MODEL OF CHOICE

COMPARISON ON ROC CURVES FOR ALL CLASSIFICATION MODELS -

Receiver operating characteristic (ROC) curves are an effective tool for assessing and contrasting the effectiveness of various models. A plot of the true positive rate (TPR) versus the false positive rate (FPR) for various categorization criteria is known as a ROC curve. The TPR measures the percentage of true positives (positives that were correctly detected) among all actual positives, whereas the FPR measures the percentage of false positives (positives that were mistakenly identified) among all actual negatives.

Below is the plot for combined ROC curves based on the validation set. This plot shows that models XGBoost, Adaboost and Decision tree have the highest AUC score of 0.84, followed by Random Forest and SVM with AUC score of 0.80. It is known that the model with highest AUC score have the best performance because the it achieves True Positive Rate (TPR) for a given False Positive Rate (FPR).

From this comparison, XGBoost is the best model because the TRP is the highest for XGBoost and the FPR then increases with an increase in TPR.

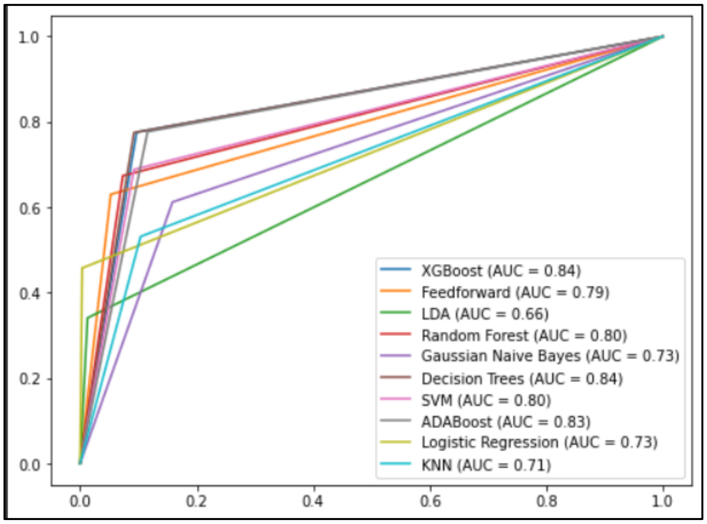

Figure 25: Combined ROC curve

COMPARISON ON ACCURACY FOR ALL CLASSIFICATION MODELS -

Model	Accuracy	F1 Score		Precision		Recall		AUC
Logistic	0.85	0	0.91	0	0.83	0	1.00	0.93
		1	0.62	1	0.98	1	0.46	
KNN	0.80	0	0.86	0	0.83	0	0.90	0.84
		1	0.59	1	0.66	1	0.53	
Naïve Bayes	0.78	0	0.85	0	0.85	0	0.84	0.84
		1	0.60	1	0.59	1	0.61	
SVM	0.85	0	0.90	0	0.89	0	0.91	0.90
		1	0.71	1	0.74	1	0.69	
LDA	0.81	0	0.88	0	0.80	0	0.99	0.88
		1	0.50	1	0.91	1	0.34	
Decision Tree	0.86	0	0.91	0	0.87	0	0.95	0.92
		1	0.71	1	0.82	1	0.63	
Random Forest	0.86	0	0.90	0	0.88	0	0.93	0.92
		1	0.72	1	0.78	1	0.67	
XGBoost	0.87	0	0.91	0	0.91	0	0.9	0.94
		1	0.76	1	0.75	1	0.78	
ADABoost	0.85	0	0.9	0	0.91	0	0.88	0.94
		1	0.75	1	0.72	1	0.78	
Feedforward Neural Networks	0.85	0	0.90	0	0.88	0	0.92	0.79
		1	0.71	1	0.77	1	0.66	

Table 24: Classification Report Summary

RECOMMENDED MODEL: XGBOOST

As per the combined classification report and the combined ROC curve given above, the XGBoost model has produced the best performance out of the 10 models tested. Thus, making it our recommended model. Summarizing its performance on all the features, it has produced an accuracy of 0.87, a precision of 0.91 for the Active (0) class and 0.75 for the Terminated (1) class, a recall of 0.90 for the Active (0) class and 0.78 for the Terminated (1) class, and an F1-score of 0.91 for the Active (0) class and 0.76 for the Terminated (1) class. The model has achieved an AUC of 0.94, indicating a strong performance in predicting the target class.

