WorstCase MIA attack result report

Introduction

This report provides a summary of a series of simulated attack experiments performed on the model outputs provided. An attack model is trained to attempt to distinguish between outputs from training (in-sample) and testing (out-of-sample) data. The metrics below describe the success of this classifier. A successful classifier indicates that the original model is unsafe and should not be allowed to be released from the TRE.
In particular, the simulation splits the data provided into test and train sets (each will in- and out-of-sample examples). The classifier is trained on the train set and evaluated on the test set. This is repeated with different train/test splits a user-specified number of times.
To help place the results in context, the code may also have run a series of baseline experiments. In these, random model outputs for hypothetical in- and out-of-sample data are generated with identical statistical properties. In these baseline cases, there is no signal that an attacker could leverage and therefore these values provide a baseline against which the actual values can be compared.
For some metrics (FDIF and AUC), we are able to compute p-values. In each case, shown below (in the Global metrics sections) is the number of repetitions that exceeded the p-value threshold both without, and with correction for multiple testing (Benjamini-Hochberg procedure).
ROC curves for all real (red) and dummy (blue) repetitions are provided. These are shown in log space (as reommended here [ADD URL]) to emphasise the region in which risk is highest -- the bottom left (are high true positive rates possible with low false positive rates).
A description of the metrics and how to interpret them within the context of an attack is given below.

Experiment summary

```
                     n_reps: 10
            reproduce_split: [5, 25, 36, 49, 64, 81, 100, 121, 144, 169]
                   p_thresh: 0.05
               n_dummy_reps: 10
                 train_beta: 1
                  test_beta: 1
                  test_prop: 0.2
                  n_rows_in: 1200
                 n_rows_out: 300
   training_preds_filename: None
       test_preds_filename: None
                 output_dir: ./SVM_unsafe_90synth
                report_name: attack_output
include_model_correct_feature: False
                 sort_probs: True
           mia_attack_model: <class
'sklearn.ensemble._forest.RandomForestClassifier'>
       mia_attack_model_hyp: {'min_samples_split': 20, 'min_samples_leaf':
10, 'max_depth': 5}
  attack_metric_success_name: P_HIGHER_AUC
attack_metric_success_thresh: 0.05
attack_metric_success_comp_type: lte
attack_metric_success_count_thresh: 5
          attack_fail_fast: False
 attack_config_json_file_name: None
               target_path: None
```

Global metrics

```
        null_auc_3sd_range: 0.3748 -> 0.6252
          n_sig_auc_p_vals: 1
n_sig_auc_p_vals_corrected: 1
```

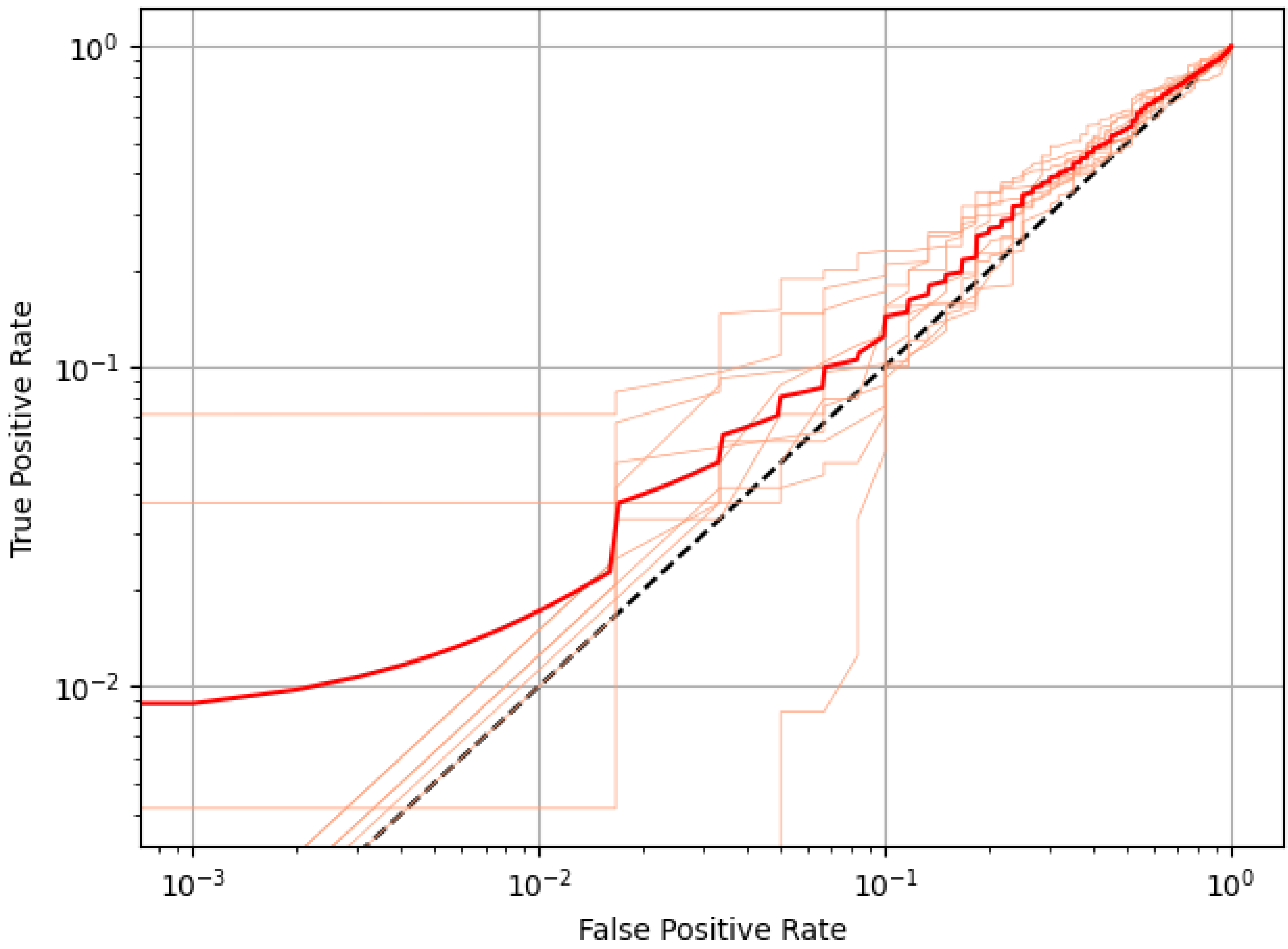
```
            n_sig_pdif_vals: 1
  n_sig_pdif_vals_corrected: 0
```

