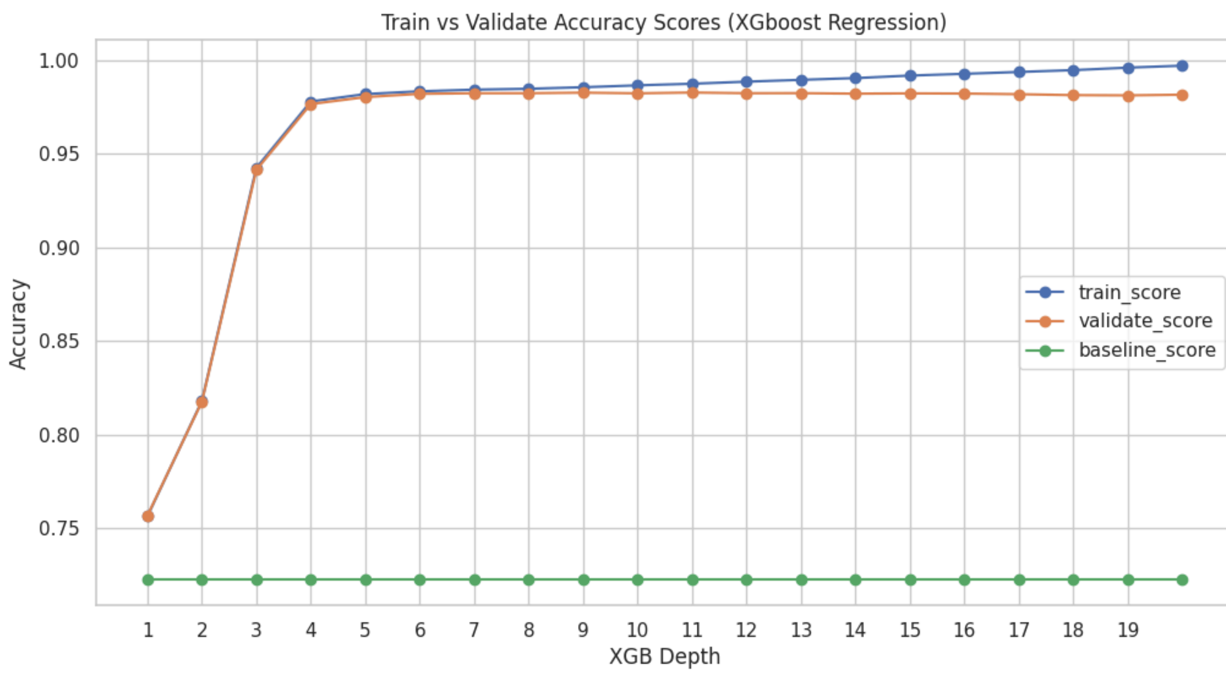
Data Visualizations

For the next phase of our analysis, we involved evaluating four classification models: XGboost, Logistic Regression, Random Forest, and Decision Tree.

XGboost:

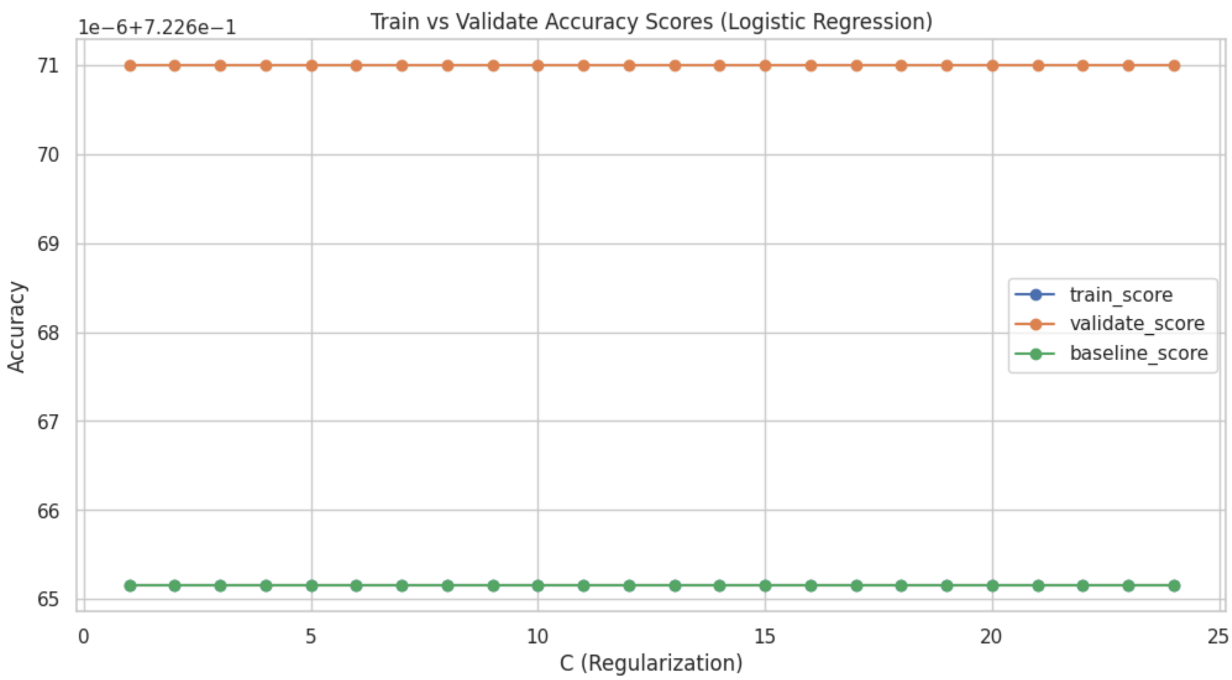
For this model, validation accuracy increased sharply between depths 1 and 4 and stabilized around 0.9833 while validation accuracy was 0.9820, compared to a baseline of accuracy of 0.7227.

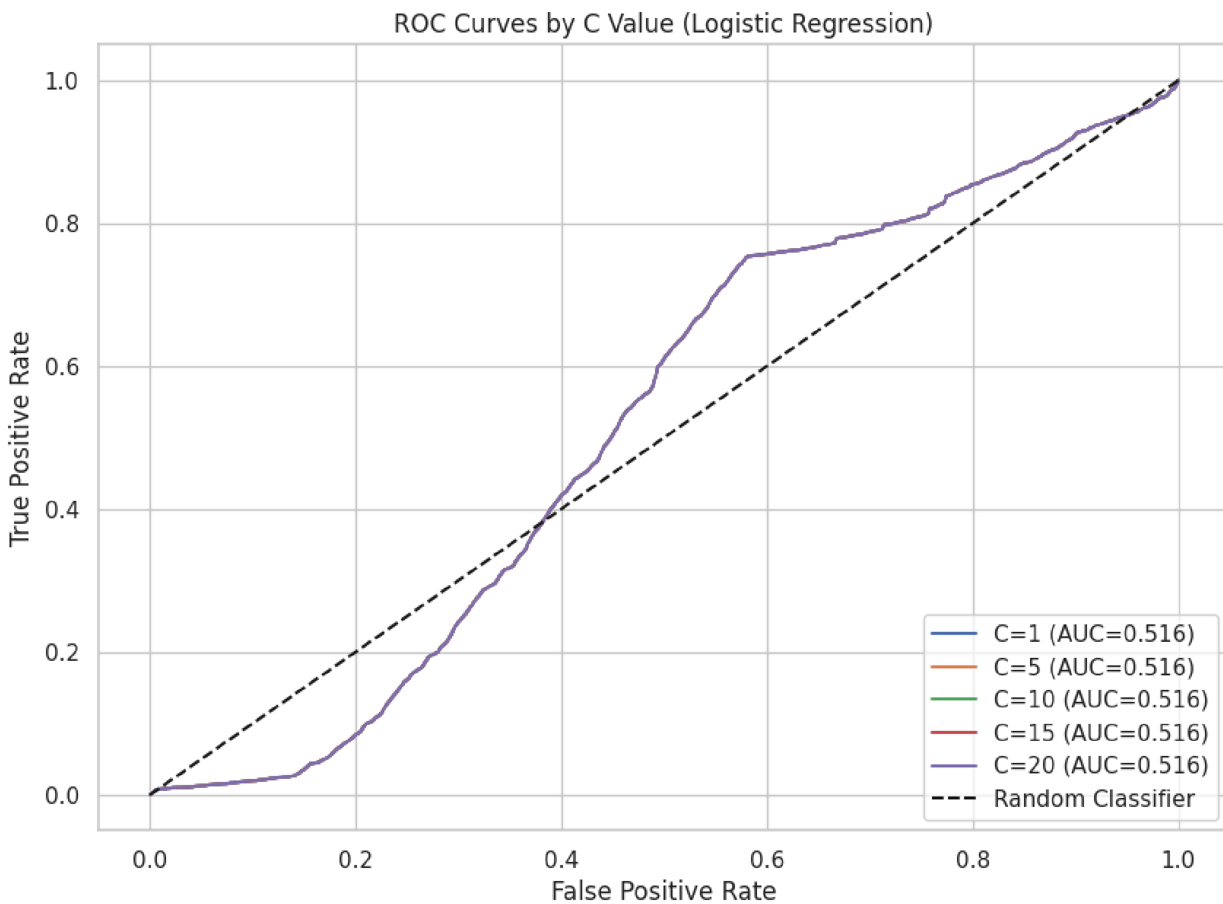
The gap between training and validation accuracy remained minimal across depths which suggests strong generalization and limited overfitting.