XGBOOST USING TOP 10 FEATURES –

Furthermore, by analyzing the feature importance, we have identified the top 10 most important features for the XGBoost model, which have contributed significantly to its predictive power. The features and their importance are listed below:

Feature	Importance
Recent Promotion to Termination Years	0.4298
Recent Promotion to Hire Years	0.3292
Years of Service	0.0604
Compa Ratio	0.0477
Age (Term by Term Date, Active by Today)	0.0388
Base Pay Mid-Point Annualized USD	0.0292
Generation_Gen Z	0.0116
Marital Status_Single	0.0085
Cost to Replace Employee Multiplier	0.0085
Termination Reason_Change of Career Direction	0.0083

Table 25: Top ten important features

Performance Evaluation:

After retraining the model on the top 10 features to further enhance its performance, it was found that the model achieved an accuracy of 0.87, a precision of 0.92 for the Active (0) class and 0.75 for the Terminated (1) class, a recall of 0.90 for the Active (0) class and 0.78 for the Terminated (1) class, and an F1-score of 0.91 for the Active (0) class and 0.77 for the Terminated (1) class. The model has achieved an AUC of 0.95 on these top 10 features, indicating that they are highly informative in predicting the target class.

Overall, the model's accuracy means that it correctly predicted the outcome for 87% of the employees. The model's ability to predict both classes (active and voluntarily terminated) is measured by the macro and weighted average of the precision, recall, and F1-score, which are all above 0.75 for both classes. Additionally, the AUC indicates that the model has good discriminative ability and can distinguish between the two classes effectively.

Based on these results, we recommend using the XGBoost model for predicting the target class, especially when using the top 10 most important features identified by the model.

	Precision	Recall	F1-Score	Support
Active (0)	0.92	0.90	0.91	2,360
Terminated (1)	0.75	0.78	0.77	892
Accuracy	0.87			
Macro avg	0.84	0.84	0.84	3,252
Weighted avg	0.87	0.87	0.87	3,252

Table 26: Classification Report (XGB-Top 10)

Confusion Matrix:

In the confusion matrix provided, the model has correctly predicted 2131 employees as active and 699 employees as likely to voluntarily terminate. On the other hand, the model has incorrectly predicted 193 active employees as likely to terminate and 229 employees likely to terminate as active. These incorrect predictions are known as false positives and false negatives, respectively.

The true positives (699) represent the number of employees who were predicted to voluntarily terminate and actually did, while the true negatives (2131) represent the number of employees who were predicted to remain active and actually did. The false positives (229) indicate the number of employees who were predicted to voluntarily terminate but actually remained active, while the false negatives (193) indicate the number of employees who were predicted to remain active but actually voluntarily terminated.

Confusion Matrix			
Predicted	Actual		
		Active (0)	Voluntary Termination (1)
	Active (0)	True Negatives 2131	False Positives 229
	Voluntary Termination (1)	False Negatives 193	True Positives 699

Table 27: Confusion Matrix (XGB-Top 10)