Metrics

The following show summaries of the attack metrics over the repetitions

```
          AUC mean = 0.55, var = 0.0009, min = 0.50, max = 0.61
          ACC mean = 0.80, var = 0.0000, min = 0.78, max = 0.80
    Advantage mean = 0.01, var = 0.0001, min = 0.00, max = 0.03
       FDIF01 mean = 0.03, var = 0.0119, min = -0.17, max = 0.23
       PDIF01 mean = 0.43, var = 0.0822, min = 0.01, max = 0.95
      TPR@0.1 mean = 0.14, var = 0.0022, min = 0.09, max = 0.23
     TPR@0.01 mean = 0.01, var = 0.0004, min = 0.00, max = 0.06
    TPR@0.001 mean = 0.01, var = 0.0002, min = 0.00, max = 0.05
    TPR@1e-05 mean = 0.01, var = 0.0002, min = 0.00, max = 0.05
```

Log ROC



This plot shows the False Positive Rate (x) versus the True Positive Rate (y). The axes are in log space enabling us to focus on areas where the False Positive Rate is low (left hand area). Curves above the y = x line (black dashes) in this region represent a disclosure risk as an attacker can obtain many more true than false positives. The solid coloured lines show the curves for the attack simulations with the true model outputs. The lighter grey lines show the curves for randomly generated outputs with no structure (i.e. in- and out-of- sample predictions are generated from the same distributions. Solid curves consistently higher than the grey curves in the left hand part of the plot are a sign of concern.

Glossary

AUC

Area

True Positive Rate (TPR)

