Synthetic Data Figures and Analysis

Nicholas Christakis, Panagiotis Tirchas and Dimitris Drikakis

1. Introduction

This supplementary section presents the full synthetic-data figure set used to evaluate mutual-information (MI) feature filtering and oracle-informed feature removal under controlled signal-to-noise regimes. Each synthetic dataset contains a fixed total *d* of continuous predictors, partitioned into an oracle-informative subset of size *k* and an oracle-irrelevant (noise) subset of size *(d-k)*. The binary outcome is generated with approximately balanced classes. Informative features are constructed to exhibit an intermediate marginal association with the label; in practice, the generator setting used in code yields an expected point-biserial correlation of approximately 0.3 for informative coordinates, while noise features are generated independently of the label (up to finite-sample fluctuations).

For each value of *k* (namely, 2, 10 and 18), figures are shown for three sample sizes (full dataset, 10% of dataset and 1% of dataset) and two training settings: (i) training on all features ("all features") and (ii) training restricted to the oracle-informative subset ("only informative"). For each configuration, we report: (a) MI values, used to screen relevance; (b) SHAP summary plots for both class outputs (class 0 on the left, class 1 on the right) comparing all-features training to informative-only training; (c) ROC curves and confusion matrices comparing the two training settings at each sample size.

The full datasets consist of 10,000 instances, 20 continuous features X1--X20 and one categorical output Y.

2. Informative Features *k*=2

Feature	Mutual Information
X1	0.3270
X2	0.3303
X3	0.0000
X4	0.0045
X5	-0.0031
X6	0.0056
X7	-0.0017
X8	-0.0013
X9	0.0000
X10	0.0024
X11	0.0016
X12	0.0013
X13	0.0055
X14	-0.0114
X15	0.0028
X16	-0.0049
X17	0.0033
X18	0.0000
X19	0.0061
X20	0.0000

Figure S2.1. Mutual information for all features. Mutual information (MI) estimates between each of the continuous predictors and the binary label for the regime with oracle-informative predictors. The informative coordinates are constructed to achieve an empirical point-biserial correlation with the label of approximately 0.3, while remaining coordinates are oracle-irrelevant noise features.