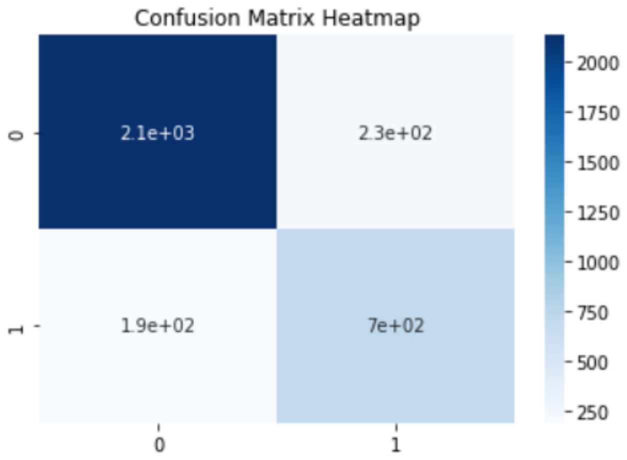

Figure 26: Confusion Matrix heatmap (XBG-Top 10)

ROC Curve:

The ROC curve edging further towards the corner indicates the better performance of the model yet again. The model has a better ability to differentiate between the two classes, which is indicated by the high AUC value. In this case, an AUC of 0.95 is quite high, indicating that the model has strong predictive performance.

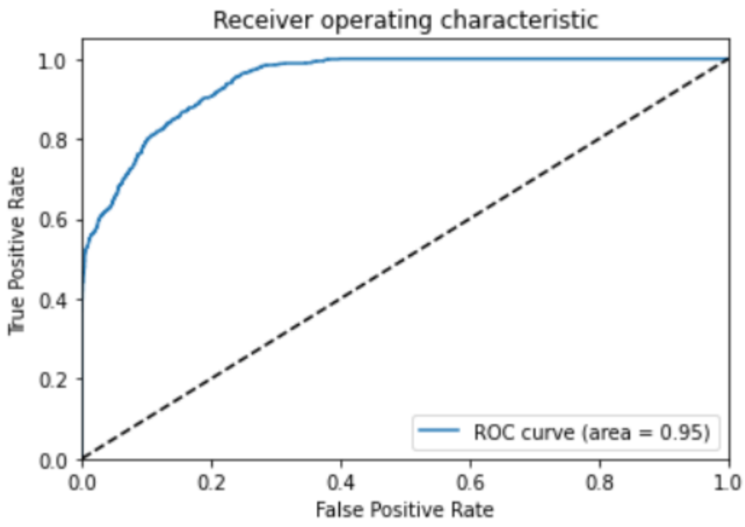

Figure 27: ROC Curve (XBG-Top 10)

Learning Curve:

The learning curve below demonstrates the training score and the cross-validation score of the model. As per the plot, the converging training and cross-validation scores indicate that the model is not overfitting or underfitting.

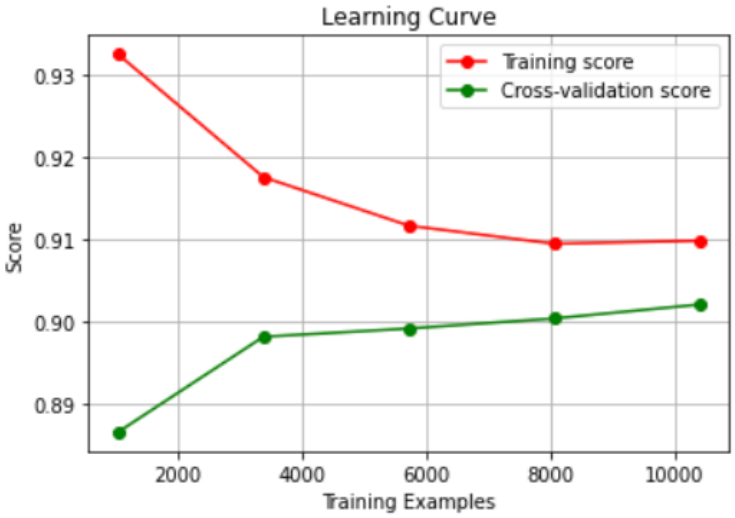

Figure 28: Learning Curve (XBG-Top 10)

Based on the results of this model, it appears that the top 10 features identified are strong predictors of employee turnover, and the model has good performance in predicting the likelihood of voluntary termination. However, it is important to note that no model can predict the future with certainty, and other factors not captured in the model may impact employee turnover.