**Phishing Detection Report Using SMOTIFIED-GAN & 10 Features**

**1. Dataset Overview**

This report consolidates the evaluation results of phishing website detection using a SMOTIFIED-GAN approach across three datasets with varying class imbalances: 80/20, 90/10, and 95/5 (Legitimate/Phishing). Five machine learning models—Logistic Regression, Random Forest, XGBoost, CatBoost, and Stacking Ensemble—were evaluated after balancing the datasets to a 50/50 class distribution using SMOTIFIED-GAN. The top 10 most informative features were selected using SelectKBest for all datasets. Performance metrics include Accuracy, Precision, Recall, F1 Score, and ROC-AUC across cross-validation, training, and test sets. This report also includes a comparison of ROC curves and confusion matrices for the test sets to provide visual insights into model performance.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Initial Class Distribution | Post-SMOTIFIED-GAN Distribution | Feature Selection |
| 80/20 | 80% Legitimate, 20% Phishing | 50% Legitimate, 50% Phishing | Top 10 (SelectKBest) |
| 90/10 | 90% Legitimate, 10% Phishing | 50% Legitimate, 50% Phishing | Top 10 (SelectKBest) |
| 95/5 | 95% Legitimate, 5% Phishing | 50% Legitimate, 50% Phishing | Top 10 (SelectKBest) |

**2. Cross-Validation Scores**

The following table compares cross-validation performance across the three datasets.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | 80/20 Accuracy | 80/20 F1 Score | 80/20 ROC-AUC | 90/10 Accuracy | 90/10 F1 Score | 90/10 ROC-AUC | 95/5 Accuracy | 95/5 F1 Score | 95/5 ROC-AUC |
| Logistic Regression | 0.9020 | 0.9100 | 0.9068 | 0.9015 | 0.9092 | 0.9816 | 0.9044 | 0.9111 | 0.9900 |
| Random Forest | 0.9126 | 0.9167 | 0.9462 | 0.9460 | 0.9442 | 0.9892 | 0.9363 | 0.9388 | 0.9659 |
| XGBoost | 0.9125 | 0.9169 | 0.9467 | 0.9468 | 0.9451 | 0.9904 | 0.9361 | 0.9387 | 0.9667 |
| CatBoost | 0.9127 | 0.9169 | 0.9473 | 0.9470 | 0.9453 | 0.9904 | 0.9370 | 0.9395 | 0.9657 |
| Stacking Ensemble | 0.9131 | 0.9171 | 0.9191 | 0.9467 | 0.9451 | 0.9921 | 0.9358 | 0.9381 | 0.9939 |

**Observations**

* **Accuracy**: The 90/10 dataset shows higher accuracy for Random Forest, XGBoost, CatBoost, and Stacking Ensemble compared to 80/20 and 95/5. Logistic Regression remains consistent (~0.9015–0.9044).
* **F1 Score**: The 95/5 dataset yields the highest F1 scores for Logistic Regression (0.9111) and CatBoost (0.9395). The 90/10 dataset performs best for Random Forest, XGBoost, and Stacking Ensemble (~0.945).
* **ROC-AUC**: The 90/10 and 95/5 datasets achieve higher ROC-AUC, with Stacking Ensemble leading (0.9921–0.9939), indicating strong discriminative ability.

**3. Training Set Scores**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | 80/20 Accuracy | 80/20 Precision | 80/20 Recall | 80/20 F1 Score | 80/20 ROC-AUC | 90/10 Accuracy | 90/10 Precision | 90/10 Recall | 90/10 F1 Score | 90/10 ROC-AUC | 95/5 Accuracy | 95/5 Precision | 95/5 Recall | 95/5 F1 Score | 95/5 ROC-AUC |
| Logistic Regression | 0.9019 | 0.8405 | 0.9921 | 0.9100 | 0.9019 | 0.9016 | 0.8436 | 0.9860 | 0.9093 | 0.9016 | 0.9043 | 0.8515 | 0.9794 | 0.9109 | 0.9043 |
| Random Forest | 0.9230 | 0.8863 | 0.9705 | 0.9265 | 0.9230 | 0.9520 | 0.9832 | 0.9197 | 0.9504 | 0.9520 | 0.9418 | 0.9074 | 0.9841 | 0.9442 | 0.9418 |
| XGBoost | 0.9174 | 0.8800 | 0.9666 | 0.9213 | 0.9174 | 0.9500 | 0.9792 | 0.9196 | 0.9484 | 0.9500 | 0.9378 | 0.9037 | 0.9800 | 0.9403 | 0.9378 |
| CatBoost | 0.9184 | 0.8806 | 0.9681 | 0.9223 | 0.9184 | 0.9506 | 0.9811 | 0.9189 | 0.9490 | 0.9506 | 0.9399 | 0.9058 | 0.9819 | 0.9423 | 0.9399 |
| Stacking Ensemble | 0.9208 | 0.8830 | 0.9700 | 0.9245 | 0.9208 | 0.9502 | 0.9766 | 0.9226 | 0.9488 | 0.9502 | 0.9387 | 0.9082 | 0.9762 | 0.9410 | 0.9387 |