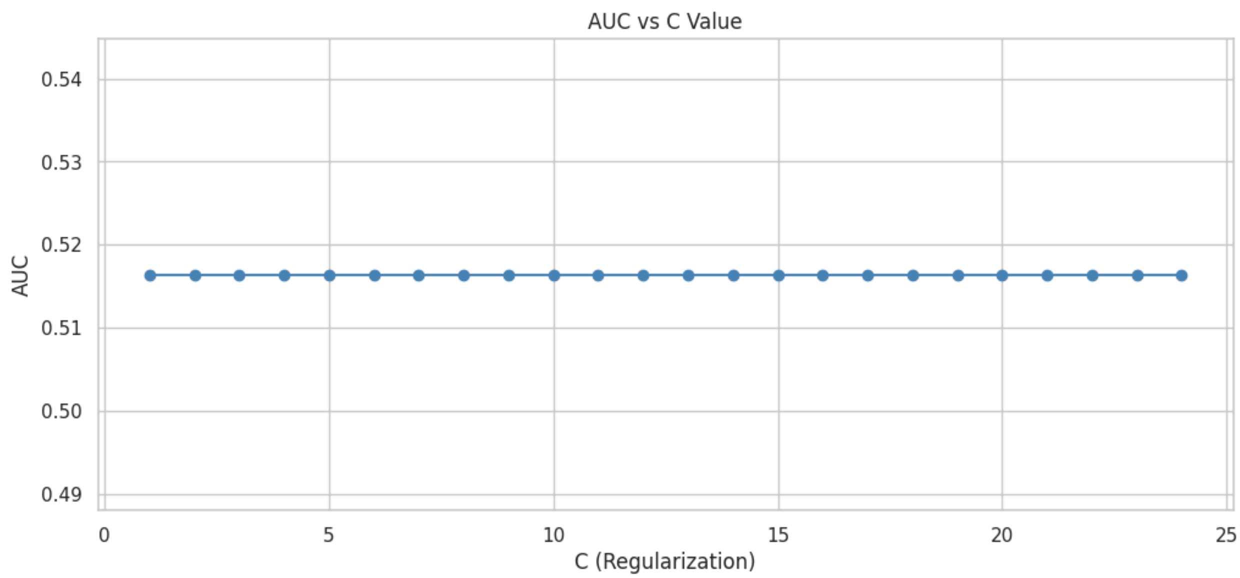


Logistic Regression:

For this model, training and validation accuracy remained stable at approximately 0.71 across all C values tested with an AUC of 0.516 throughout. The consistently low AUC indicates that despite reasonable accuracy, the model performs barely better than random chance at ranking predictions. This suggests that a linear decision boundary is insufficient to capture the nonlinear relationships present in the dataset.

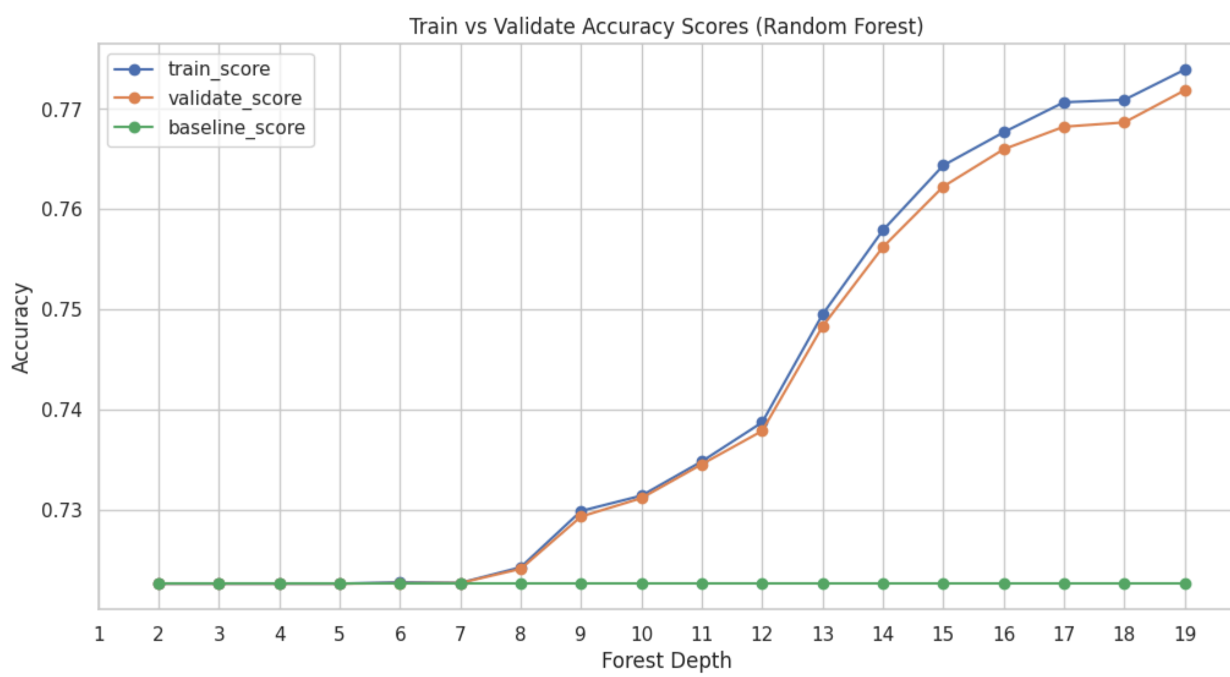

Train vs Validate Accuracy Scores (Logistic Regression)
1e−6+7.226e−1
71
70
69
68
67
66
65
Accuracy
0
5
10
15
20
25
C (Regularization)
train_score
validate_score
baseline_score


ROC Curves by C Value (Logistic Regression)
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
C=1 (AUC=0.516)
C=5 (AUC=0.516)
C=10 (AUC=0.516)
C=15 (AUC=0.516)
C=20 (AUC=0.516)
Random Classifier



