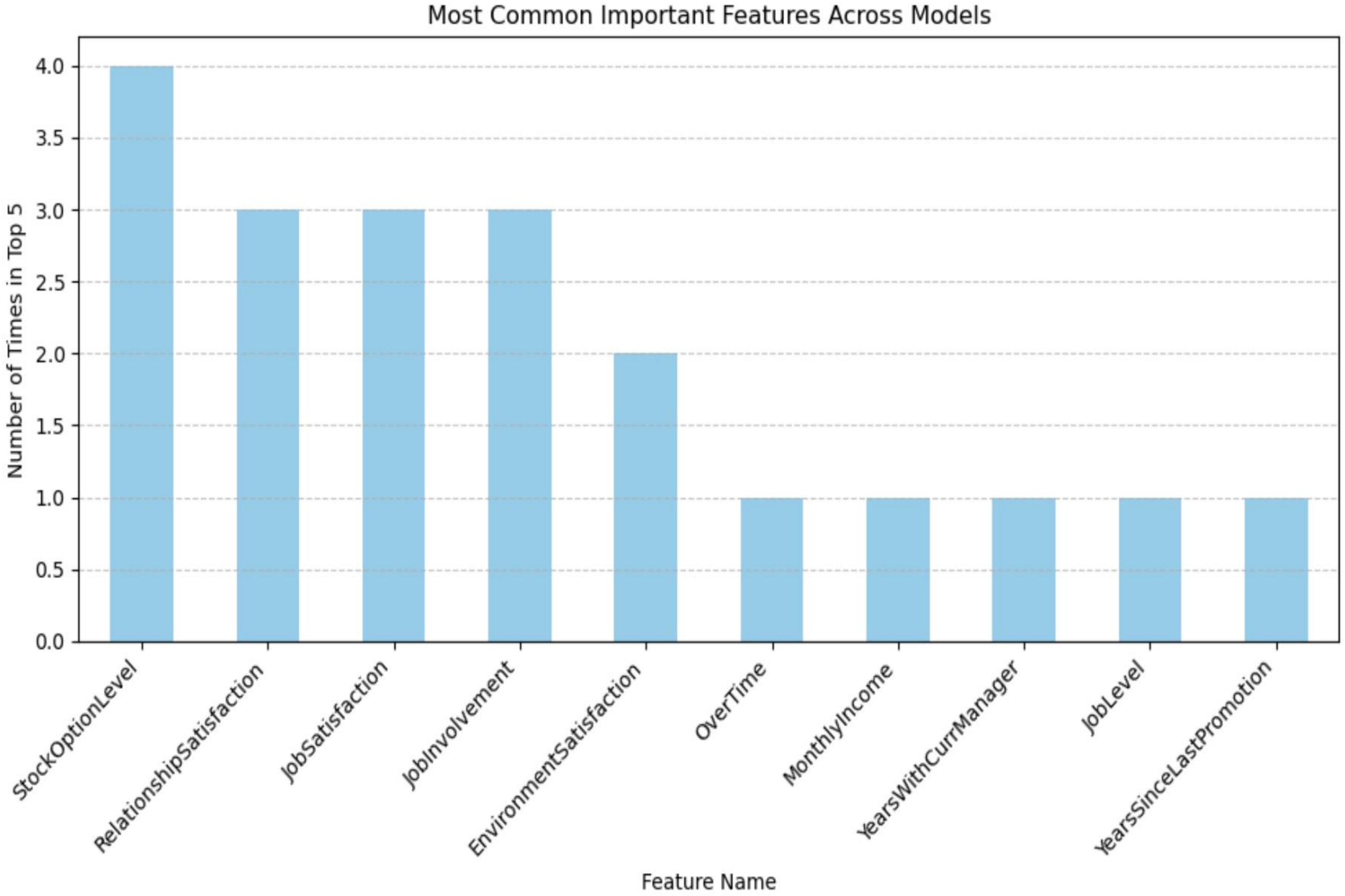


Measure of Importance

Recall

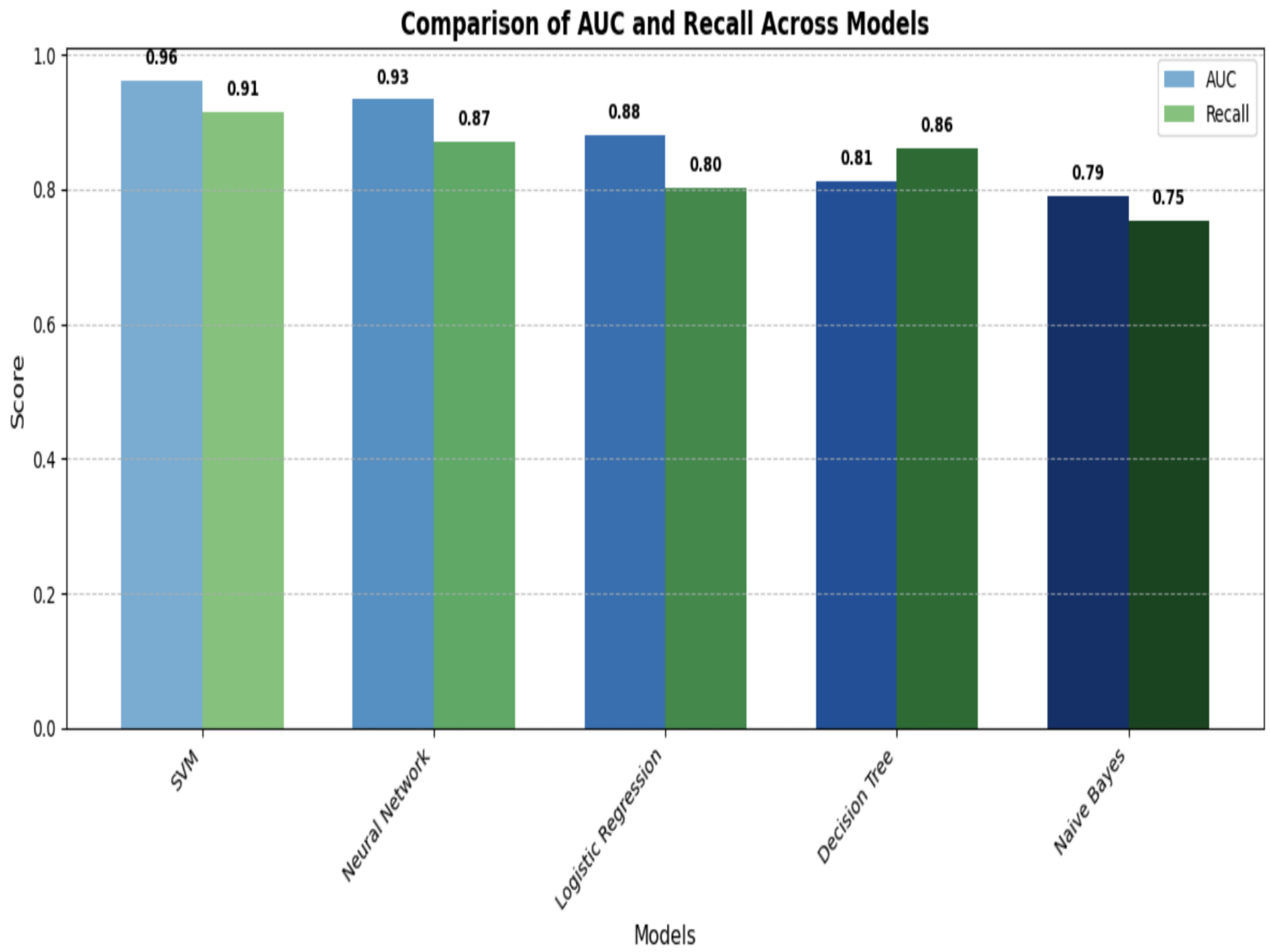


Top Features



So far..