**Observations**

* **Accuracy**: The 90/10 dataset outperforms others for Random Forest, XGBoost, CatBoost, and Stacking Ensemble (~0.950–0.952).
* **Precision**: The 90/10 dataset achieves the highest precision for Random Forest, XGBoost, and CatBoost (~0.979–0.983).
* **Recall**: Logistic Regression consistently achieves the highest recall (0.9794–0.9921).
* **F1 Score and ROC-AUC**: The 90/10 dataset shows the highest F1 scores and ROC-AUC for ensemble-based models.

**4. Test Set Scores**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | 80/20 Accuracy | 80/20 Precision | 80/20 Recall | 80/20 F1 Score | 80/20 ROC-AUC | 90/10 Accuracy | 90/10 Precision | 90/10 Recall | 90/10 F1 Score | 90/10 ROC-AUC | 95/5 Accuracy | 95/5 Precision | 95/5 Recall | 95/5 F1 Score | 95/5 ROC-AUC |
| Logistic Regression | 0.8385 | 0.5556 | 0.9625 | 0.7045 | 0.8850 | 0.8195 | 0.3438 | 0.8857 | 0.4953 | 0.8489 | 0.8155 | 0.1620 | 0.6441 | 0.2588 | 0.7343 |
| Random Forest | 0.8571 | 0.6027 | 0.8368 | 0.7007 | 0.8495 | 0.9389 | 0.7525 | 0.5789 | 0.6544 | 0.7789 | 0.8729 | 0.2170 | 0.5906 | 0.3174 | 0.7392 |
| XGBoost | 0.8579 | 0.6033 | 0.8457 | 0.7042 | 0.8534 | 0.9397 | 0.7454 | 0.6027 | 0.6665 | 0.7899 | 0.8736 | 0.2155 | 0.5786 | 0.3141 | 0.7339 |
| CatBoost | 0.8577 | 0.6027 | 0.8457 | 0.7038 | 0.8532 | 0.9395 | 0.7513 | 0.5908 | 0.6615 | 0.7845 | 0.8753 | 0.2237 | 0.6048 | 0.3267 | 0.7472 |
| Stacking Ensemble | 0.8581 | 0.6035 | 0.8462 | 0.7046 | 0.8536 | 0.9385 | 0.7224 | 0.6260 | 0.6707 | 0.7996 | 0.8723 | 0.1930 | 0.4880 | 0.2766 | 0.6903 |

