

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 recommended 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: 1
train_beta: 1
test_beta: 1
test_prop: 0.5
n_rows_in: 1094
n_rows_out: 186
training_preds_filename: None
test_preds_filename: None
output_dir: ./2
report_name: attack_results
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: 0.4028 -> 0.5972
n_sig_auc_p_vals: 0
n_sig_auc_p_vals_corrected: 0
```

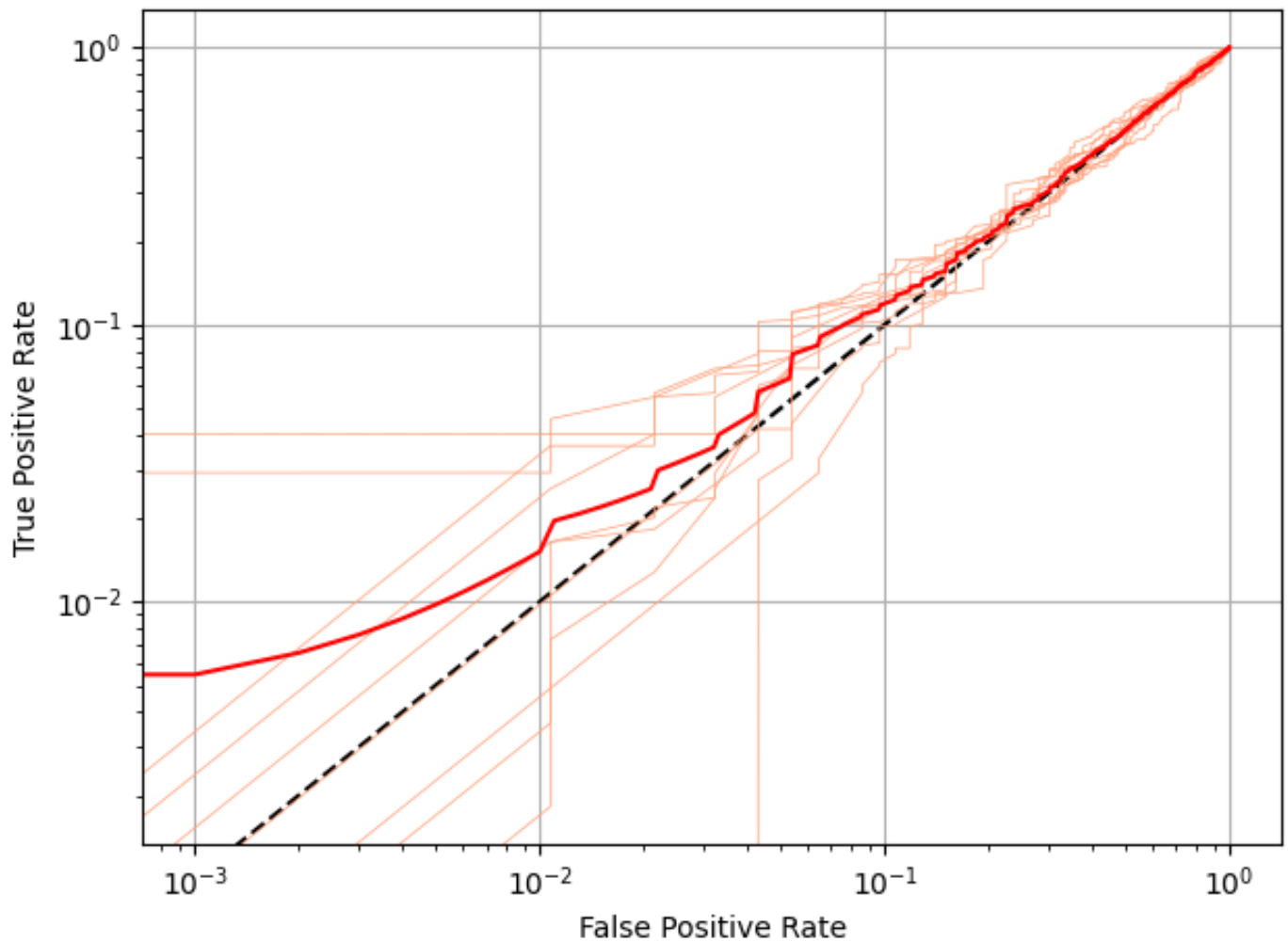
```
n_sig_pdif_vals: 1
n_sig_pdif_vals_corrected: 0
```

Metrics

The following show summaries of the attack metrics over the repetitions

```
AUC mean = 0.51, var = 0.0004, min = 0.46, max = 0.54
ACC mean = 0.85, var = 0.0000, min = 0.85, max = 0.85
Advantage mean = 0.00, var = 0.0000, min = 0.00, max = 0.00
FDIF01 mean = 0.04, var = 0.0032, min = -0.06, max = 0.12
PDIF01 mean = 0.31, var = 0.0734, min = 0.02, max = 0.84
TPR@0.1 mean = 0.12, var = 0.0005, min = 0.07, max = 0.15
TPR@0.01 mean = 0.01, var = 0.0001, min = 0.00, max = 0.03
TPR@0.001 mean = 0.00, var = 0.0001, min = 0.00, max = 0.02
TPR@1e-05 mean = 0.00, 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
thes