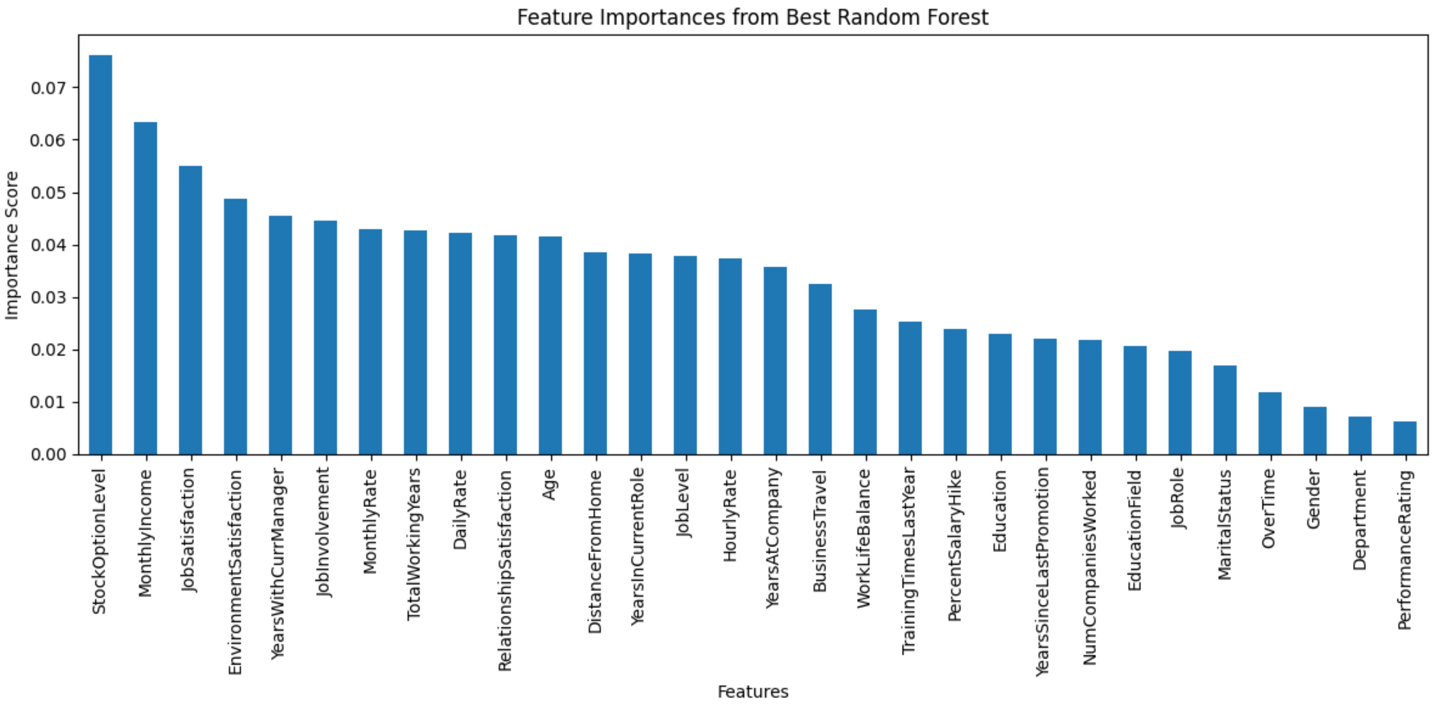
Model	HyperParameters
SVM	C=10
Neural Network	hidden_layer_sizes': (100, 50), 'activation': 'relu'
Logistic Regression	Backward Elimination (17 features)
Decision Tree	max_depth=10
Naive Bayes	No Hyperparameters Tuned

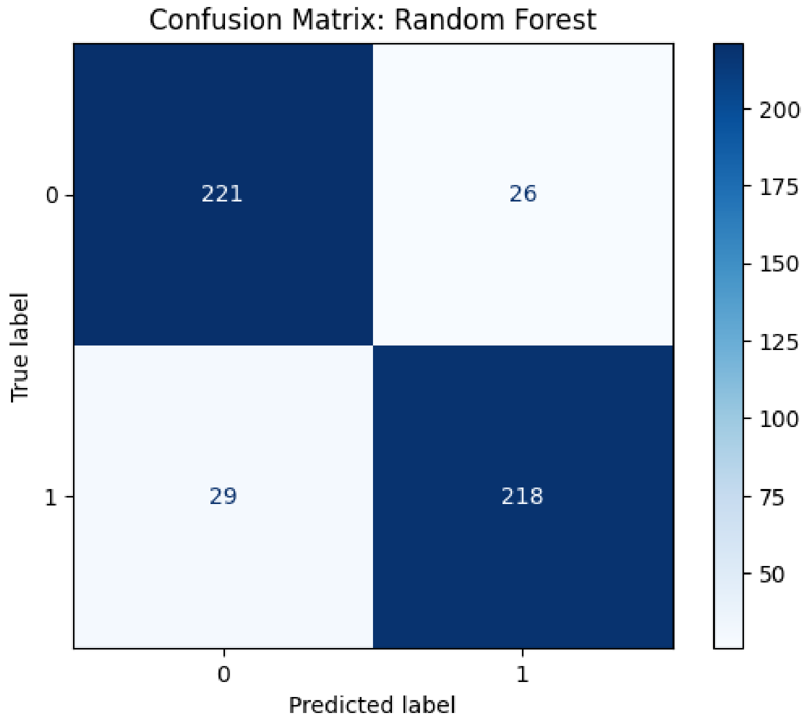


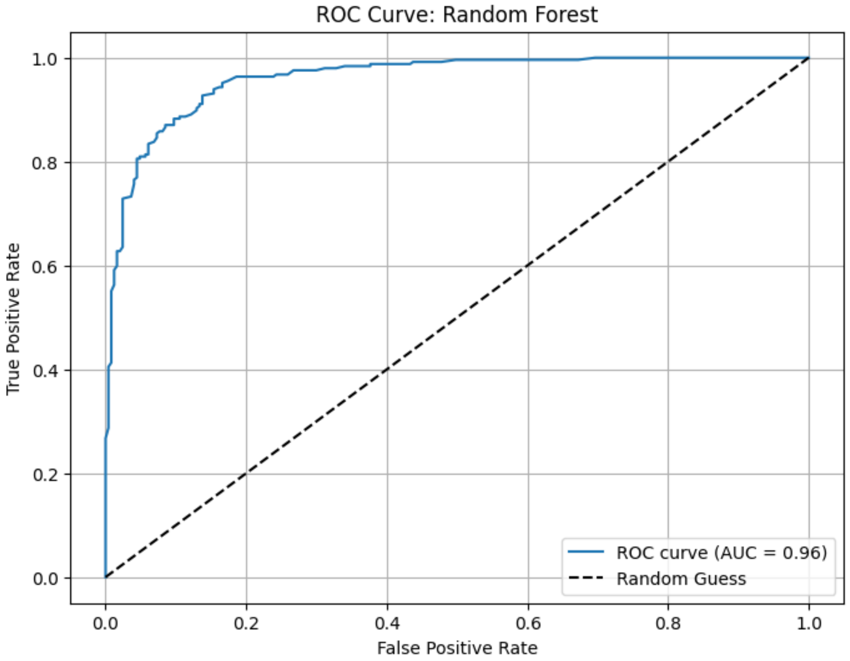
Random Forest

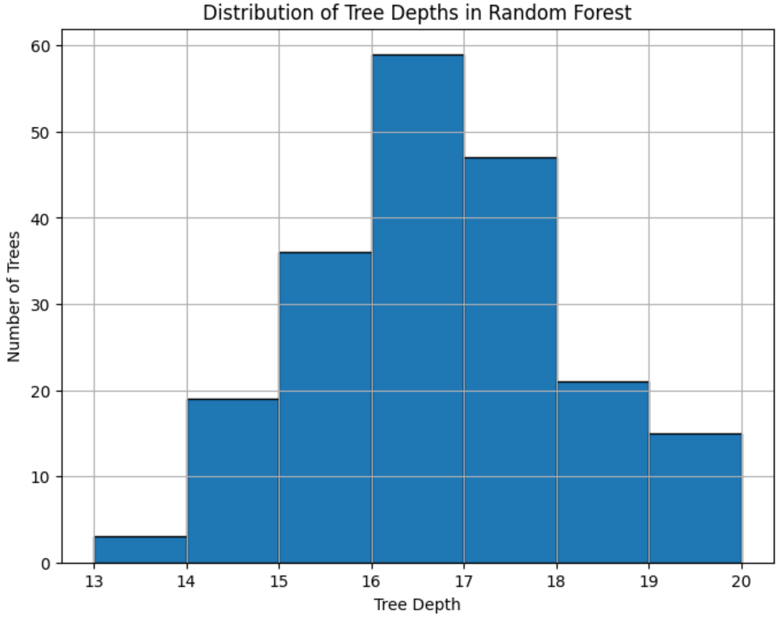
Performance Metrics:
- Accuracy: 88.9%
- AUC Score: 0.96
- Recall: 88.3% (Strong at identifying at-risk employees)