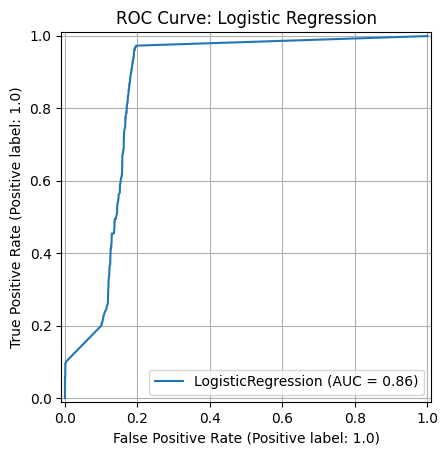
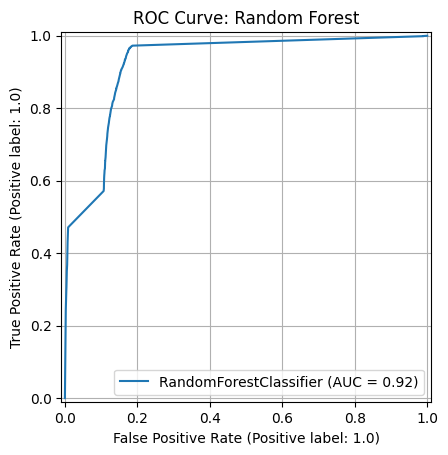
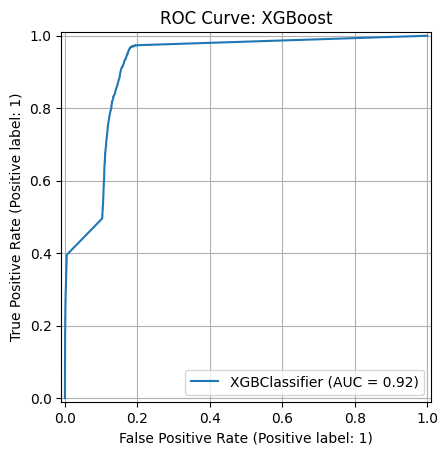
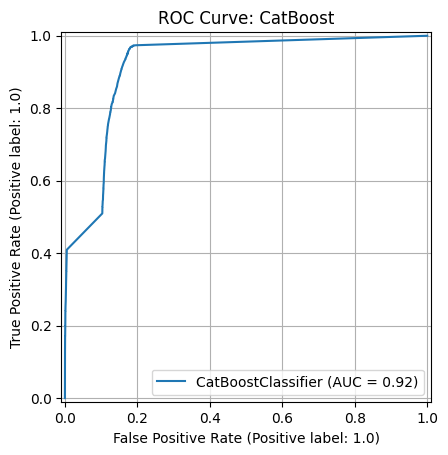
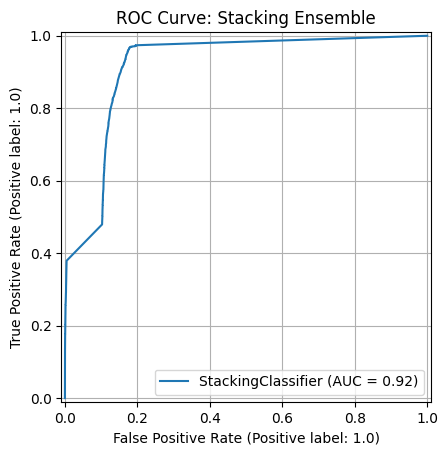
**Observations**

* **Accuracy**: The 90/10 dataset achieves the highest accuracy (0.9385–0.9397) for ensemble-based models, while 95/5 outperforms 80/20 slightly for Random Forest, XGBoost, and CatBoost.
* **Precision**: Precision decreases with increasing imbalance (80/20 > 90/10 > 95/5), with 80/20 showing the highest (0.5556–0.6035) and 95/5 the lowest (0.1620–0.2237).
* **Recall**: Logistic Regression maintains high recall (0.6441–0.9625), but other models see reduced recall in 90/10 and 95/5, with Stacking Ensemble dropping to 0.4880 in 95/5.
* **F1 Score**: The 80/20 dataset yields the highest F1 scores (0.7007–0.7046), while 95/5 has the lowest (0.2588–0.3267).
* **ROC-AUC**: The 80/20 dataset achieves the highest ROC-AUC (0.8495–0.8850), followed by 90/10 (0.7789–0.8489) and 95/5 (0.6903–0.7472).