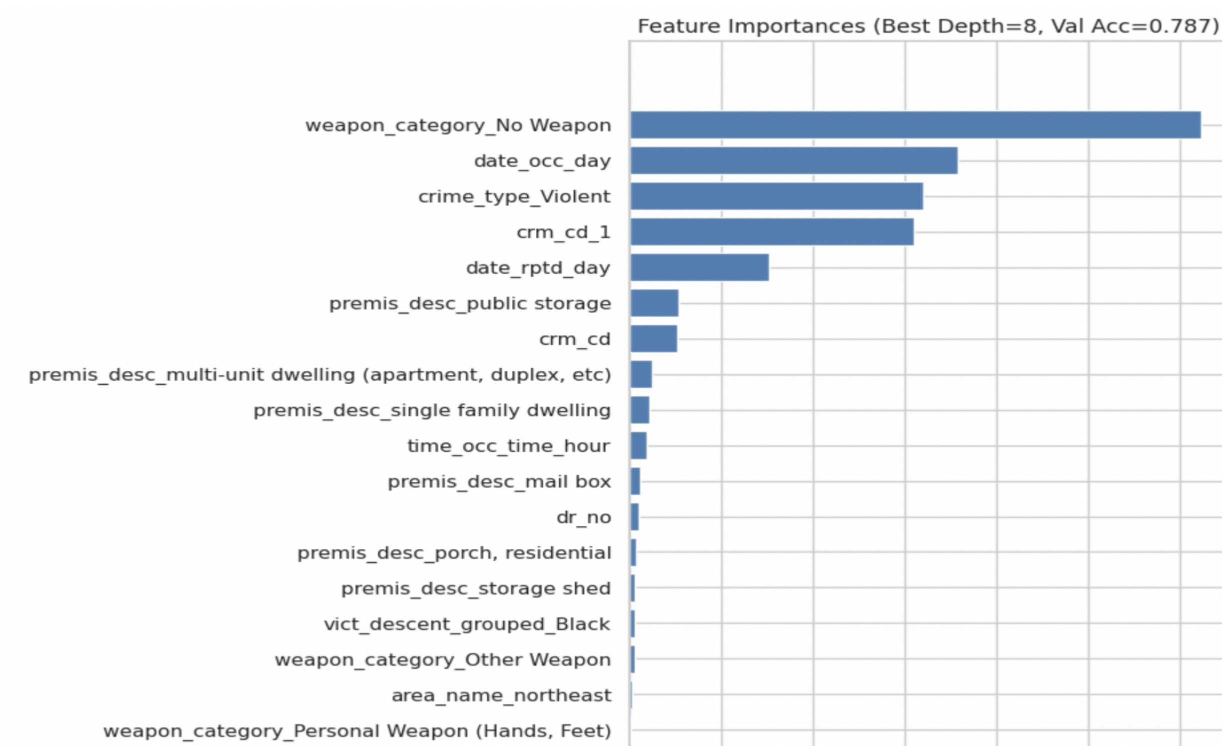
Random Forest:

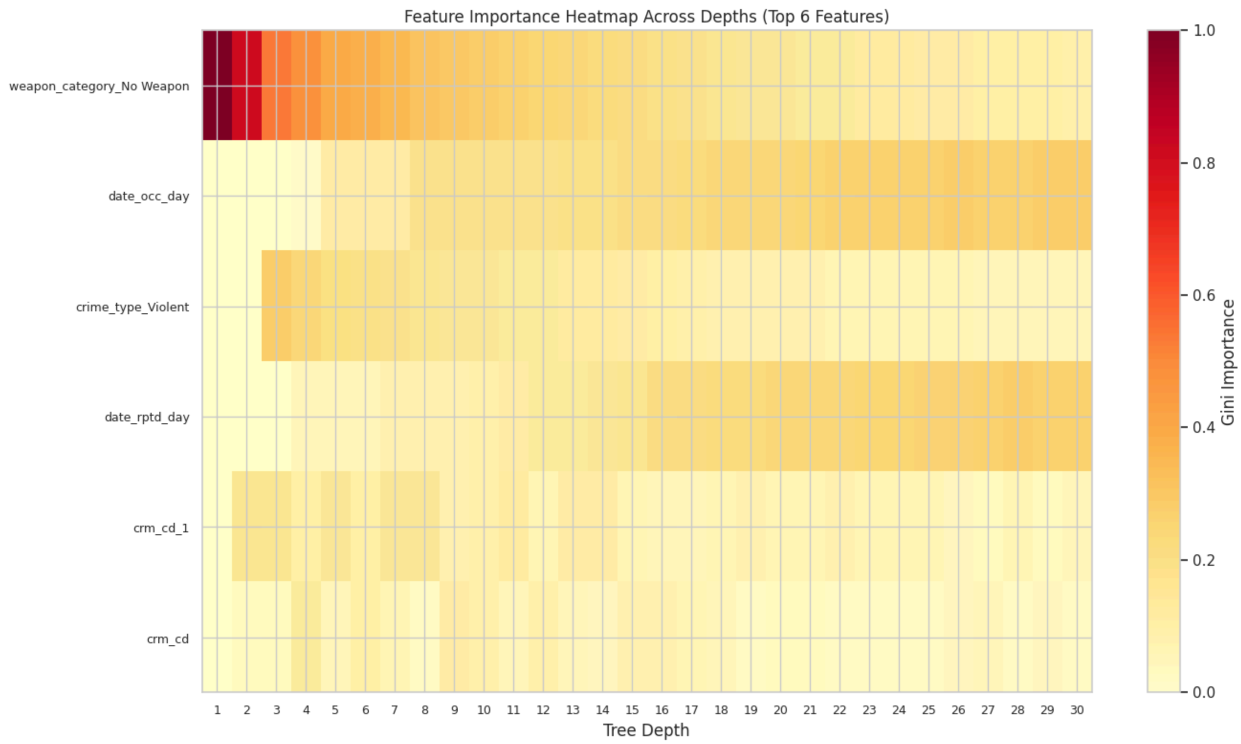
This model demonstrated gradual improvement in validation accuracy as forest depth increased. Validation accuracy improved from about 0.72 to about 0.77 at higher depths. Training accuracy increased at a similar rate with a small gap between training and validation results.



Decision Tree:

For this model, validation accuracy increased steadily as depth increased, reaching about 0.86-0.87 at higher depths. Though, training accuracy increased more rapidly, exceeding 0.95 at deeper levels. The widening gap between training and validation accuracy suggests increasing overfitting as tree depths grow. The model lacks stability, even though it captures a nonlinear structure effectively.





Post Analysis:

The XGBoost model performed very well, reaching an overall accuracy of 98% on the test set. This is a 35.5% improvement over the baseline accuracy of 72.27%. The classification report shows that XGBoost worked well for both classes. For low reporting delay (class 0), recall was perfect at 100%, and precision was high at 98%. For high reporting delay (class 1), precision was 99%, and recall was 94%. Both macro average and weighted average F1-scores were 98%, indicating the model effectively managed the class imbalance.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	64145
1	0.99	0.94	0.97	24616
accuracy			0.98	88761
macro avg	0.99	0.97	0.98	88761
weighted avg	0.98	0.98	0.98	88761

The confusion matrix supports these findings, with 64,025 true negatives and 23,141 true positives, and very few misclassifications. The model had only 120 false positives (0.19% error rate for low reporting delay) and 1,475 false negatives (6% error rate for high reporting delay).

The low false-positive rate and strong true-positive results indicate that XGBoost learned to distinguish between the classes effectively. This makes it a strong choice for deployment, as it generalizes well and avoids overfitting.