The t
posit
exam
these

ACC

The p

Likelihood Ratio Attack Report

Introduction

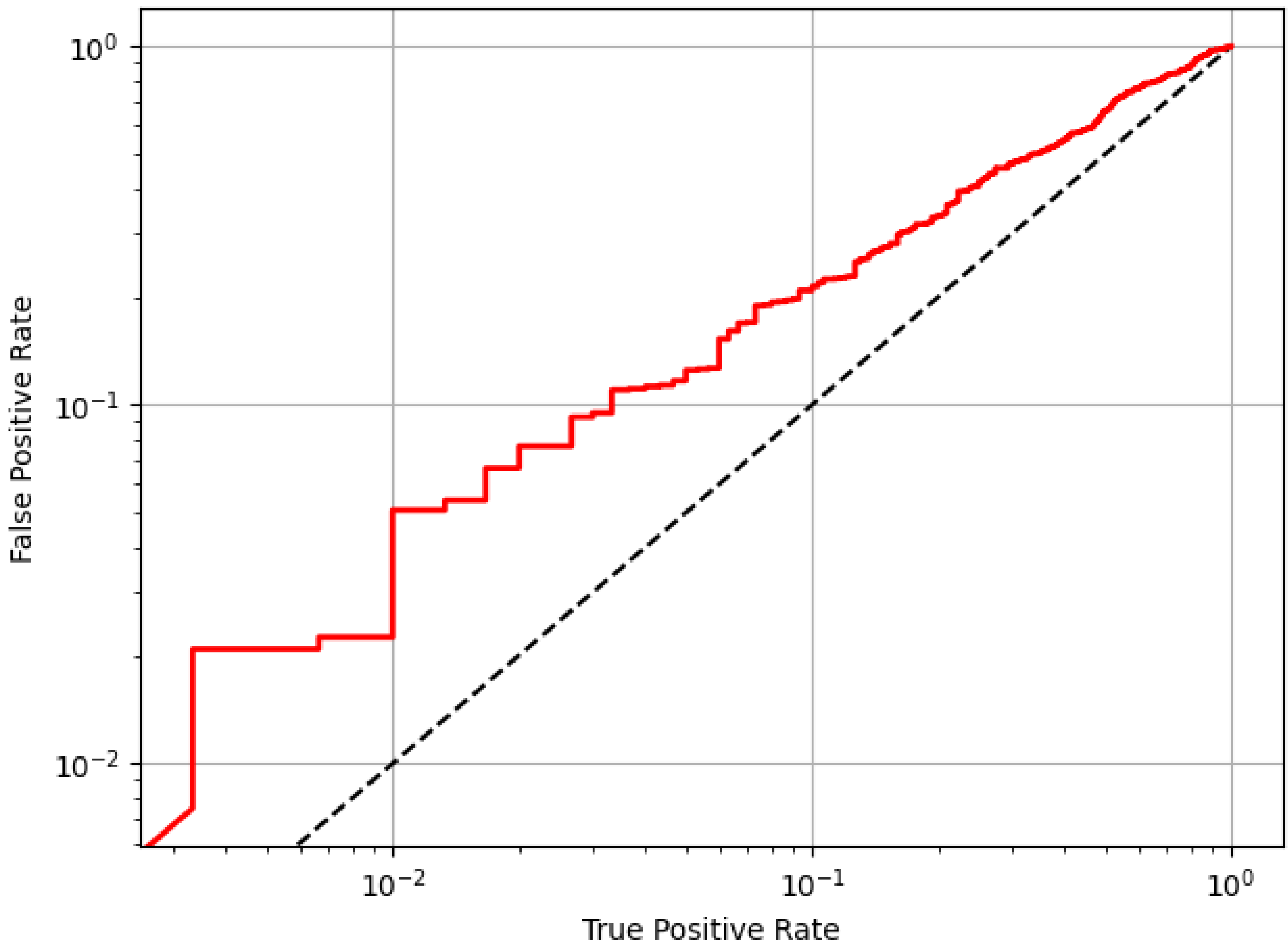
Metadata

```
             n_shadow_models: 100
                    p_thresh: 0.05
                  output_dir: ./SVM_unsafe_90synth
                 report_name: attack_output
      training_data_filename: train_data.csv
          test_data_filename: test_data.csv
     training_preds_filename: train_preds.csv
         test_preds_filename: test_preds.csv
                target_model: ['sklearn.svm']
            target_model_hyp: {'C': 10, 'gamma': 'scale'}
  attack_config_json_file_name: ./SVM_unsafe_90synth\lira_config.json
n_shadow_rows_confidences_min: 10
      shadow_models_fail_fast: False
                 target_path: None
                    PDIF_sig: Significant at p=0.05
                     AUC_sig: Significant at p=0.05
          null_auc_3sd_range: 0.44407966976730795 -> 0.555920330232692
```

Metrics

```
         TPR: 0.6692
         FPR: 0.5067
         FAR: 0.1592
         TNR: 0.4933
         PPV: 0.8408
         NPV: 0.2716
         FNR: 0.3308
         ACC: 0.6340
     F1score: 0.7452
   Advantage: 0.1625
         AUC: 0.6196
P_HIGHER_AUC: 0.0000
      FMAX01: 0.9067
      FMIN01: 0.6533
      FDIF01: 0.2533
      PDIF01: 0.0000
      FMAX02: 0.8933
      FMIN02: 0.7000
      FDIF02: 0.1933
      PDIF02: 20.2448
     FMAX001: 0.9333
     FMIN001: 0.5333
     FDIF001: 0.4000
     PDIF001: 5.7812
pred_prob_var: 0.1461
```

ROC Curve



10^0
10^{-1}
10^{-2}
False Positive Rate
10^{-2}
10^{-1}
10^0
True Positive Rate

WorstCase MIA attack result report

Introduction

This report provides a summary of a series of simulated attack experiments performed on the model outputs provided. An attack model is trained to attempt to distinguish between outputs from training (in-sample) and testing (out-of-sample) data. The metrics below describe the success of this classifier. A successful classifier indicates that the original model is unsafe and should not be allowed to be released from the TRE.
In particular, the simulation splits the data provided into test and train sets (each will in- and out-of-sample examples). The classifier is trained on the train set and evaluated on the test set. This is repeated with different train/test splits a user-specified number of times.
To help place the results in context, the code may also have run a series of baseline experiments. In these, random model outputs for hypothetical in- and out-of-sample data are generated with identical statistical properties. In these baseline cases, there is no signal that an attacker could leverage and therefore these values provide a baseline against which the actual values can be compared.
For some metrics (FDIF and AUC), we are able to compute p-values. In each case, shown below (in the Global metrics sections) is the number of repetitions that exceeded the p-value threshold both without, and with correction for multiple testing (Benjamini-Hochberg procedure).
ROC curves for all real (red) and dummy (blue) repetitions are provided. These are shown in log space (as reommended here [ADD URL]) to emphasise the region in which risk is highest -- the bottom left (are high true positive rates possible with low false positive rates).
A description of the metrics and how to interpret them within the context of an attack is given below.