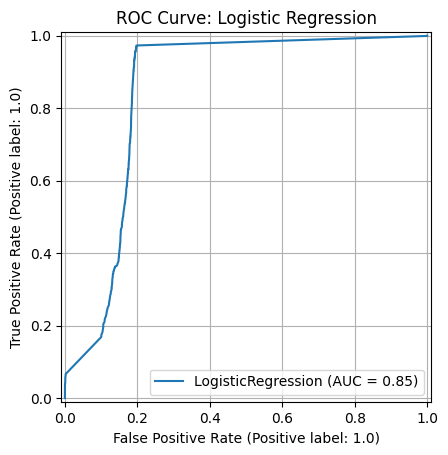
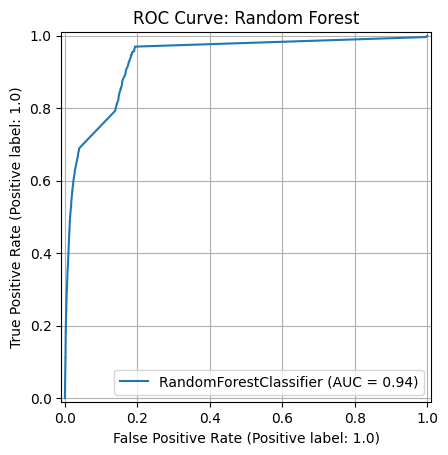
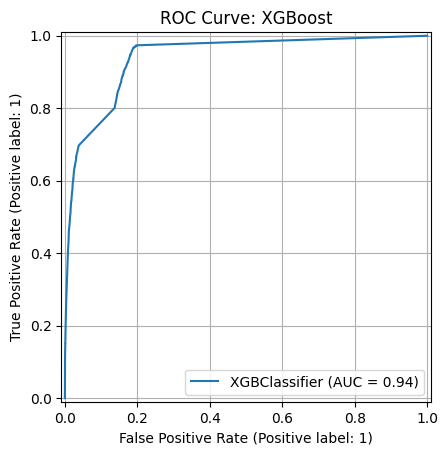
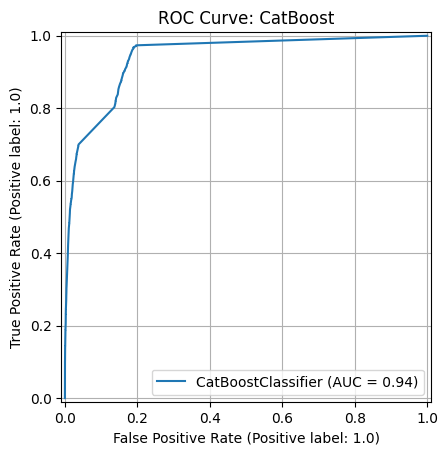
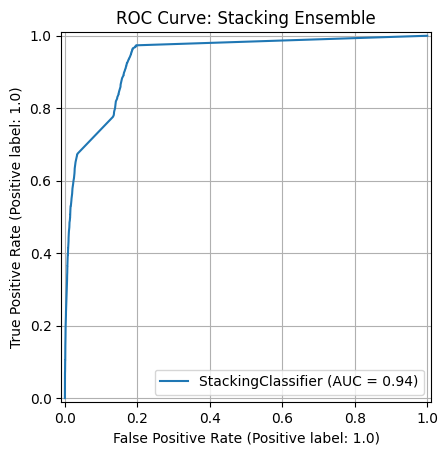
**5. ROC Curve Comparison**

The ROC curves for each model on the test sets of the three datasets are referenced in the original reports as follows:

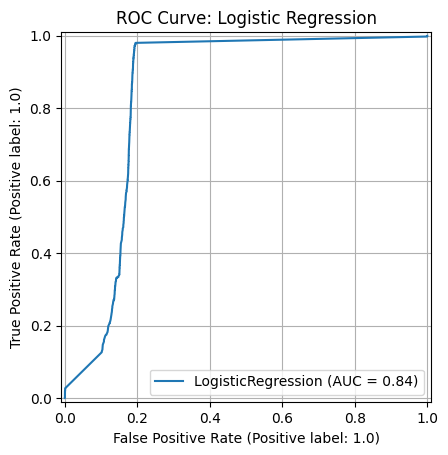
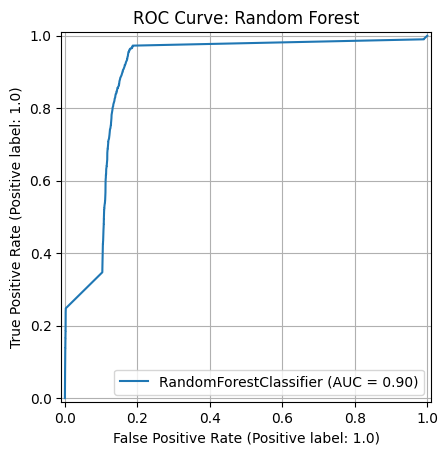
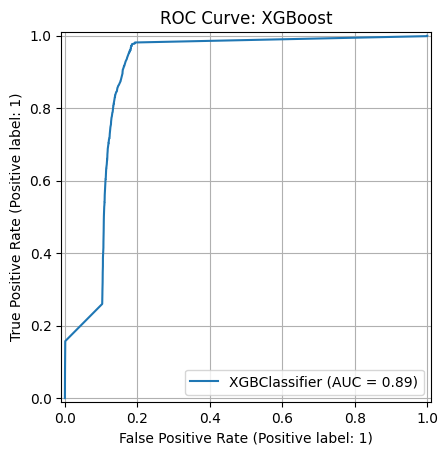
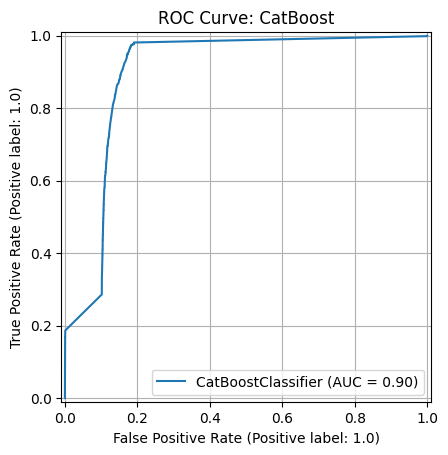
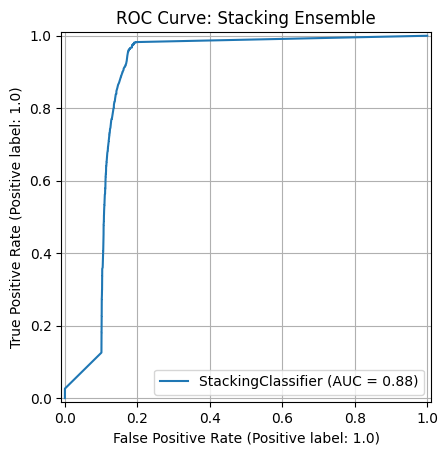
* **80/20 Dataset**:



* **90/10 Dataset**:



* **95/5 Dataset**:

****

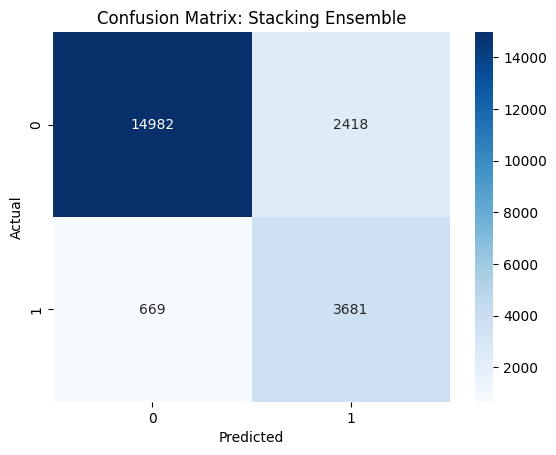
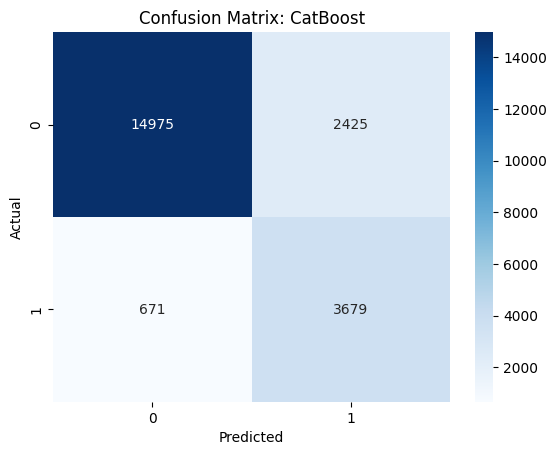
**Qualitative Comparison**