Optimal Parameters:
- 200 Estimators
- Max Depth = 20
- Min Samples Split = 2

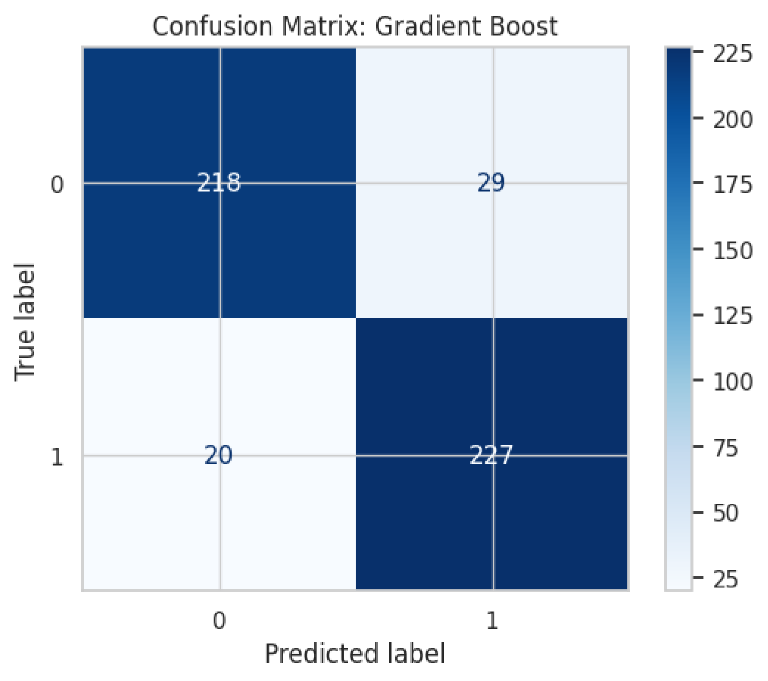

Feature Importances from Best Random Forest
Importance Score
Features
StockOptionLevel
MonthlyIncome
JobSatisfaction
EnvironmentSatisfaction
YearsWithCurrManager
JobInvolvement
MonthlyRate
TotalWorkingYears
DailyRate
RelationshipSatisfaction
Age
DistanceFromHome
YearsInCurrentRole
JobLevel
HourlyRate
YearsAtCompany
BusinessTravel
WorkLifeBalance
TrainingTimesLastYear
PercentSalaryHike
Education
YearsSinceLastPromotion
NumCompaniesWorked
EducationField
JobRole
MaritalStatus
OverTime
Gender
Department
PerformanceRating

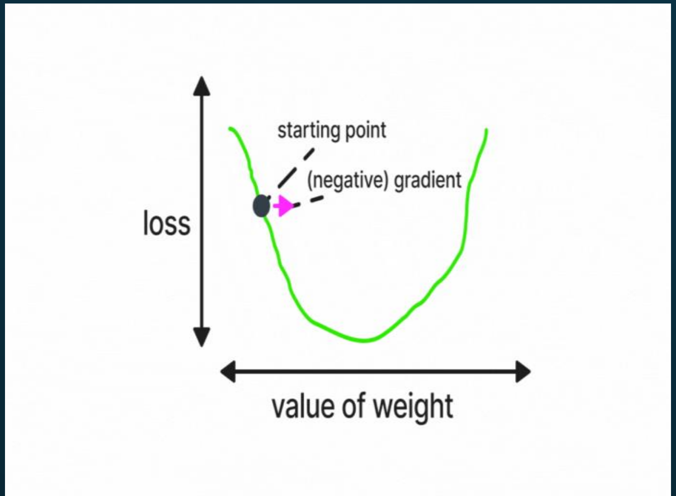

Confusion Matrix: Random Forest
True label
Predicted label
221
26
29
218

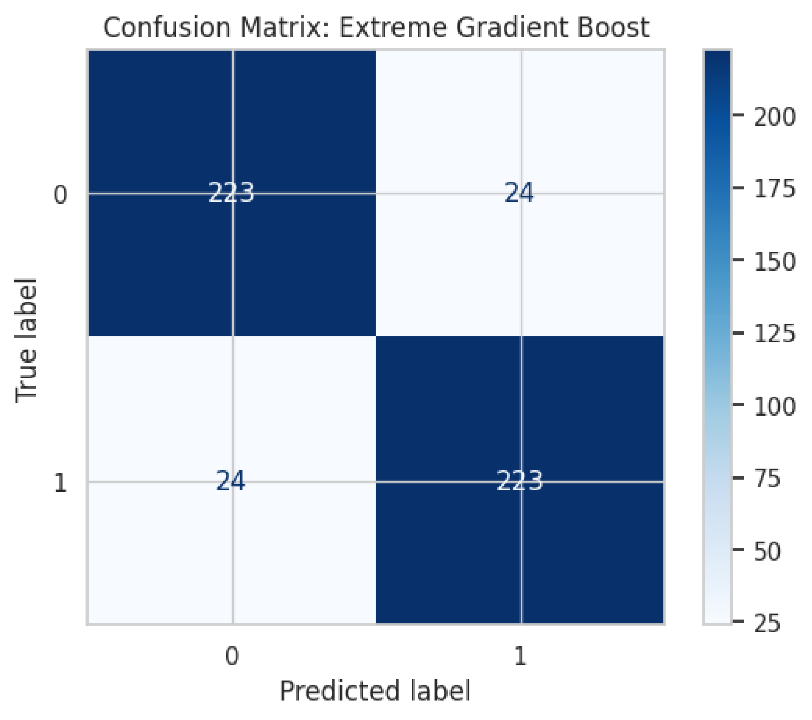

ROC Curve: Random Forest
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.96)
Random Guess


Distribution of Tree Depths in Random Forest
Number of Trees
Tree Depth

Gradient Boosting & XGBoost







Contrasting the two

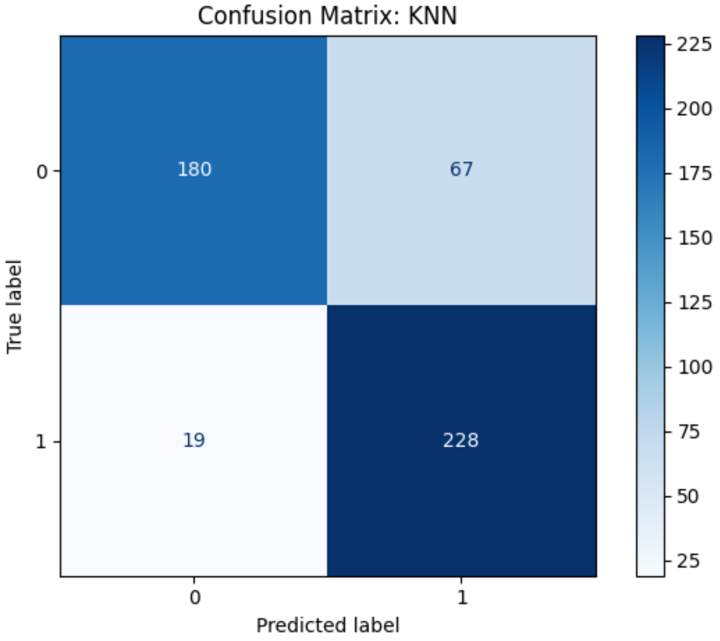
Aspect	Gradient Boost	XGBoost
Accuracy	90%	90%
Recall	92%	90%
Precision	88%	90%
AUC	97%	97.02%
Simplicity	Simpler Hyperparameters (3 tuned)	More Tuned Hyperparameters (6 tuned)
Regularization	None	L1 (alpha) and L2 (lambda) regularization
Feature Focus	Income & Satisfaction	Position, Career Progression