Experiment summary

```
                      n_reps: 10
             reproduce_split: [5, 25, 36, 49, 64, 81, 100, 121, 144, 169]
                    p_thresh: 0.05
                n_dummy_reps: 10
                  train_beta: 1
                   test_beta: 1
                   test_prop: 0.2
                   n_rows_in: 1200
                  n_rows_out: 186
       training_preds_filename: None
           test_preds_filename: None
                  output_dir: ./SVM_unsafe_90synth
                 report_name: attack_output
 include_model_correct_feature: False
                  sort_probs: True
            mia_attack_model: <class
'sklearn.ensemble._forest.RandomForestClassifier'>
        mia_attack_model_hyp: {'min_samples_split': 20, 'min_samples_leaf':
10, 'max_depth': 5}
    attack_metric_success_name: P_HIGHER_AUC
  attack_metric_success_thresh: 0.05
attack_metric_success_comp_type: lte
attack_metric_success_count_thresh: 5
            attack_fail_fast: False
  attack_config_json_file_name: None
                 target_path: None
```

Global metrics

```
          null_auc_3sd_range: 0.3468 -> 0.6532
            n_sig_auc_p_vals: 2
    n_sig_auc_p_vals_corrected: 1
```

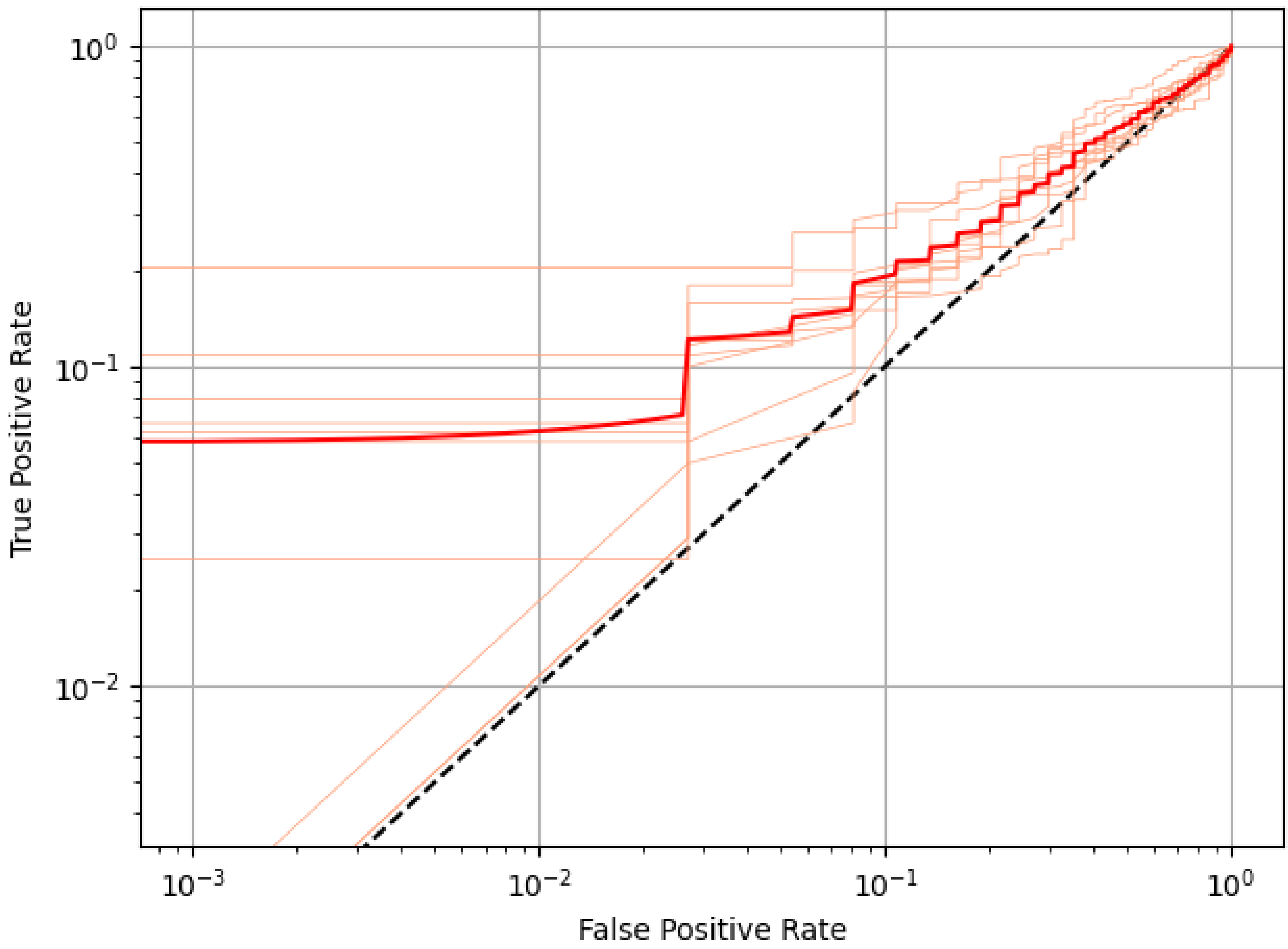
```
           n_sig_pdif_vals: 1
 n_sig_pdif_vals_corrected: 1
```