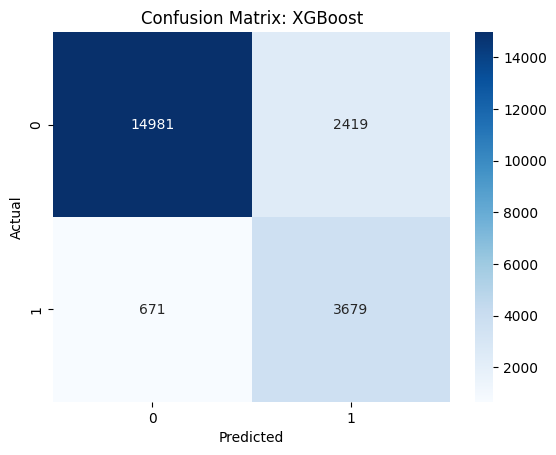
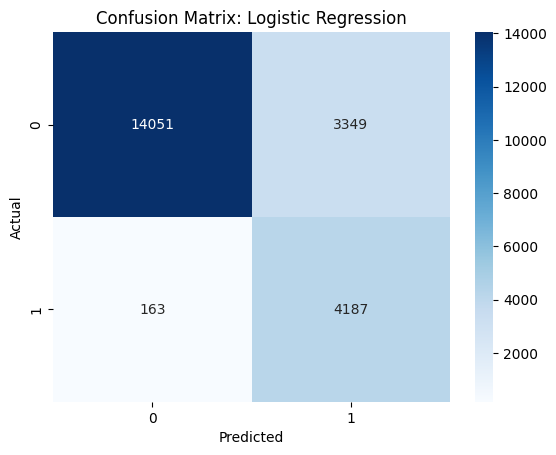
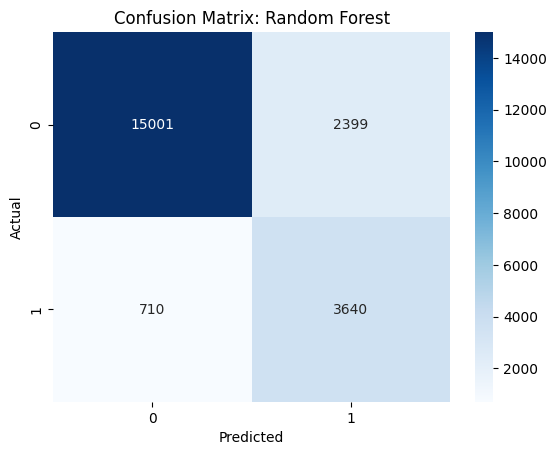
* **80/20 Dataset**:
  + **Logistic Regression**: High ROC-AUC (0.8850) suggests a steep curve with good separation, reflecting high recall (0.9625).
  + **Random Forest, XGBoost, CatBoost, Stacking Ensemble**: Similar ROC-AUC values (~0.8495–0.8536) indicate closely clustered curves with strong discriminative ability.
  + **Observation**: Best balance between true positive rate (TPR) and false positive rate (FPR).
* **90/10 Dataset**:
  + **Logistic Regression**: ROC-AUC (0.8489) suggests a slightly flatter curve due to low precision (0.3438).
  + **Random Forest, XGBoost, CatBoost, Stacking Ensemble**: ROC-AUC values (0.7789–0.7996) indicate reduced TPR at higher FPRs due to lower recall (0.5789–0.6260).
  + **Observation**: Increased imbalance reduces ROC performance.
* **95/5 Dataset**:
  + **Logistic Regression**: ROC-AUC (0.7343) shows high TPR but poor precision (0.1620), leading to higher FPR.
  + **Random Forest, XGBoost, CatBoost**: ROC-AUC (~0.7339–0.7472) suggests moderate TPR and higher FPR.
  + **Stacking Ensemble**: Lowest ROC-AUC (0.6903) indicates the flattest curve.
  + **Observation**: Extreme imbalance results in the least favorable ROC curves.

**6. Confusion Matrix Comparison**

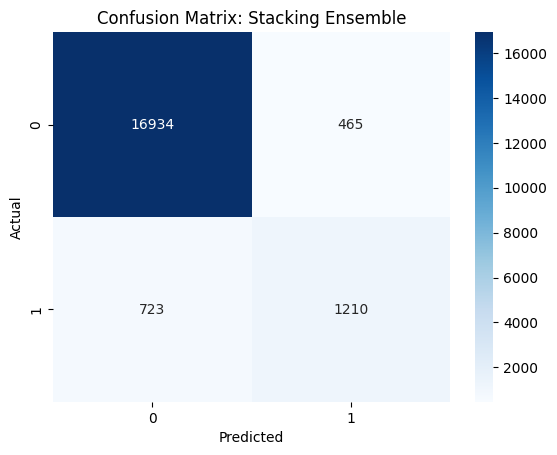
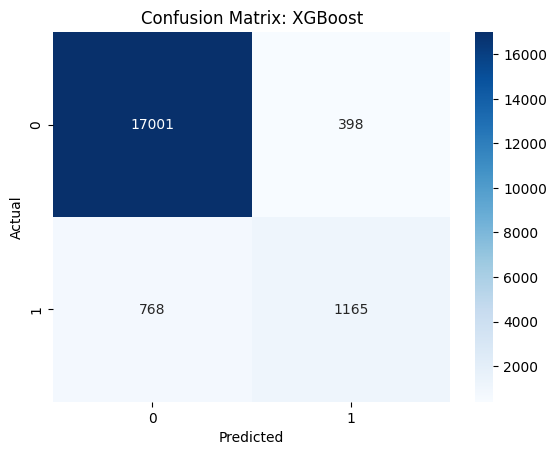
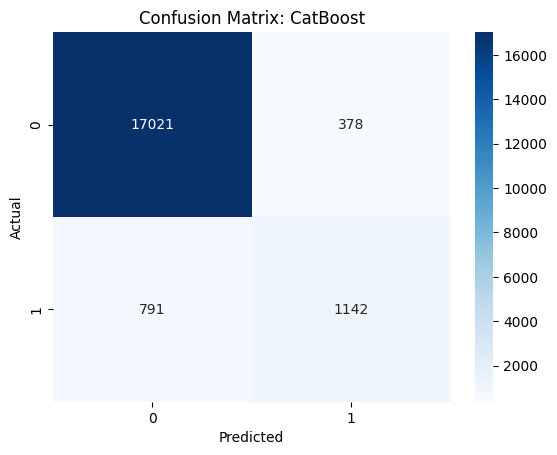
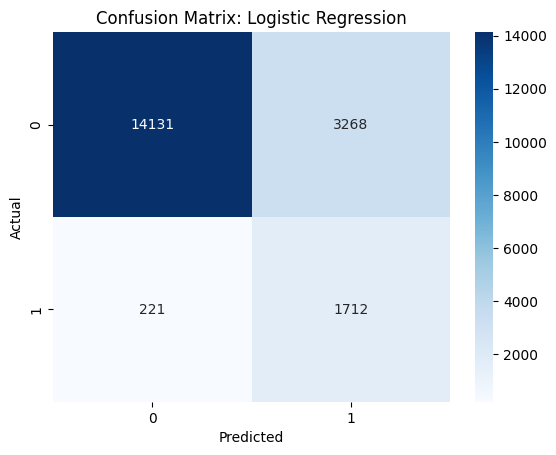
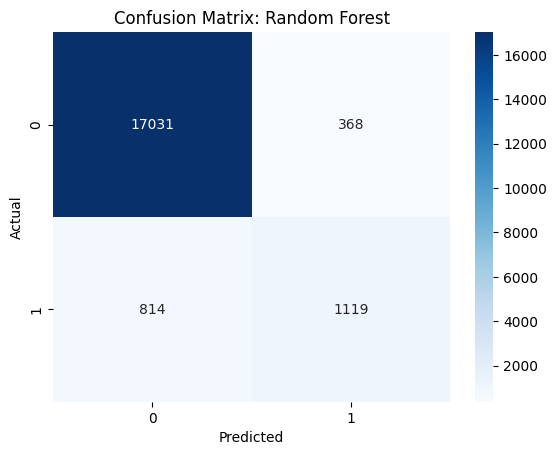
The confusion matrices for each model on the test sets are referenced in the original reports as follows:

* **80/20 Dataset**:

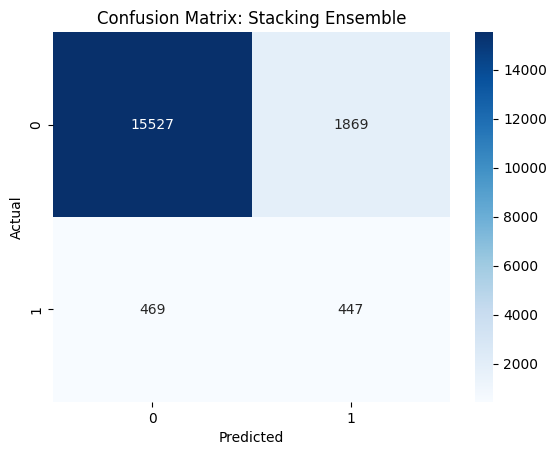
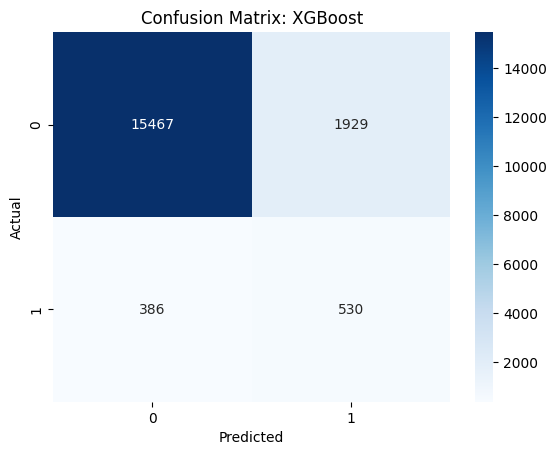
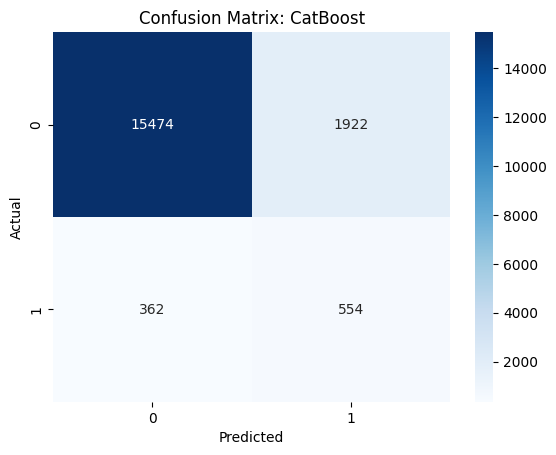
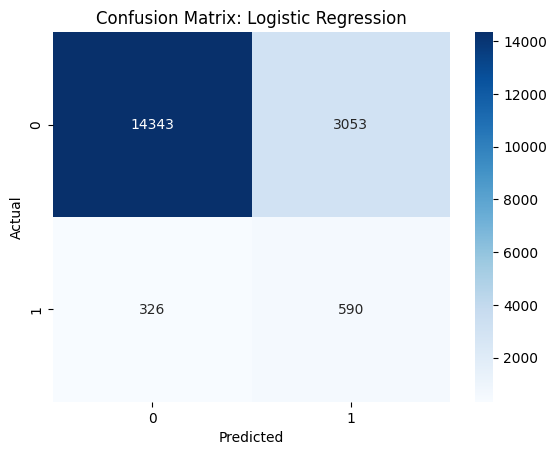
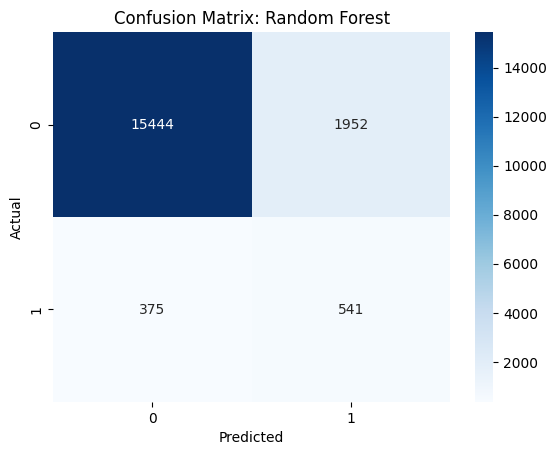
 