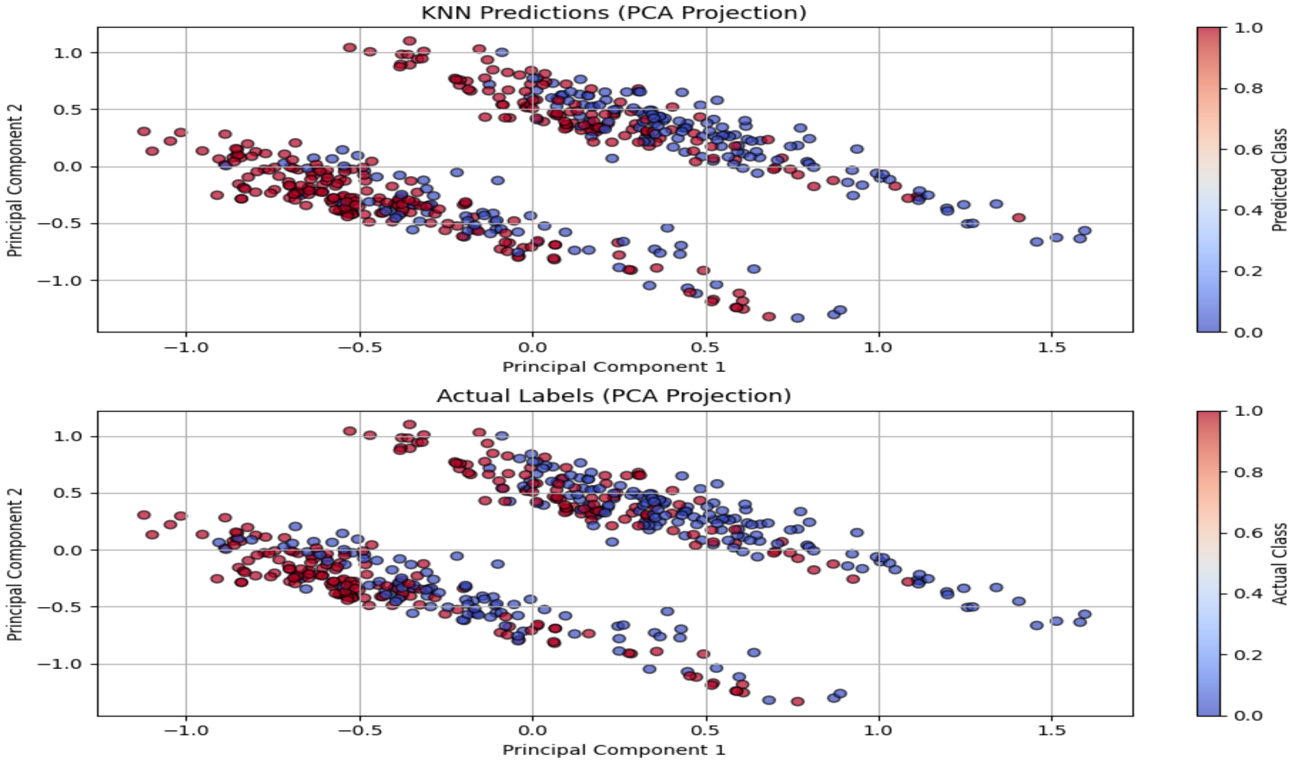
- **Gradient Boost** offers slightly better recall with a simpler setup.
- **XGBoost** provides stronger regularization, slightly better AUC, and more balanced generalization across features.

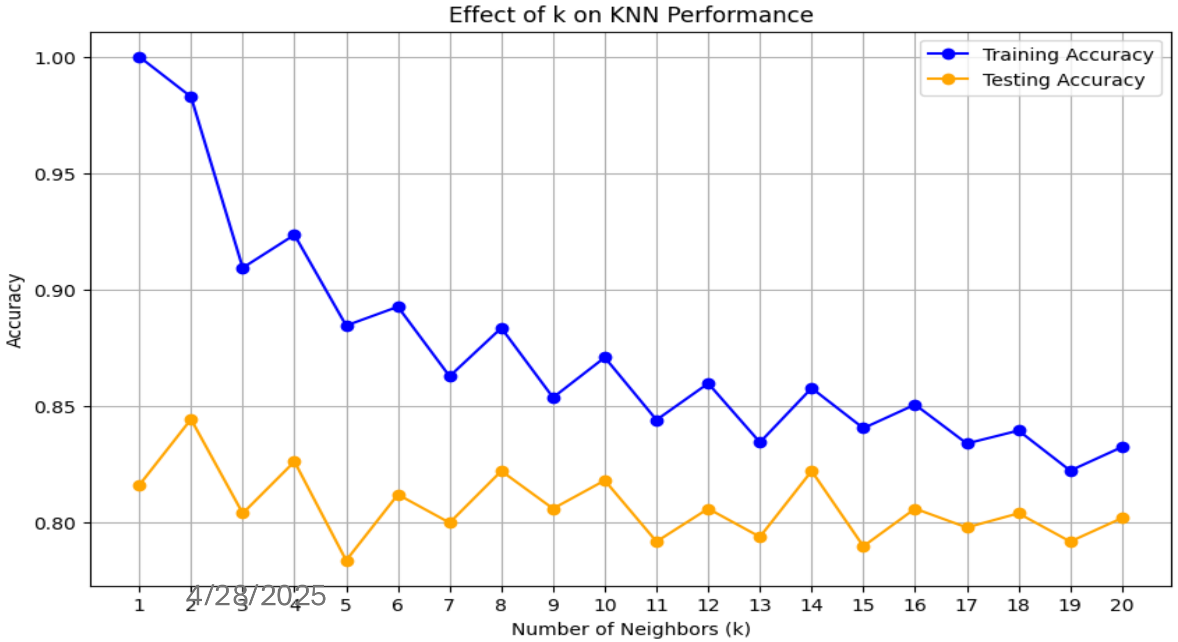
KNN

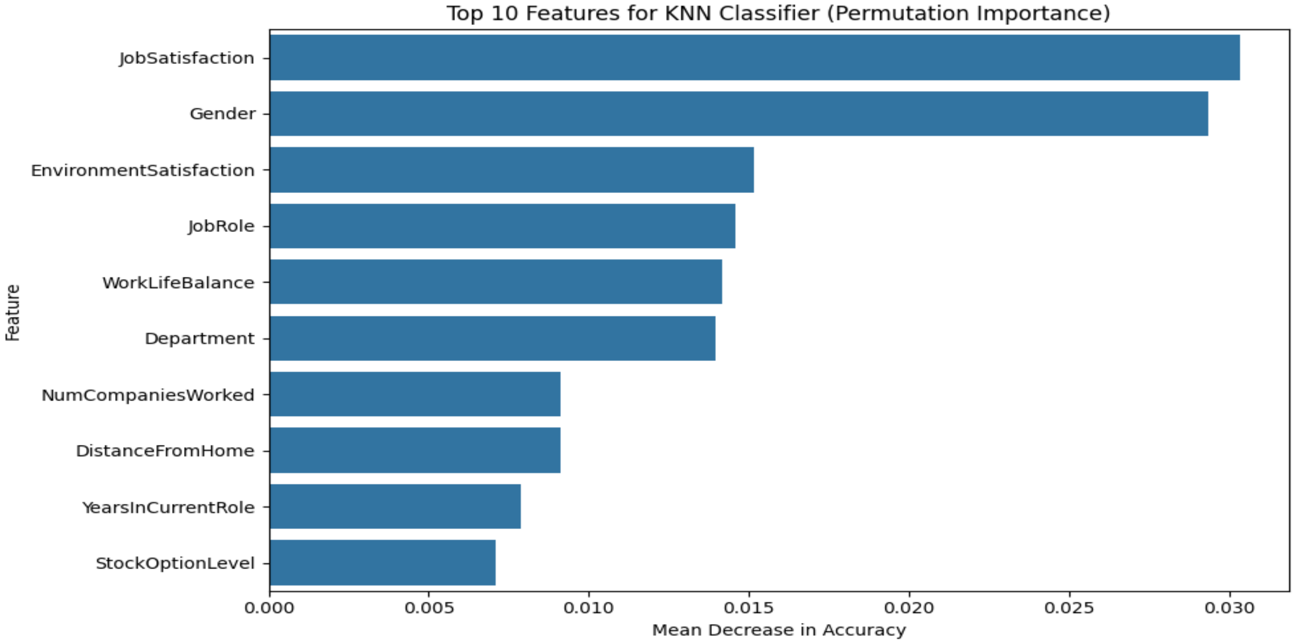
- KNN achieved 82.6% accuracy and 0.90 AUC score, making it one of our top performers
- Optimal model uses 4 nearest neighbors to make predictions
- Exceptional at identifying at-risk employees with 92.3% recall rate
- Feature importance analysis revealed JobSatisfaction and Gender as the key predictors
- Works well with our dataset without requiring complex parameter tuning
- Provides excellent balance between computational efficiency and predictive power

K-NN


Confusion Matrix: KNN
True label
Predicted label
0
1
180
67
19
228
225
200
175
150
125
100
75
50
25