Metrics

The following show summaries of the attack metrics over the repetitions

```
      AUC mean = 0.55, var = 0.0020, min = 0.50, max = 0.63
      ACC mean = 0.87, var = 0.0000, min = 0.87, max = 0.87
Advantage mean = 0.00, var = 0.0000, min = 0.00, max = 0.00
   FDIF01 mean = 0.07, var = 0.0060, min = 0.00, max = 0.29
   PDIF01 mean = 0.29, var = 0.0182, min = 0.00, max = 0.50
  TPR@0.1 mean = 0.19, var = 0.0027, min = 0.12, max = 0.30
 TPR@0.01 mean = 0.06, var = 0.0036, min = 0.00, max = 0.20
TPR@0.001 mean = 0.06, var = 0.0036, min = 0.00, max = 0.20
TPR@1e-05 mean = 0.06, var = 0.0036, min = 0.00, max = 0.20
```

Log ROC



This plot shows the False Positive Rate (x) versus the True Positive Rate (y). The axes are in log space enabling us to focus on areas where the False Positive Rate is low (left hand area). Curves above the y = x line (black dashes) in this region represent a disclosure risk as an attacker can obtain many more true than false positives. The solid coloured lines show the curves for the attack simulations with the true model outputs. The lighter grey lines show the curves for randomly generated outputs with no structure (i.e. in- and out-of- sample predictions are generated from the same distributions. Solid curves consistently higher than the grey curves in the left hand part of the plot are a sign of concern.

Glossary

AUC

Area

True Positive Rate (TPR)

The t
posit
exam
these

ACC

The p

Likelihood Ratio Attack Report

Introduction

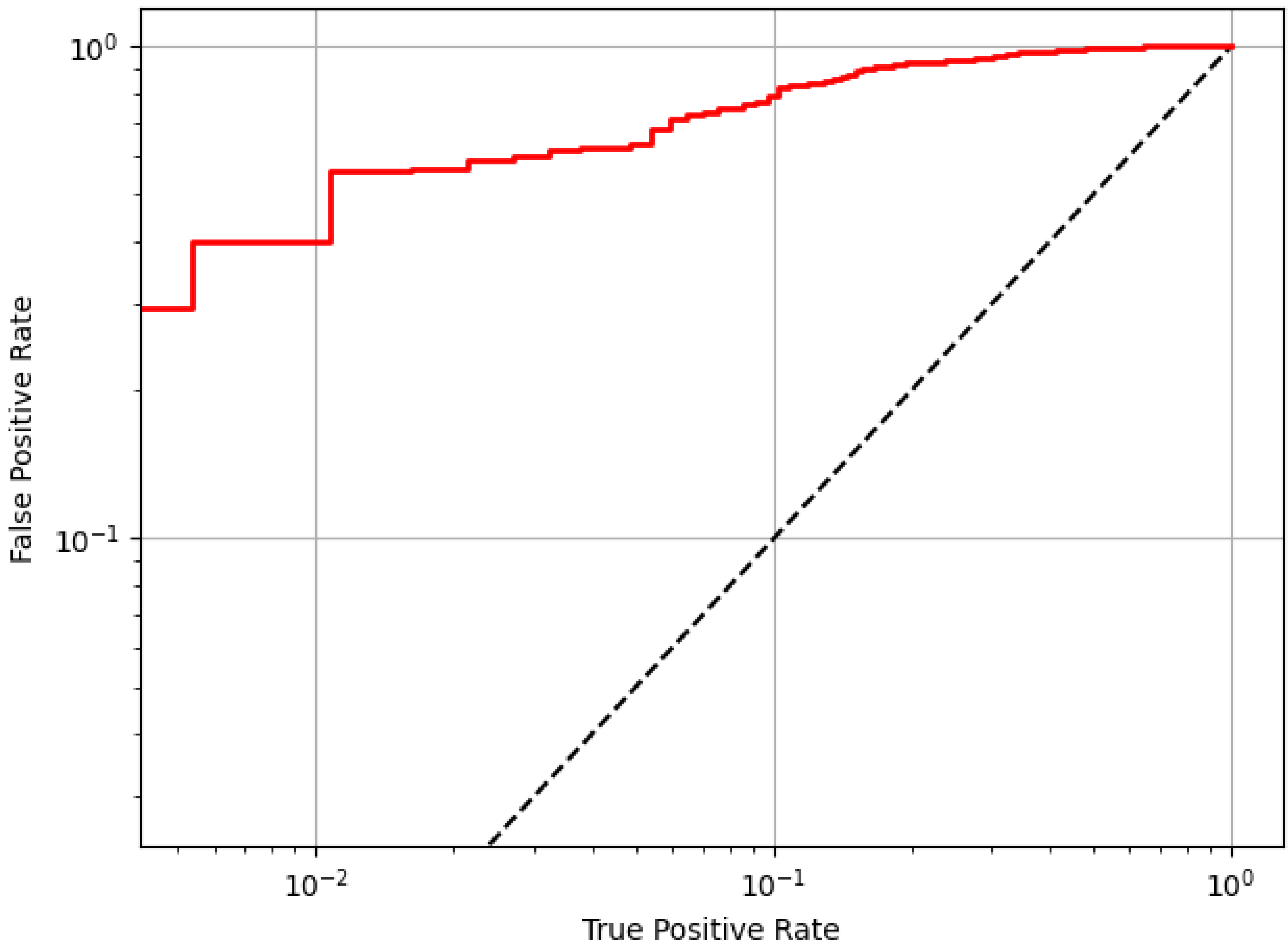
Metadata

```
           n_shadow_models: 100
                  p_thresh: 0.05
                output_dir: ./SVM_unsafe_90synth
               report_name: attack_output
    training_data_filename: train_data.csv
        test_data_filename: test_data.csv
   training_preds_filename: train_preds.csv
       test_preds_filename: test_preds.csv
              target_model: ['sklearn.svm']
          target_model_hyp: {'C': 10, 'gamma': 'scale'}
attack_config_json_file_name: ./SVM_unsafe_90synth\lira_config.json
n_shadow_rows_confidences_min: 10
   shadow_models_fail_fast: False
               target_path: None
                  PDIF_sig: Significant at p=0.05
                   AUC_sig: Significant at p=0.05
        null_auc_3sd_range: 0.4317312789077612 -> 0.5682687210922388
```