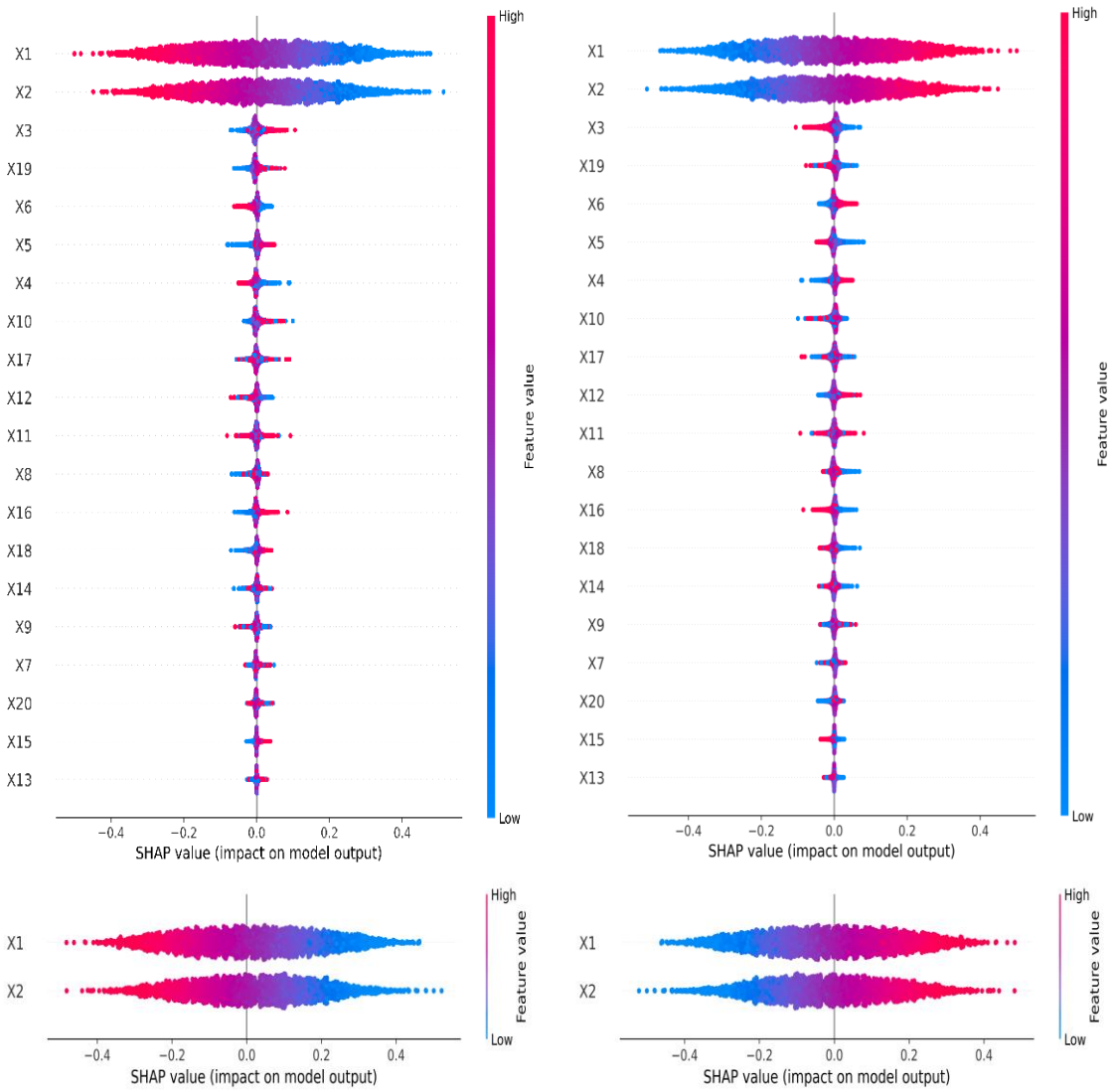

Figure S2.2. SHAP summary plots for informative features (all features vs only informative). SHAP summary plots shown separately for class 0 (left) and class 1 (right). Top row: model trained on all features. Bottom row: model trained only on the oracle-informative subset.

2.1 Full Dataset

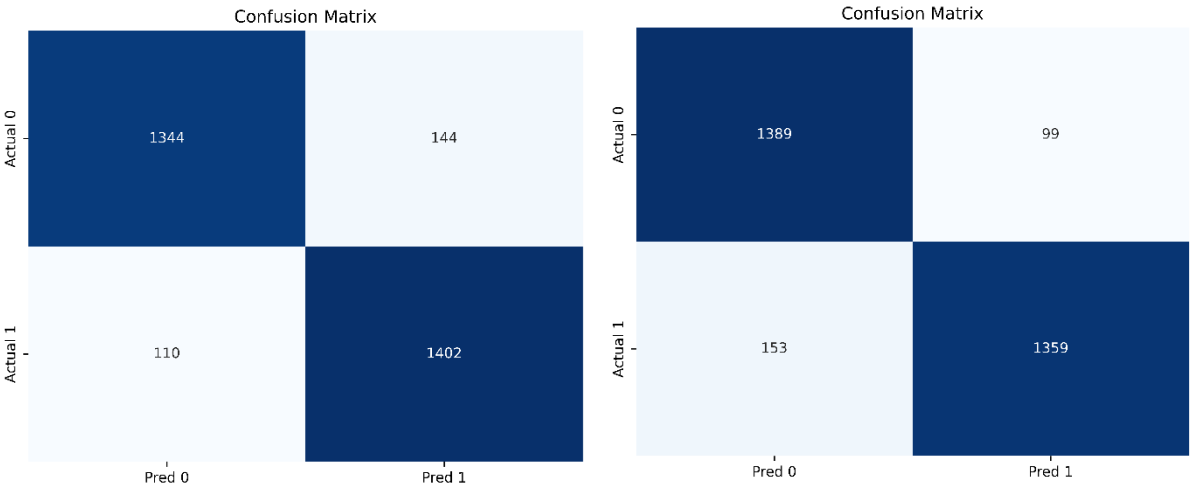

Figure S2.3. Confusion matrices (left: all features; right: only informative) for the binary classifier trained with observations.