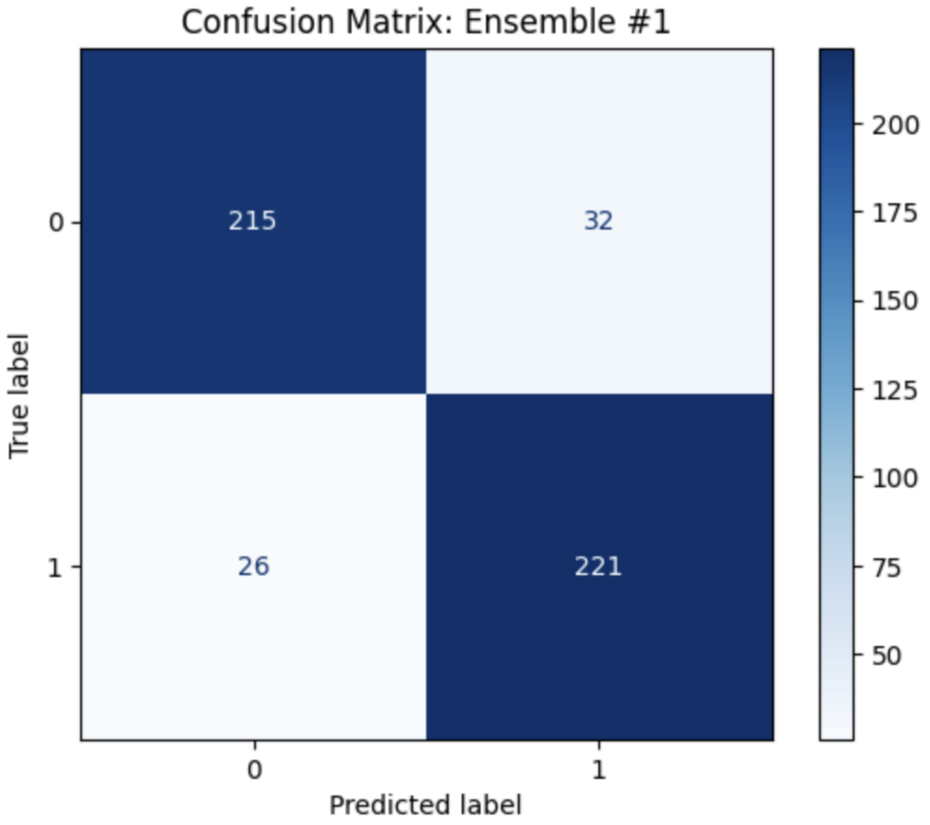

KNN Predictions (PCA Projection)
Principal Component 1
Principal Component 2
Predicted Class
Actual Labels (PCA Projection)
Principal Component 1
Principal Component 2
Actual Class

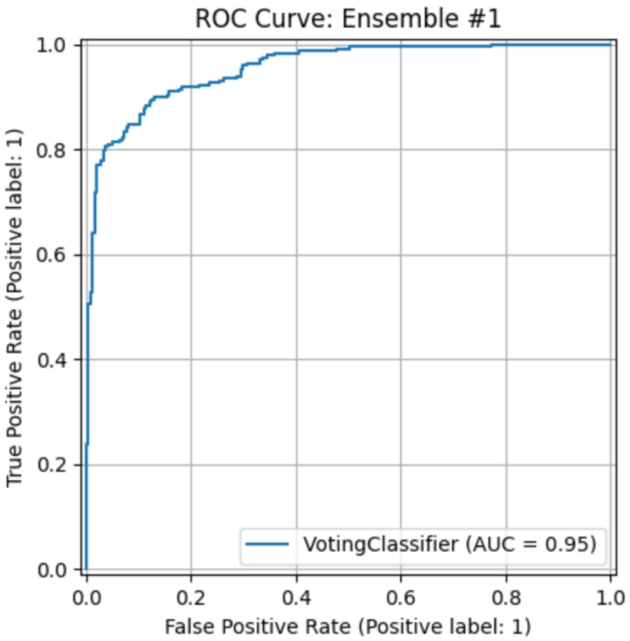

Effect of k on KNN Performance
Training Accuracy
Testing Accuracy
Accuracy
Number of Neighbors (k)

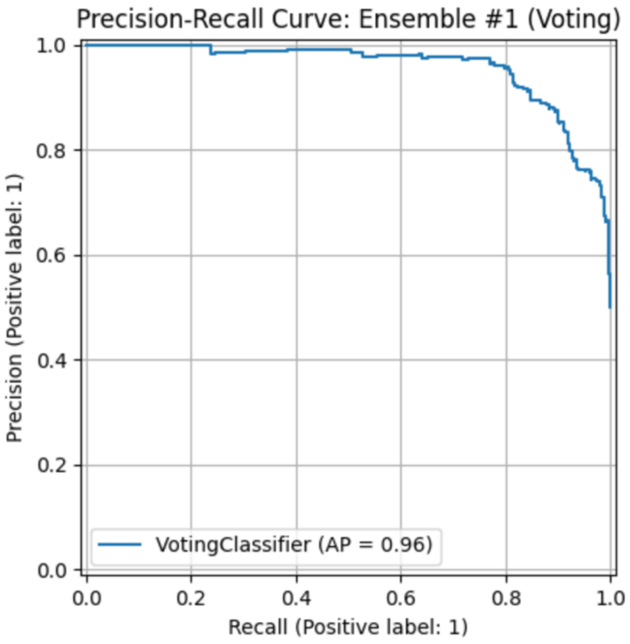

Top 10 Features for KNN Classifier (Permutation Importance)
JobSatisfaction
Gender
EnvironmentSatisfaction
JobRole
WorkLifeBalance
Department
NumCompaniesWorked
DistanceFromHome
YearsInCurrentRole
StockOptionLevel
Feature
Mean Decrease in Accuracy

Ensemble Method #1 (SVM, DT, NN)