Metrics

```
          TPR: 0.6650
          FPR: 0.0538
          FAR: 0.0124
          TNR: 0.9462
          PPV: 0.9876
          NPV: 0.3045
          FNR: 0.3350
          ACC: 0.7027
      F1score: 0.7948
    Advantage: 0.6112
          AUC: 0.9391
 P_HIGHER_AUC: 0.0000
       FMAX01: 1.0000
       FMIN01: 0.1942
       FDIF01: 0.8058
       PDIF01: 0.0000
       FMAX02: 1.0000
       FMIN02: 0.4353
       FDIF02: 0.5647
       PDIF02: 115.1300
      FMAX001: 1.0000
      FMIN001: 0.0000
      FDIF001: 1.0000
      PDIF001: 33.0992
pred_prob_var: 0.1764
```

ROC Curve



WorstCase MIA attack result report

Introduction

This report provides a summary of a series of simulated attack experiments performed on the model outputs provided. An attack model is trained to attempt to distinguish between outputs from training (in-sample) and testing (out-of-sample) data. The metrics below describe the success of this classifier. A successful classifier indicates that the original model is unsafe and should not be allowed to be released from the TRE.
In particular, the simulation splits the data provided into test and train sets (each will in- and out-of-sample examples). The classifier is trained on the train set and evaluated on the test set. This is repeated with different train/test splits a user-specified number of times.
To help place the results in context, the code may also have run a series of baseline experiments. In these, random model outputs for hypothetical in- and out-of-sample data are generated with identical statistical properties. In these baseline cases, there is no signal that an attacker could leverage and therefore these values provide a baseline against which the actual values can be compared.
For some metrics (FDIF and AUC), we are able to compute p-values. In each case, shown below (in the Global metrics sections) is the number of repetitions that exceeded the p-value threshold both without, and with correction for multiple testing (Benjamini-Hochberg procedure).
ROC curves for all real (red) and dummy (blue) repetitions are provided. These are shown in log space (as reommended here [ADD URL]) to emphasise the region in which risk is highest -- the bottom left (are high true positive rates possible with low false positive rates).
A description of the metrics and how to interpret them within the context of an attack is given below.

Experiment summary

```
                      n_reps: 10
             reproduce_split: [5, 25, 36, 49, 64, 81, 100, 121, 144, 169]
                    p_thresh: 0.05
                n_dummy_reps: 10
                  train_beta: 1
                   test_beta: 1
                   test_prop: 0.2
                   n_rows_in: 1094
                  n_rows_out: 186
     training_preds_filename: None
         test_preds_filename: None
                  output_dir: ./SVM_unsafe_90synth
                 report_name: attack_output
 include_model_correct_feature: False
                  sort_probs: True
            mia_attack_model: <class
'sklearn.ensemble._forest.RandomForestClassifier'>
        mia_attack_model_hyp: {'min_samples_split': 20, 'min_samples_leaf':
10, 'max_depth': 5}
    attack_metric_success_name: P_HIGHER_AUC
  attack_metric_success_thresh: 0.05
attack_metric_success_comp_type: lte
attack_metric_success_count_thresh: 5
           attack_fail_fast: False
  attack_config_json_file_name: None
                 target_path: None
```

Global metrics

```
          null_auc_3sd_range: 0.3458 -> 0.6542
            n_sig_auc_p_vals: 0
    n_sig_auc_p_vals_corrected: 0
```

