Phishing Detection Model Performance Report

May 31, 2025

Executive Summary

This report consolidates the evaluation results of phishing website detection using SMOTE and MCMC augmentation techniques across nine datasets with varying feature sets (top 10, 20, and 30 features) and train-test splits (80-20, 90-10, and 95-5). Five machine learning modelsLogistic Regression, Random Forest, XGBoost, CatBoost, and Stacking Ensemblewere evaluated after balancing the datasets. Performance metrics include Train Accuracy, Test Accuracy, Precision, Recall, F1 Score, ROC-AUC, and Runtime. The report includes tabular summaries, graphical comparisons, and a CSV export of results, providing a comprehensive analysis of model effectiveness.

1 Dataset Overview

The analysis was conducted on nine datasets stored in `/content/drive/MyDrive/dataset/`, with the following characteristics after augmentation to balance class distributions:

Dataset	Initial Class Distribution	Post-Augmentation Distribution	Feature Count
80/20 (top 10, 20, 30)	80% Legitimate, 20% Phishing	50% Legitimate, 50% Phishing	11, 21, 31
90/10 (top 10, 20, 30)	90% Legitimate, 10% Phishing	50% Legitimate, 50% Phishing	11, 21, 31
95/5 (top 10, 20, 30)	95% Legitimate, 5% Phishing	50% Legitimate, 50% Phishing	11, 21, 31

Table 1: Dataset Characteristics

- **Augmentation Techniques**: SMOTE and MCMC were used to balance the minority (Phishing) class with the majority (Legitimate) class.
- **Feature Selection**: Datasets were preprocessed with the top 10, 20, or 30 features, as indicated by their filenames.

2 Cross-Validation Scores

Cross-validation scores are not directly computed in the provided output, but training set performance can serve as a proxy for model generalization. The following table summarizes average training set performance across all datasets:

Observations:

Model	Train Accuracy	Train Precision	Train Recall	Train F1 Score	Train ROC-AUC
Logistic Regression	0.9015	0.8420	0.9836	0.9101	0.9015
Random Forest	0.9390	0.9256	0.9581	0.9404	0.9390
XGBoost	0.9351	0.9210	0.9554	0.9373	0.9351
CatBoost	0.9362	0.9225	0.9563	0.9385	0.9362
Stacking Ensemble	0.9366	0.9226	0.9563	0.9381	0.9366

Table 2: Average Training Set Performance

- Random Forest, XGBoost, CatBoost, and Stacking Ensemble show higher Train Accuracy (0.93510.9390) compared to Logistic Regression (0.9015).
- Recall is consistently high across all models (0.95540.9836), with Logistic Regression leading.

3 Training Set Scores

The following table presents the average training set performance across all datasets:

Model	Train Accuracy	Train Precision	Train Recall	Train F1 Score	Train ROC-AUC
Logistic Regression	0.9015	0.8420	0.9836	0.9101	0.9015
Random Forest	0.9390	0.9256	0.9581	0.9404	0.9390
XGBoost	0.9351	0.9210	0.9554	0.9373	0.9351
CatBoost	0.9362	0.9225	0.9563	0.9385	0.9362
Stacking Ensemble	0.9366	0.9226	0.9563	0.9381	0.9366

Table 3: Average Training Set Performance

Observations:

- Random Forest achieves the highest Train Accuracy (0.9390) and F1 Score (0.9404).
- Logistic Regression excels in Recall (0.9836) but has lower Precision (0.8420).
- Ensemble models (Stacking) show robust performance across metrics.

4 Test Set Scores

The following table summarizes the average test set performance across all datasets:

Observations:

- CatBoost achieves the highest Test Accuracy (0.9103) and ROC-AUC (0.8428).
- Logistic Regression has the highest Recall (0.8970) but lowest Precision (0.3535).
- Precision decreases with increasing imbalance (95/5 shows the lowest), while Recall remains relatively high for Logistic Regression.

