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
                    p_thresh: 0.05
                n_dummy_reps: 1
                  train_beta: 5
                   test_beta: 2
                   test_prop: 0.5
                   n_rows_in: 398
                  n_rows_out: 171
      training_preds_filename: None
          test_preds_filename: None
                   output_dir: outputs_worstcase
                  report_name: report_worstcase
 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: 2
             attack_fail_fast: True
  attack_config_json_file_name: None
                  target_path: None
```

Global metrics

```
          null_auc_3sd_range: 0.3880 -> 0.6120
            n_sig_auc_p_vals: 2
   n_sig_auc_p_vals_corrected: 2
             n_sig_pdif_vals: 2
```

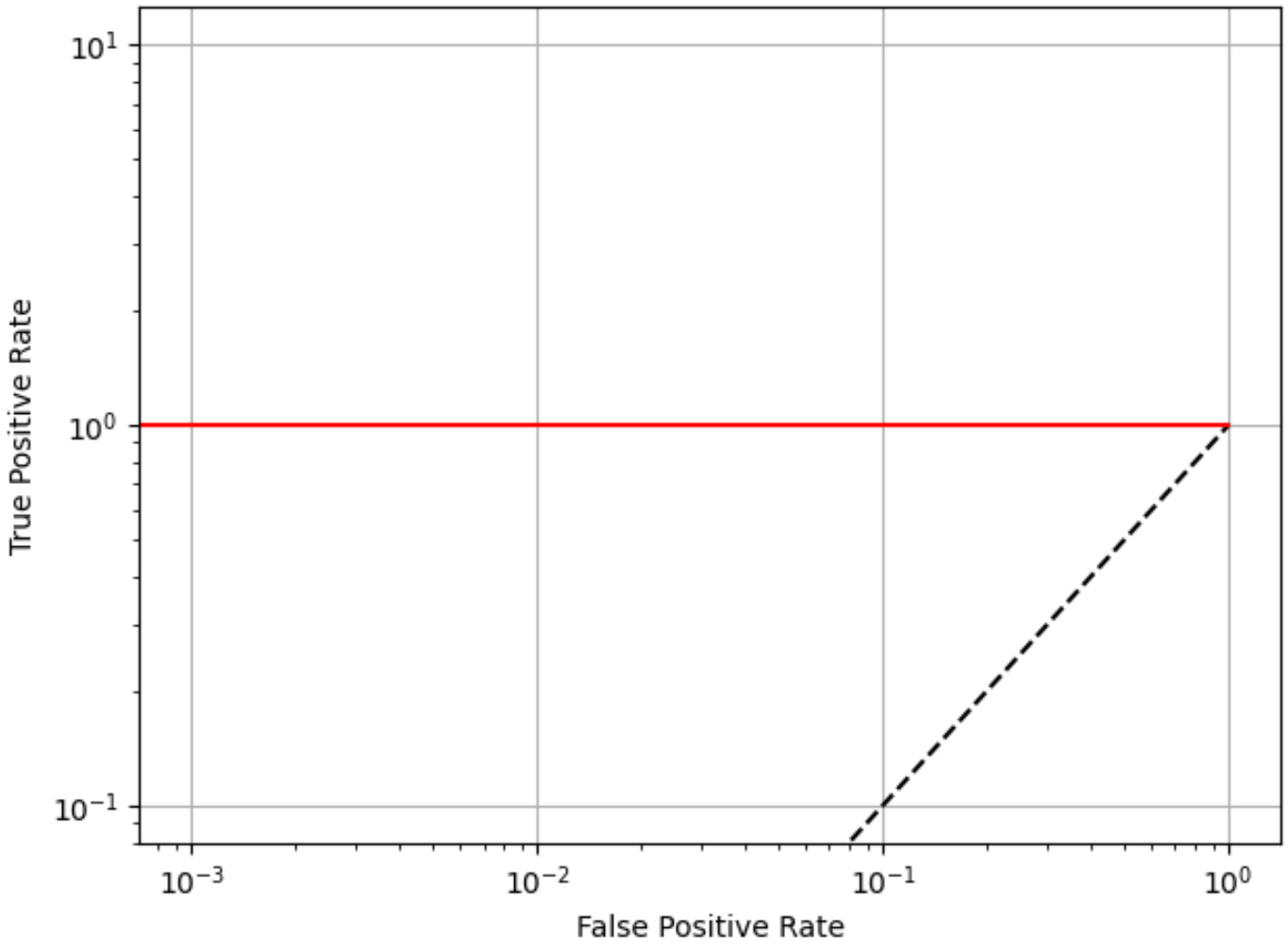
```
n_sig_pdif_vals_corrected: 2
```

Metrics

The following show summaries of the attack metrics over the repetitions