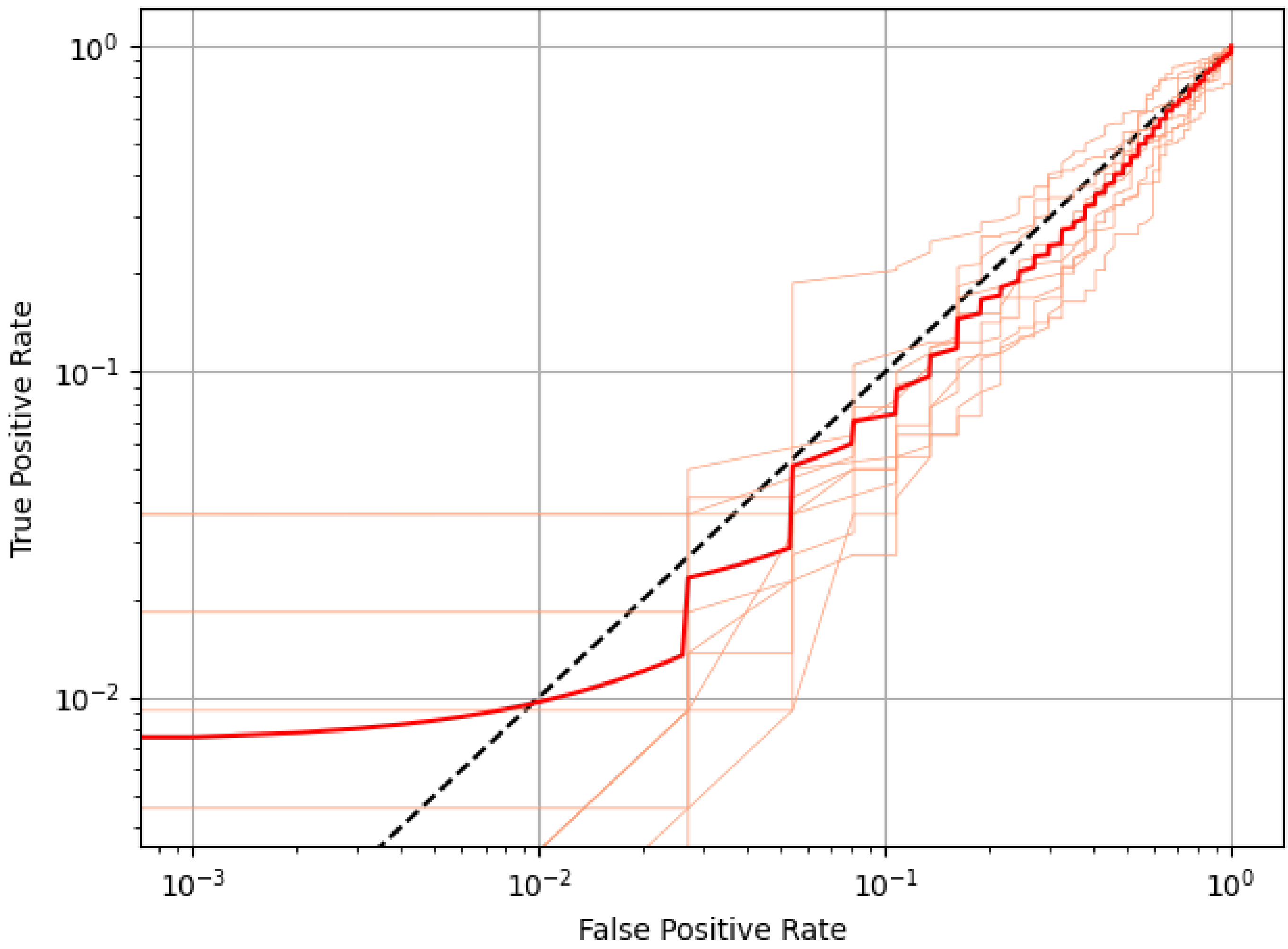
```
              n_sig_pdif_vals: 0
    n_sig_pdif_vals_corrected: 0
```

Metrics

The following show summaries of the attack metrics over the repetitions

```
         AUC mean = 0.46, var = 0.0038, min = 0.38, max = 0.58
         ACC mean = 0.86, var = 0.0000, min = 0.86, max = 0.86
   Advantage mean = 0.00, var = 0.0000, min = 0.00, max = 0.00
      FDIF01 mean = -0.06, var = 0.0095, min = -0.19, max = 0.08
      PDIF01 mean = 0.66, var = 0.0823, min = 0.22, max = 0.98
     TPR@0.1 mean = 0.07, var = 0.0024, min = 0.03, max = 0.20
    TPR@0.01 mean = 0.01, var = 0.0001, min = 0.00, max = 0.03
   TPR@0.001 mean = 0.01, var = 0.0001, min = 0.00, max = 0.02
   TPR@1e-05 mean = 0.01, var = 0.0001, min = 0.00, max = 0.02
```

Log ROC



This plot shows the False Positive Rate (x) versus the True Positive Rate (y). The axes are in log space enabling us to focus on areas where the False Positive Rate is low (left hand area). Curves above the y = x line (black dashes) in this region represent a disclosure risk as an attacker can obtain many more true than false positives. The solid coloured lines show the curves for the attack simulations with the true model outputs. The lighter grey lines show the curves for randomly generated outputs with no structure (i.e. in- and out-of- sample predictions are generated from the same distributions. Solid curves consistently higher than the grey curves in the left hand part of the plot are a sign of concern.