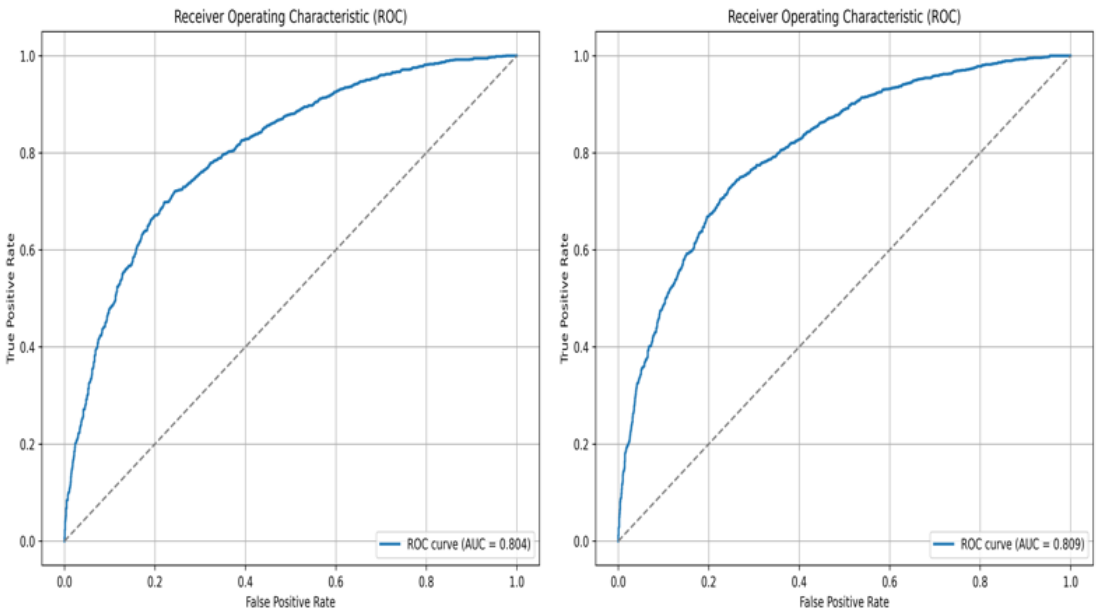

Figure S2.4. ROC curves and (left: all features; right: only informative) for the binary classifier trained with observations.

2.2 Reduced Data: 10% of Dataset

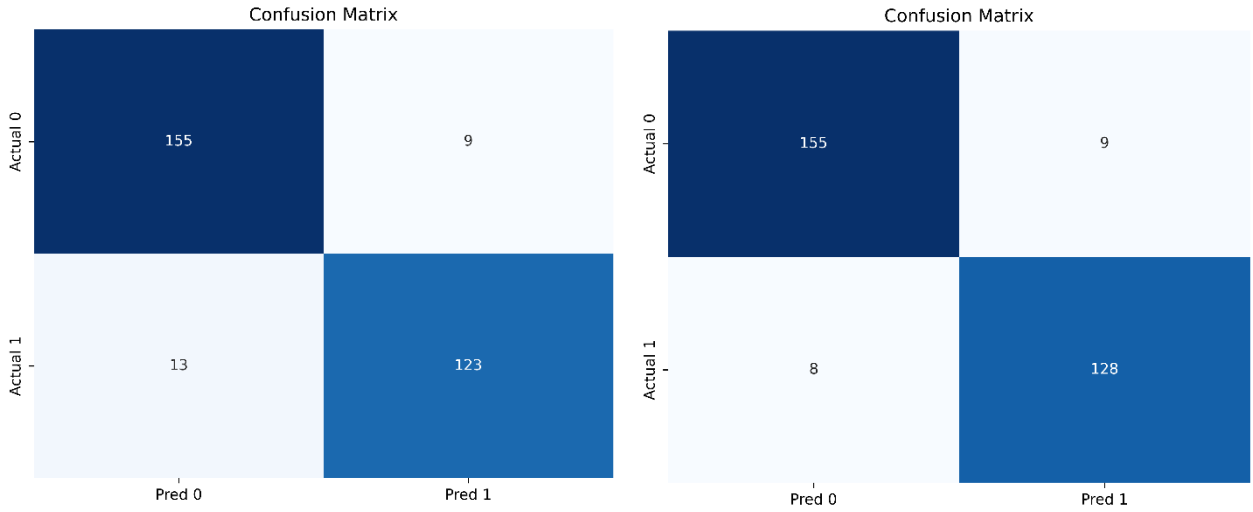

Figure S2.5. Confusion matrices (left: all features; right: only informative) for the binary classifier trained with observations.