- 'Accuracy': 0.88
- 'Precision': 0.87
- 'Recall': 0.89
- 'AUC': 0.95

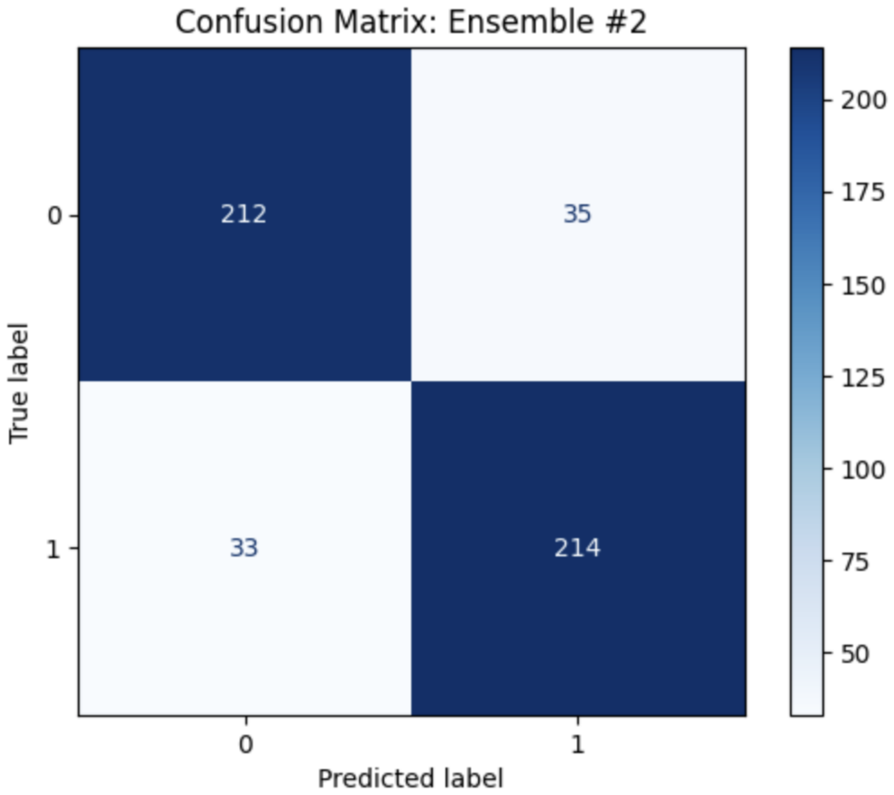


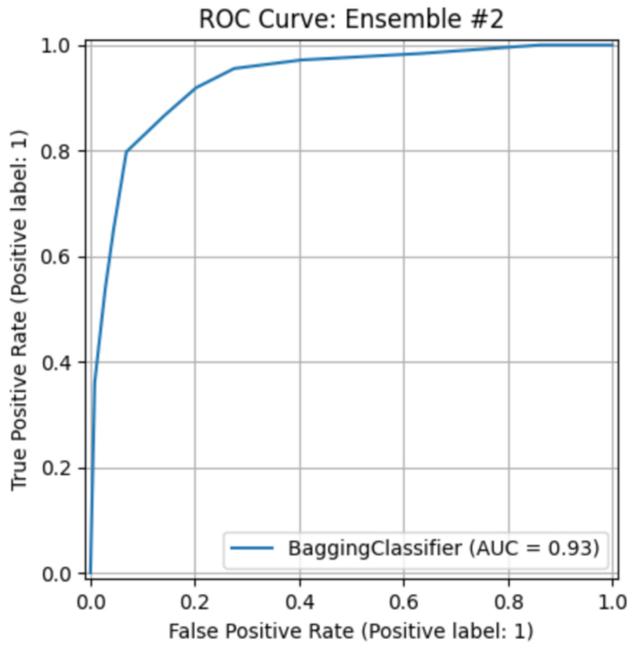


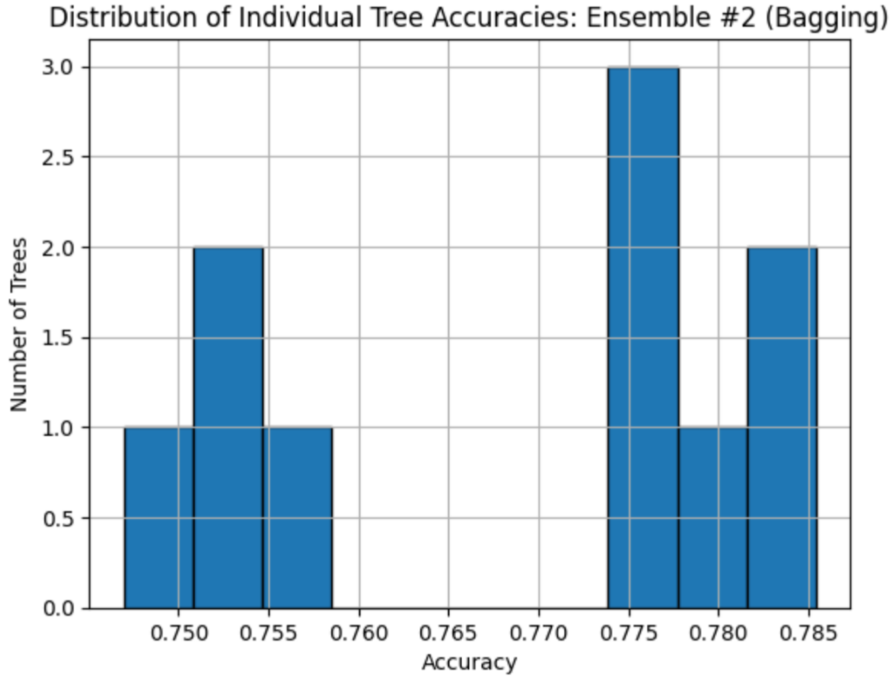


Ensemble Method #2 (Bagging with Decision Trees)

- 'Accuracy': 0.86
- 'Precision': 0.86
- 'Recall': 0.87
- 'AUC': 0.93

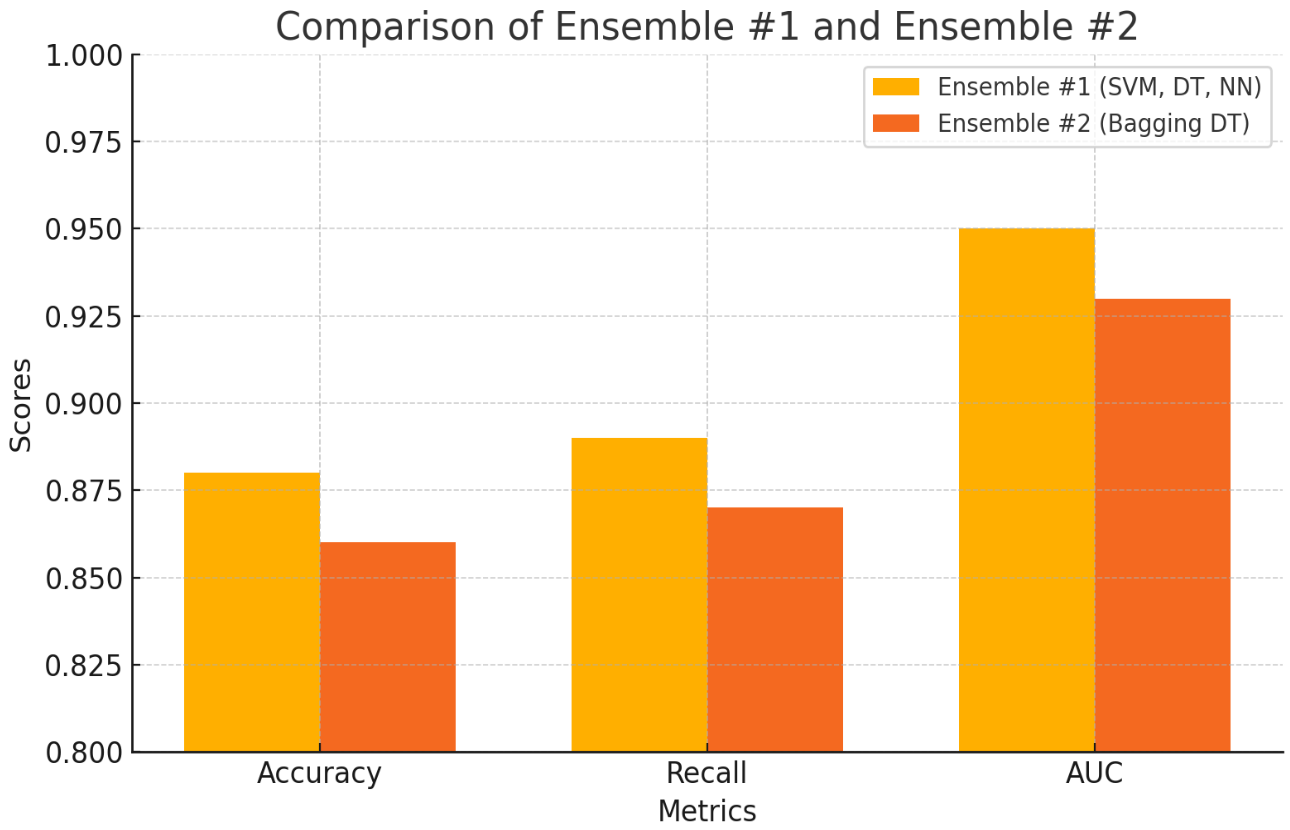






Comparison of 2 Ensemble models

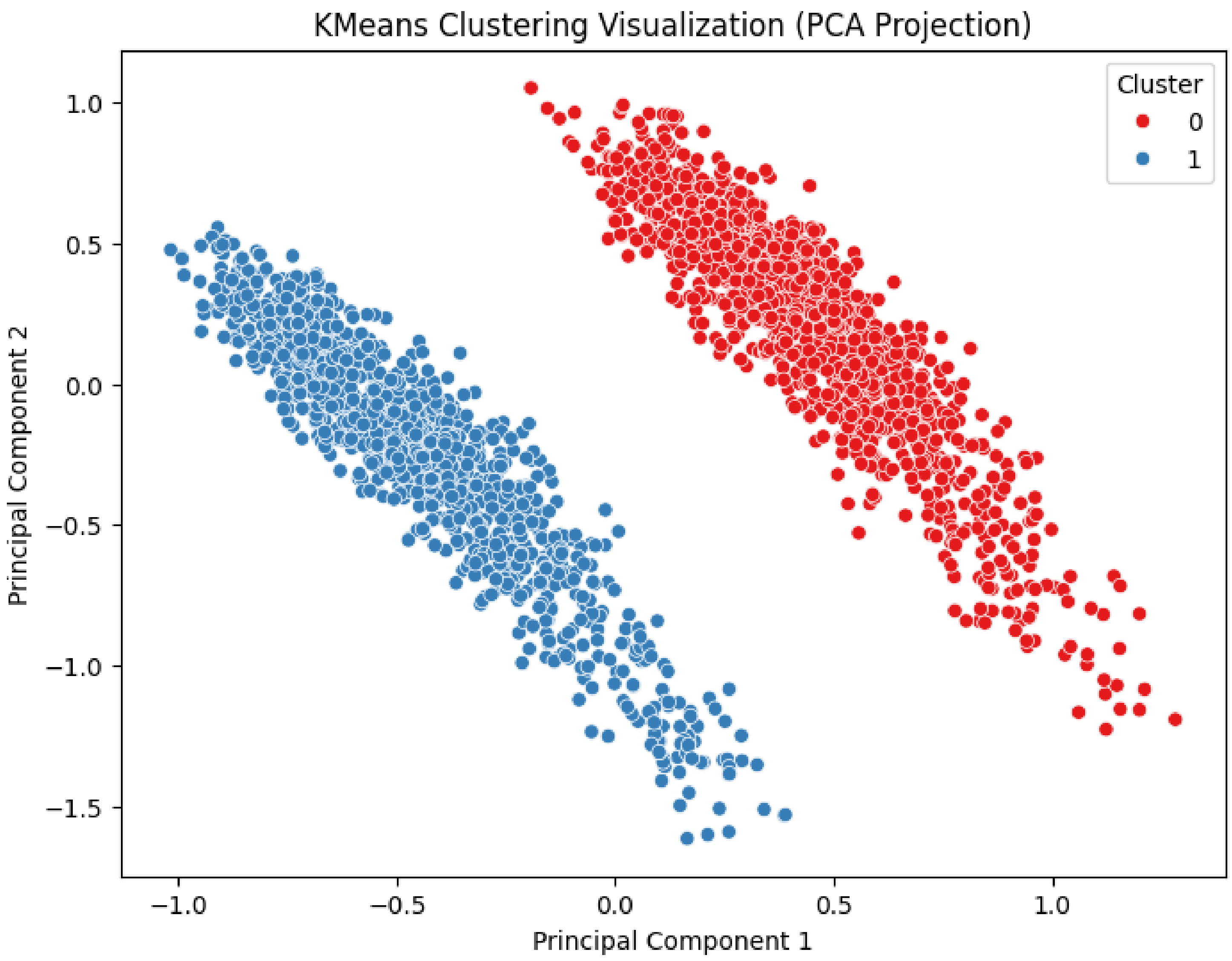
- Ensemble #1 has **2% higher accuracy**, meaning it correctly predicts attrition slightly better overall
- Ensemble #1 captures **more true positives** (employees leaving) — important for the business because missing leavers is costly.
- Ensemble #1 again shows **stronger separation** between employees who leave vs. those who stay.