Glossary

AUC

Area

True Positive Rate (TPR)

The t
posit
exam
these

ACC

The p

Likelihood Ratio Attack Report

Introduction

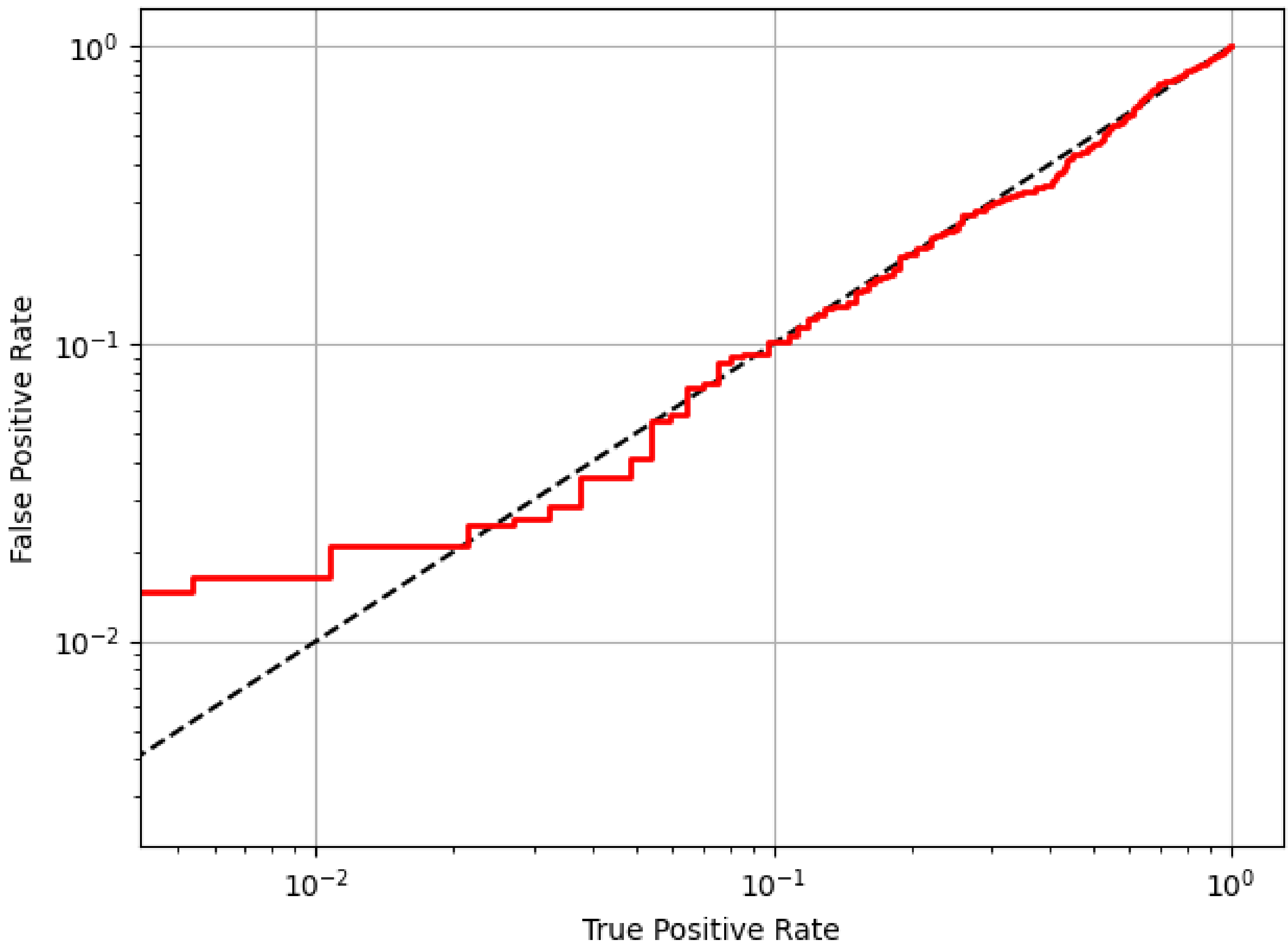
Metadata

```
           n_shadow_models: 100
                  p_thresh: 0.05
                output_dir: ./SVM_unsafe_90synth
               report_name: attack_output
    training_data_filename: train_data.csv
        test_data_filename: test_data.csv
   training_preds_filename: train_preds.csv
       test_preds_filename: test_preds.csv
              target_model: ['sklearn.svm']
          target_model_hyp: {'C': 10, 'gamma': 'scale'}
attack_config_json_file_name: ./SVM_unsafe_90synth\lira_config.json
n_shadow_rows_confidences_min: 10
   shadow_models_fail_fast: False
               target_path: None
                  PDIF_sig: Not significant at p=0.05
                   AUC_sig: Not significant at p=0.05
        null_auc_3sd_range: 0.4312868175323611 -> 0.5687131824676389
```

Metrics

```
          TPR: 0.0402
          FPR: 0.0484
          FAR: 0.1698
          TNR: 0.9516
          PPV: 0.8302
          NPV: 0.1443
          FNR: 0.9598
          ACC: 0.1727
      F1score: 0.0767
    Advantage: 0.0082
          AUC: 0.4971
 P_HIGHER_AUC: 0.5507
       FMAX01: 0.8594
       FMIN01: 0.8594
       FDIF01: 0.0000
       PDIF01: 0.5000
       FMAX02: 0.8555
       FMIN02: 0.8477
       FDIF02: 0.0078
       PDIF02: 0.9138
      FMAX001: 1.0000
      FMIN001: 0.7692
      FDIF001: 0.2308
      PDIF001: 3.0468
 pred_prob_var: 0.0316
```

ROC Curve



10^0
10^{-1}
10^{-2}
False Positive Rate
10^{-2}
10^{-1}
10^0
True Positive Rate