* **90/10 Dataset**:

* **95/5 Dataset**:

**Qualitative Comparison**

* **80/20 Dataset**:
  + **Logistic Regression**: High recall (0.9625) suggests high true positives (TP), but moderate precision (0.5556) indicates notable false positives (FP).
  + **Random Forest, XGBoost, CatBoost, Stacking Ensemble**: Balanced TP and FP, with recall (0.6027–0.6035).
  + **Observation**: Strong diagonal (high TP and true negatives, TN) with moderate FPs.
* **90/10 Dataset**:
  + **Logistic Regression**: High recall (0.8857) but low precision (0.3438) suggests many TPs but high FPs.
  + **Random Forest, XGBoost, CatBoost, Stacking Ensemble**: Higher precision (0.5789–0.6260) indicates fewer FPs but fewer TPs.
  + **Observation**: Strong TNs due to high accuracy (0.9385–0.9397), but reduced TPs.
* **95/5 Dataset**:
  + **Logistic Regression**: Moderate recall (0.6441) and low precision (0.1620) suggest high FPs.
  + **Random Forest, XGBoost, CatBoost**: Low precision (0.5786–0.6048) indicate low TPs and high FPs.
  + **Stacking Ensemble**: Lowest recall (0.4880) and precision (0.1930) suggest fewest TPs and high FPs.
  + **Observation**: Very weak TPs and high FPs due to extreme imbalance.

**7. Key Insights**

* **Effect of Class Imbalance**: Increasing imbalance (80/20 to 95/5) reduces test set precision and F1 scores, with 95/5 showing the poorest performance due to low TPs and high FPs in confusion matrices.
* **Model Performance**:
  + **Logistic Regression**: High recall but low precision in imbalanced settings, leading to high FPs and flatter ROC curves in 95/5.
  + **Random Forest, XGBoost, CatBoost**: Perform well in 80/20 and 90/10, with balanced confusion matrices and steeper ROC curves, but struggle in 95/5.
  + **Stacking Ensemble**: Robust across datasets but performs poorly in 95/5, with low TPs and high FPs.
* **SMOTIFIED-GAN Effectiveness**: Balances datasets effectively for training, but test set performance on imbalanced data highlights generalization challenges.
* **ROC and Confusion Matrix Insights**: The 80/20 dataset shows the best ROC curves and balanced confusion matrices, while 95/5 exhibits the weakest performance.

**8. Conclusion**

The SMOTIFIED-GAN approach enables robust training and cross-validation performance. The 80/20 dataset achieves the best test set performance (F1: 0.7007–0.7046, ROC-AUC: 0.8495–0.8850), with balanced confusion matrices and steep ROC curves. The 90/10 dataset excels in accuracy and precision for ensemble models, while the 95/5 dataset poses significant challenges, with low precision and F1 scores reflected in skewed confusion matrices and flatter ROC curves. The Stacking Ensemble and CatBoost are recommended for their robustness, though performance degrades in extreme imbalances like 95/5.