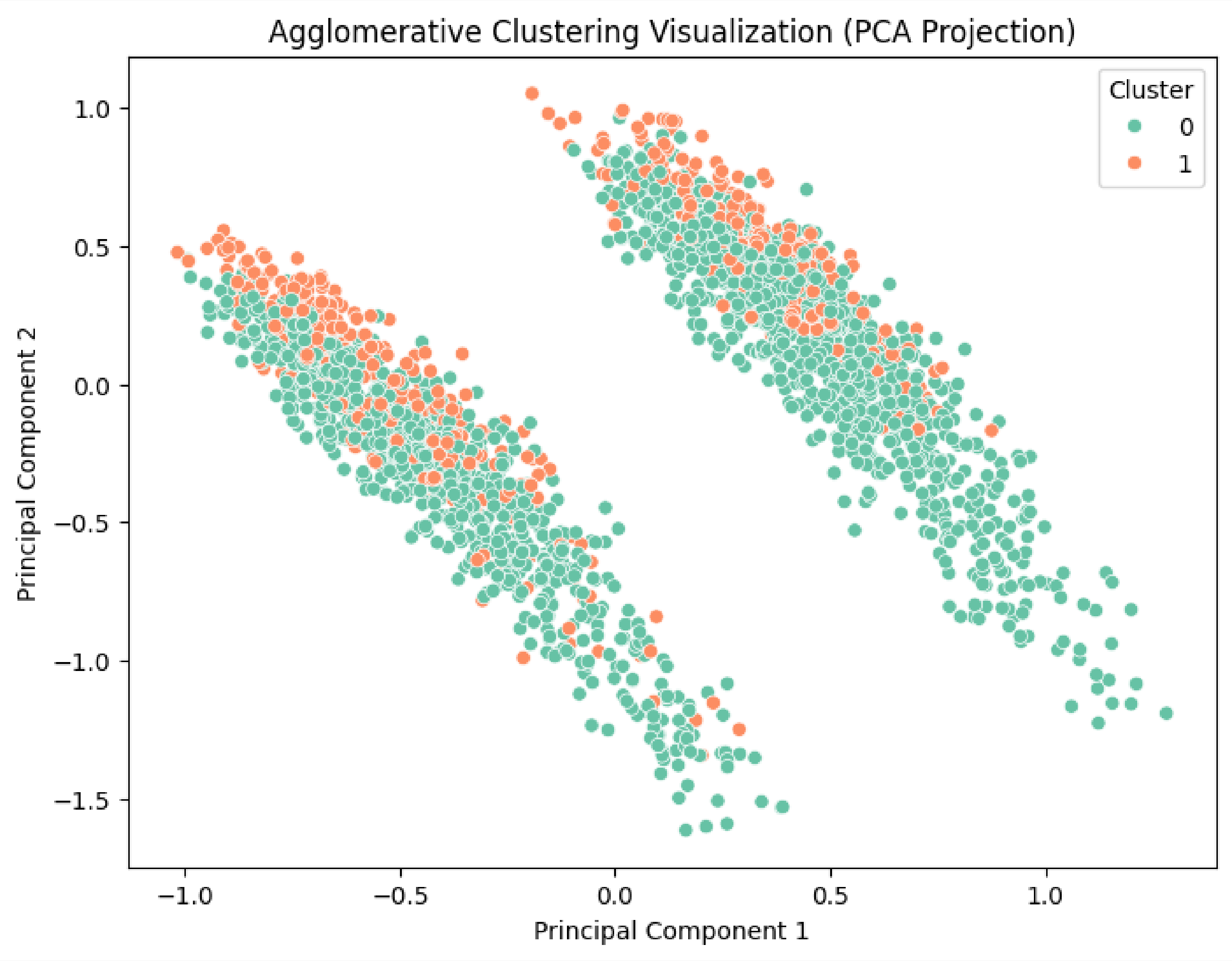


K Means

- **KMeans applied** with **k=2** clusters (because attrition is a binary phenomenon here).

- **Silhouette Score: 0.1023** — low score, suggesting moderate or weak separation between clusters.





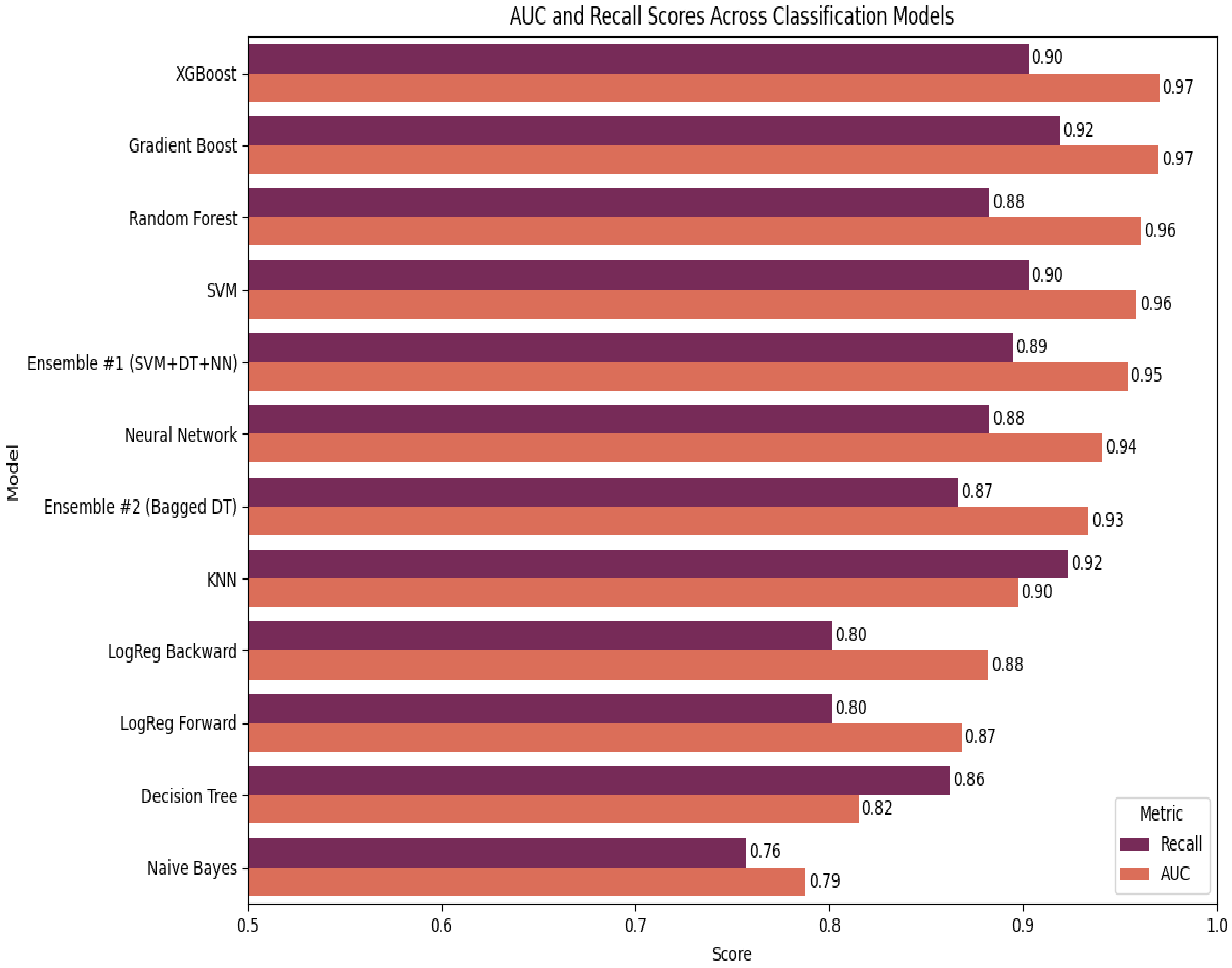
Agglomerative (Hierarchical) Clustering

- **Agglomerative Clustering applied** with **n_clusters=2**.
- **Silhouette Score**: **0.0726** — even lower than KMeans, meaning the cluster separation is weaker.
- **Visualization**: PCA scatter plot + **Dendrogram** built using Ward's method.
- **Dendrogram**: Shows cluster merging patterns; large vertical gaps suggest natural cluster separation points.
-