ACC

The p

Likelihood Ratio Attack Report

Introduction

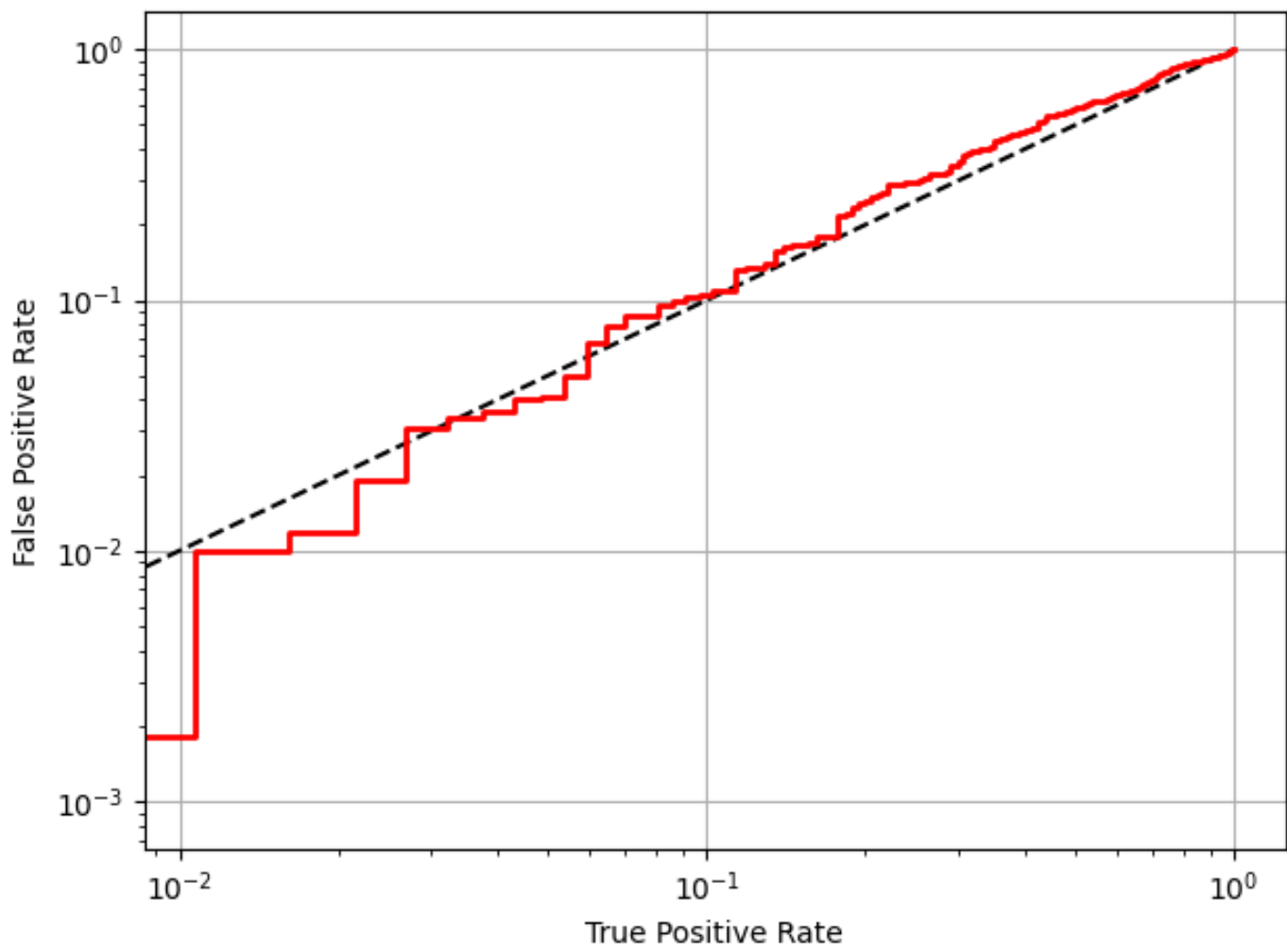
Metadata

```
n_shadow_models: 100
  p_thresh: 0.05
  output_dir: ./2
  report_name: attack_results
training_data_filename: None
  test_data_filename: None
training_preds_filename: None
  test_preds_filename: None
  target_model: None
  target_model_hyp: None
attack_config_json_file_name: None
n_shadow_rows_confidences_min: 10
  shadow_models_fail_fast: False
  target_path: None
  PDIF_sig: Not significant at p=0.05
  AUC_sig: Significant at p=0.05
null_auc_3sd_range: 0.4312868175323611 -> 0.5687131824676389
```

Metrics

```
TPR: 0.3400
FPR: 0.2903
FAR: 0.1268
TNR: 0.7097
PPV: 0.8732
NPV: 0.1546
FNR: 0.6600
ACC: 0.3937
Flscore: 0.4895
Advantage: 0.0497
AUC: 0.5456
P_HIGHER_AUC: 0.0234
  FMAX01: 0.8672
  FMIN01: 0.8125
  FDIF01: 0.0547
  PDIF01: 0.1072
  FMAX02: 0.8711
  FMIN02: 0.8125
  FDIF02: 0.0586
  PDIF02: 3.5071
  FMAX001: 0.8462
  FMIN001: 0.8462
  FDIF001: 0.0000
  PDIF001: 0.6931
pred_prob_var: 0.0181
```

ROC Curve



Attribute Inference Attack Report

Introduction

Metadata

```
output_dir: ./2
report_name: attack_results
n_cpu: 11
attack_config_json_file_name: None
target_path: None
```

Metrics

Categorical Features:

Quantitative Features:

Plots