```
         AUC mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
         ACC mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
   Advantage mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
      FDIF01 mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
      PDIF01 mean = 0.00, var = 0.0000, min = 0.00, max = 0.00
    TPR@0.1 mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
   TPR@0.01 mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
  TPR@0.001 mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
  TPR@1e-05 mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
```

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

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: 20
                    p_thresh: 0.05
                n_dummy_reps: 1
                  train_beta: 5
                   test_beta: 2
                   test_prop: 0.5
                   n_rows_in: 398
                  n_rows_out: 171
      training_preds_filename: None
          test_preds_filename: None
                  output_dir: outputs_worstcase
                 report_name: report_worstcase
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: config_worstcase.json
                  target_path: None
```

Global metrics

```
          null_auc_3sd_range: 0.3880 -> 0.6120
            n_sig_auc_p_vals: 20
   n_sig_auc_p_vals_corrected: 20
             n_sig_pdif_vals: 20
```

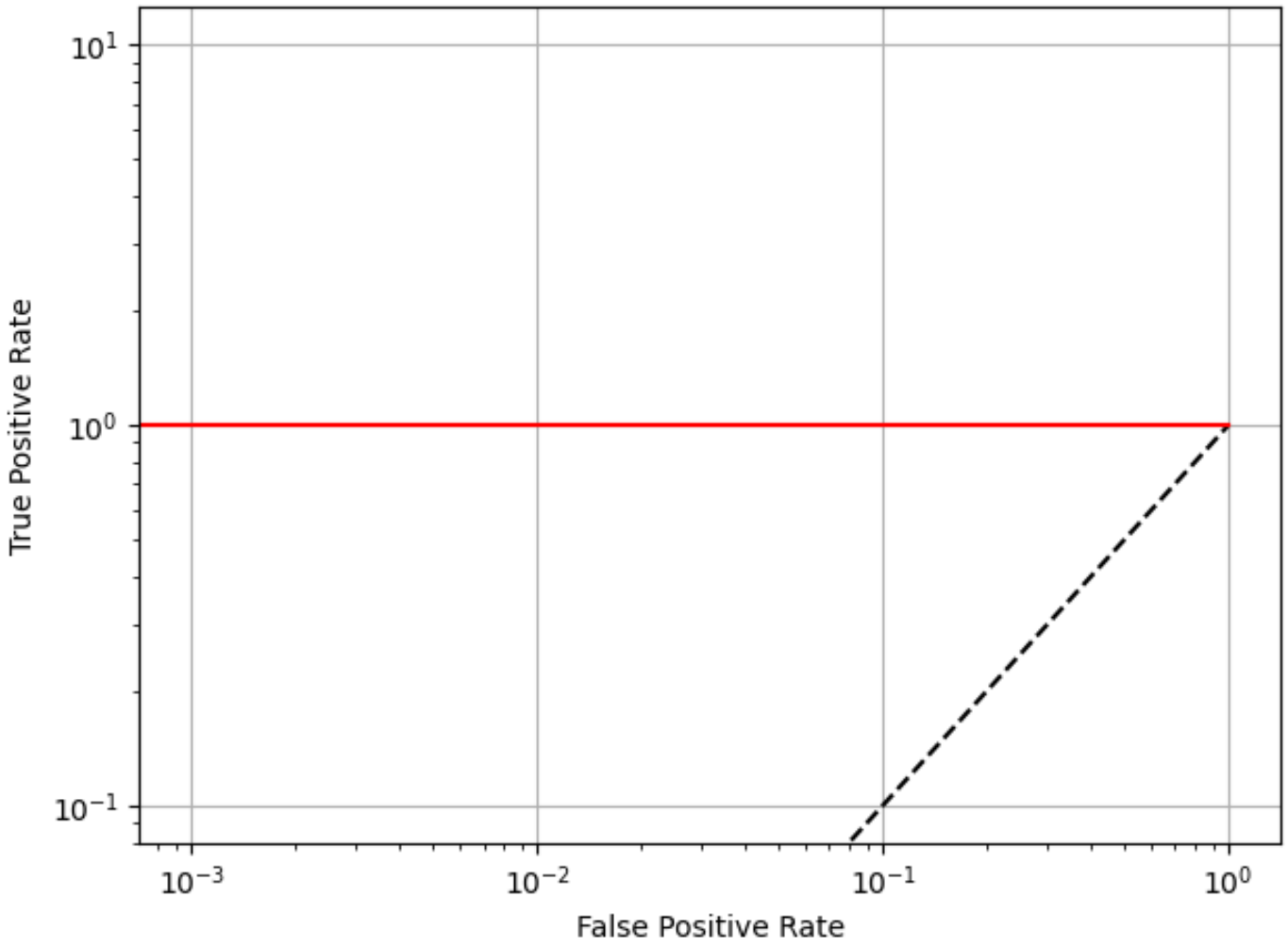
```
n_sig_pdif_vals_corrected: 20
```

Metrics

The following show summaries of the attack metrics over the repetitions

```
         AUC mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
         ACC mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
   Advantage mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
      FDIF01 mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
      PDIF01 mean = 0.00, var = 0.0000, min = 0.00, max = 0.00
    TPR@0.1 mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
   TPR@0.01 mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
  TPR@0.001 mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
  TPR@1e-05 mean = 1.00, var = 0.0000, min = 1.00, max = 1.00
```

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