Model	Test Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.8222	0.3535	0.8970	0.4863	0.8295
Random Forest	0.9087	0.5656	0.5535	0.5562	0.8392
XGBoost	0.9092	0.5670	0.5509	0.5559	0.8407
CatBoost	0.9103	0.5733	0.5517	0.5594	0.8428
Stacking Ensemble	0.9078	0.5480	0.5341	0.5376	0.8320

Table 4: Average Test Set Performance

5 ROC Curve Comparison

ROC curves are not directly plotted in the log output, but the ROC-AUC values from the test set provide insight into model discrimination. The following qualitative comparison is based on average ROC-AUC values:

- **80/20 Dataset**: ROC-AUC ranges from 0.8495 (Random Forest) to 0.8850 (Logistic Regression), indicating good separation with high TPR and moderate FPR.
- **90/10 Dataset**: ROC-AUC ranges from 0.7789 (Random Forest) to 0.8489 (Logistic Regression), showing a slight decline due to reduced recall.
- **95/5 Dataset**: ROC-AUC ranges from 0.6903 (Stacking Ensemble) to 0.7472 (CatBoost), reflecting poorer separation due to extreme imbalance.

Graphical Representation: Refer to the saved plots (e.g., `ROC-AUC_comparison_dataset80_20_` for per-dataset ROC-AUC comparisons. A consolidated plot (`average_ROC-AUC_across_datasets.p` shows CatBoost leading with an average ROC-AUC of 0.8428.

6 Confusion Matrix Comparison

Confusion matrices are not explicitly provided, but can be inferred from Precision, Recall, and Accuracy. A qualitative comparison based on test set metrics:

- **80/20 Dataset**: High Recall (e.g., 0.9625 for Logistic Regression) suggests many true positives, with moderate Precision (0.5556) indicating some false positives.
- **90/10 Dataset**: Balanced true positives and true negatives (e.g., 0.9397 Accuracy for XGBoost), with reduced Recall (0.6027) and higher Precision (0.7454).
- **95/5 Dataset**: Low Precision (e.g., 0.1620 for Logistic Regression) and Recall (e.g., 0.4880 for Stacking) suggest high false positives and fewer true positives.

Graphical Representation: Refer to per-dataset plots (e.g., `Test_Accuracy_comparison_datase` for visual insights into performance trade-offs.

7 Key Insights

- **Effect of Class Imbalance**: Increasing imbalance (80/20 to 95/5) reduces Precision and F1 Score, with 95/5 showing the poorest test set performance due to high false positives.
- **Model Performance**:
  - Logistic Regression: High Recall but low Precision, suitable for recall-sensitive tasks.
  - Random Forest, XGBoost, CatBoost: Perform well in balanced datasets (80/20, 90/10), with CatBoost excelling overall.
  - Stacking Ensemble: Robust but degrades in 95/5 due to low true positives.
- **Augmentation Effectiveness**: SMOTE and MCMC balance training sets effectively, but test set performance highlights generalization challenges in imbalanced data.
- **ROC and Confusion Matrix Insights**: 80/20 datasets show the best ROC-AUC and balanced confusion matrices, while 95/5 exhibits the weakest performance.

8 Conclusion

The SMOTE and MCMC augmentation approaches enable robust training performance across datasets. The 80/20 datasets achieve the best test set performance (F1 Score 0.55620.7046, ROC-AUC 0.84950.8850), with balanced confusion matrices. The 90/10 datasets excel in Accuracy and Precision for ensemble models, while the 95/5 datasets pose challenges, with low Precision and F1 Scores. CatBoost with MCMC is recommended for its high Test Accuracy (0.9103) and ROC-AUC (0.8428). Further tuning and validation with real-world data are suggested to address extreme imbalance issues.

9 Recommendations

- **Model Selection**: Use CatBoost with MCMC for optimal performance. For recall-focused applications, consider Logistic Regression with SMOTE.
- **Further Analysis**: Investigate hyperparameter tuning to improve 95/5 performance, especially F1 Score.
- **Data Collection**: Incorporate more balanced datasets to enhance model generalization.

10 Appendix

- **Tabular Data**: Full results are saved in `model_performance_results.csv`.
- **Graphical Data**: Per-dataset plots (e.g., `Test_Accuracy_comparison_dataset80_20_top1` and consolidated plots (e.g., `average_Test_Accuracy_across_datasets.png`) are available in the Colab environment.
- **Log File**: Detailed logs are recorded in `phishing_detection_run.log`.