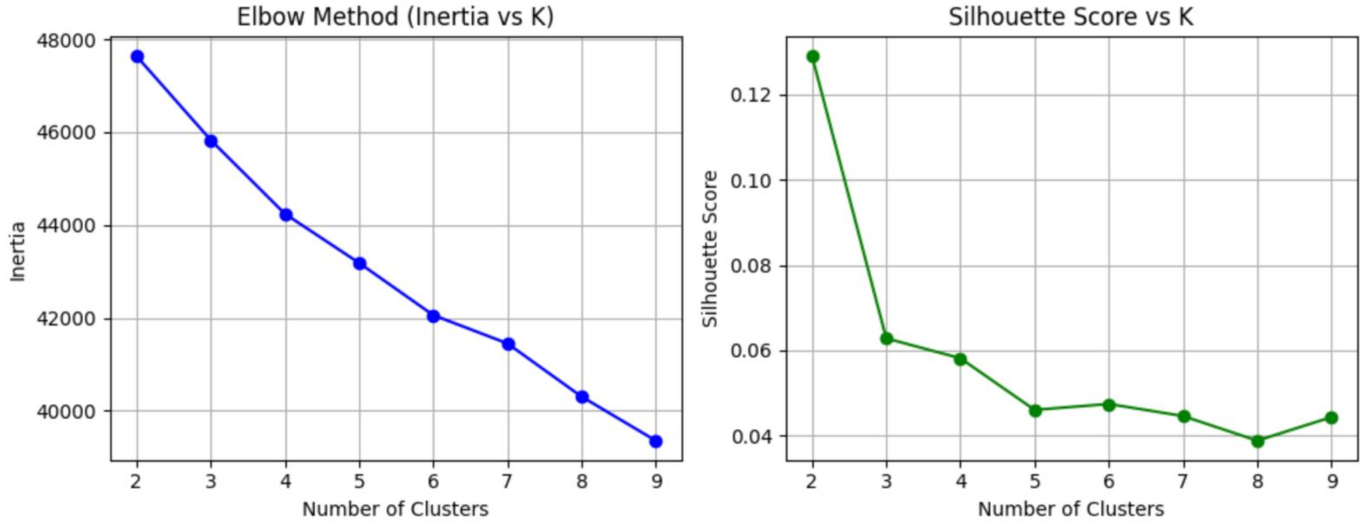
Chosen Hyperparameters for each model

	Model	Best Hyperparameters
0	Random Forest	n_estimators=200, max_depth=20
1	Extreme Gradient Boost	n_estimators=200, max_depth=5, learning_rate=0.1
2	Gradient Boost	n_estimators=200, max_depth=5, learning_rate=0.1
3	Ensemble #1 (SVM+DT+NN)	Soft voting (SVM, DT, NN)
4	Neural Network	hidden_layers=(100,), activation=relu, solver=adam
5	Ensemble #2 (Bagged DT)	Bagging (base model: Decision Tree)
6	LogReg Backward	penalty=l2, C=1.0
7	KNN	n_neighbors=4
8	SVM	C=10, kernel=rbf
9	LogReg Forward	penalty=l2, C=1.0
10	Naive Bayes	Default parameters (GaussianNB)
11	Decision Tree	max_depth=10, min_samples_split=10

Classification Models Comparison



Clustering Models Interpretation:



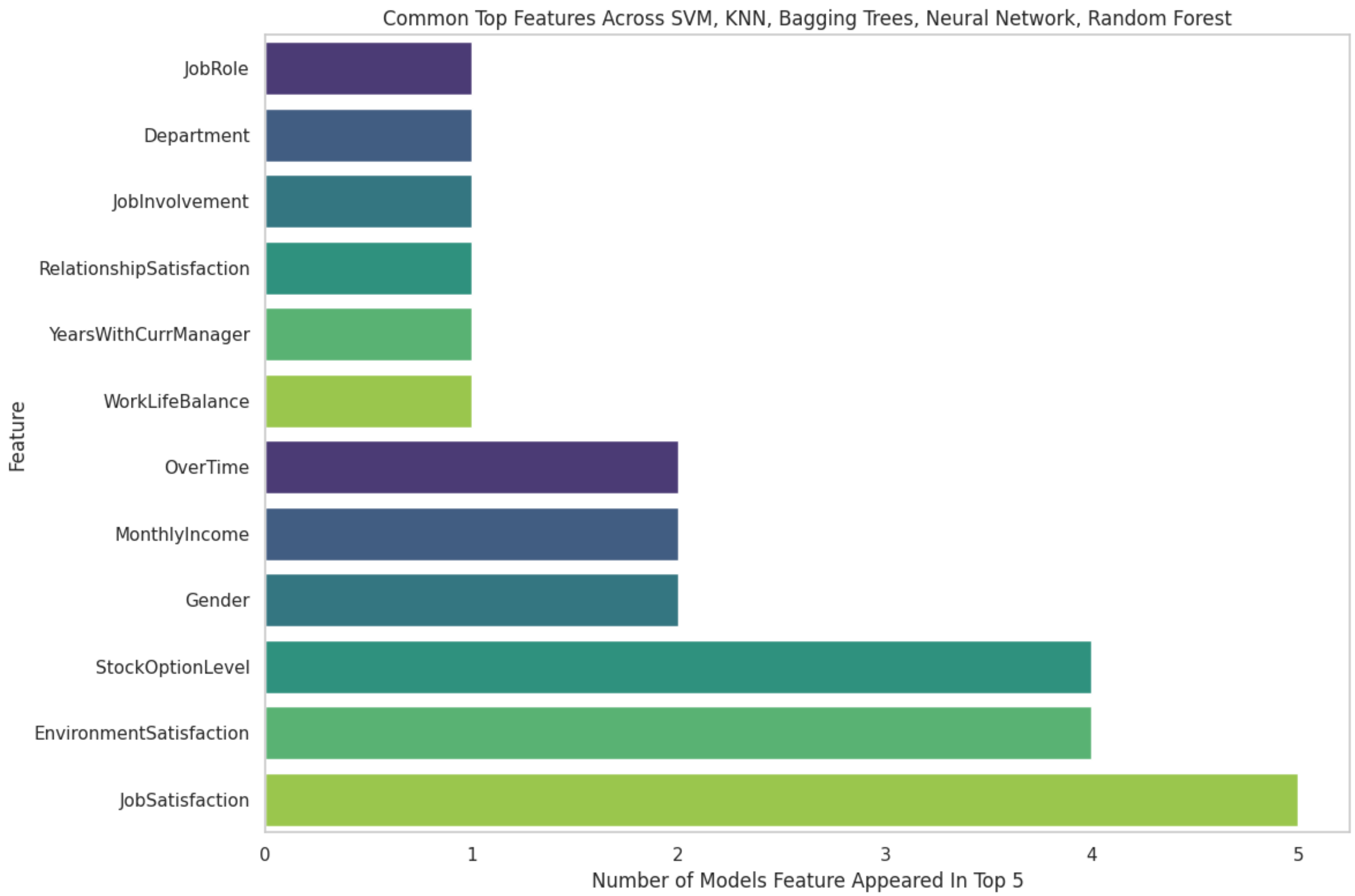
Aspect	KMeans	Agglomerative Clustering
Silhouette Score	0.1023	0.0726
Cluster Separation (PCA)	Better but still overlaps	Weaker, less clear
Interpretability	Good	Good
Best k identified?	Tentatively, k=2 or 3	Same

- K-Means performed slightly better than Agglomerative Clustering on this dataset.

Models and Their AUC, Recall and Accuracy

Model	Accuracy	Recall	AUC
Gradient Boost	0.900	0.919	0.970
Extreme Gradient Boost	0.900	0.902	0.970
SVM	0.896	0.903	0.959
Random Forest	0.889	0.883	0.961
Ensemble #1 (SVM+DT+NN)	0.883	0.895	0.954
Ensemble #2 (Bagged DT)	0.862	0.866	0.934
Neural Network	0.858	0.883	0.941
Logistic Regression (Backward)	0.806	0.802	0.882
KNN	0.826	0.923	0.898
Decision Tree	0.802	0.862	0.815
Naïve Bayes	0.678	0.802	0.788

Top features and top model in predicting attrition



Top Performing Model:

Gradient Boost

- Accuracy: 90.0%
- Recall: 91.9%
- AUC: 0.970

Conclusion:

Attrition Risk Strategy

- Our predictive model identified job satisfaction, compensation, and manager relationships as the most critical drivers of attrition risk.
- Targeted action on these factors can significantly strengthen employee retention, reduce turnover costs, and enhance organizational performance.