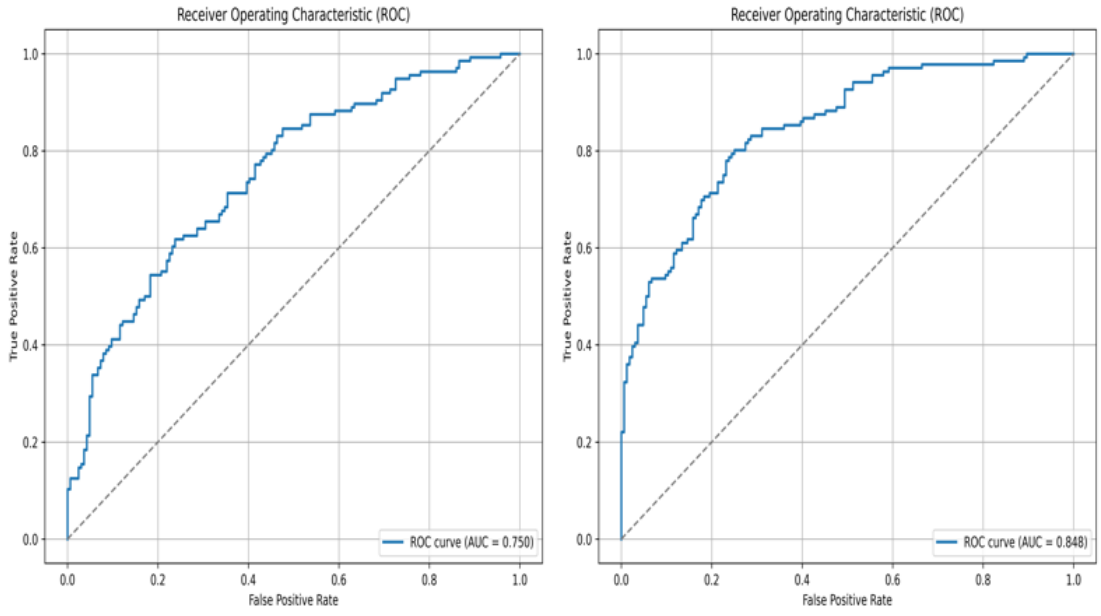

Figure S2.6. ROC curves and (left: all features; right: only informative) for the binary classifier trained with observations.

2.3 Sparse Data: 1% of Dataset

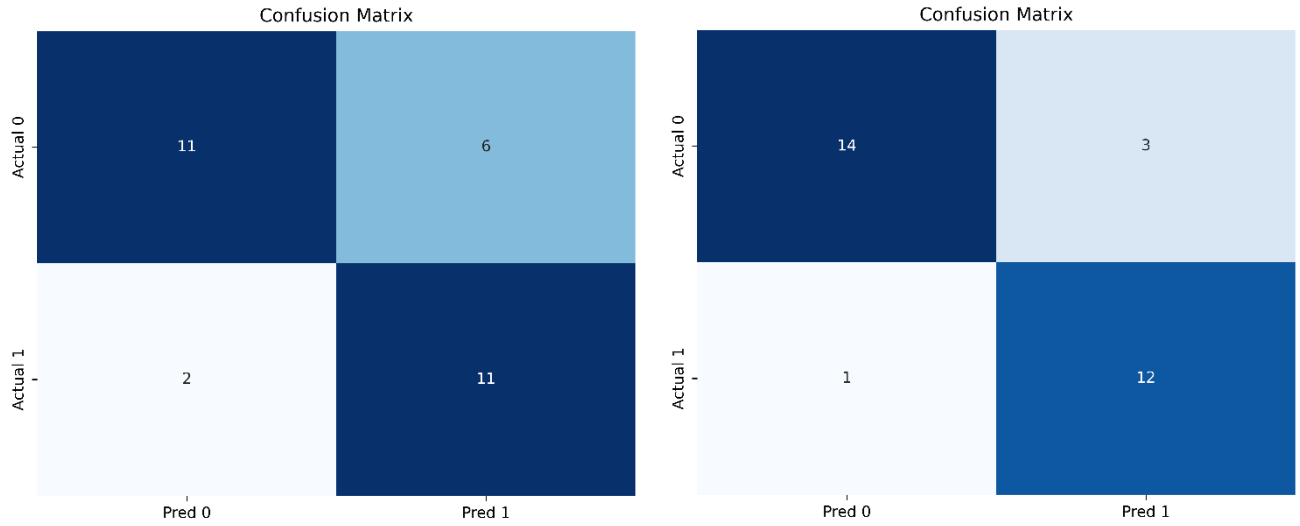

Figure S2.7. Confusion matrices (left: all features; right: only informative) for the binary classifier trained with observations.

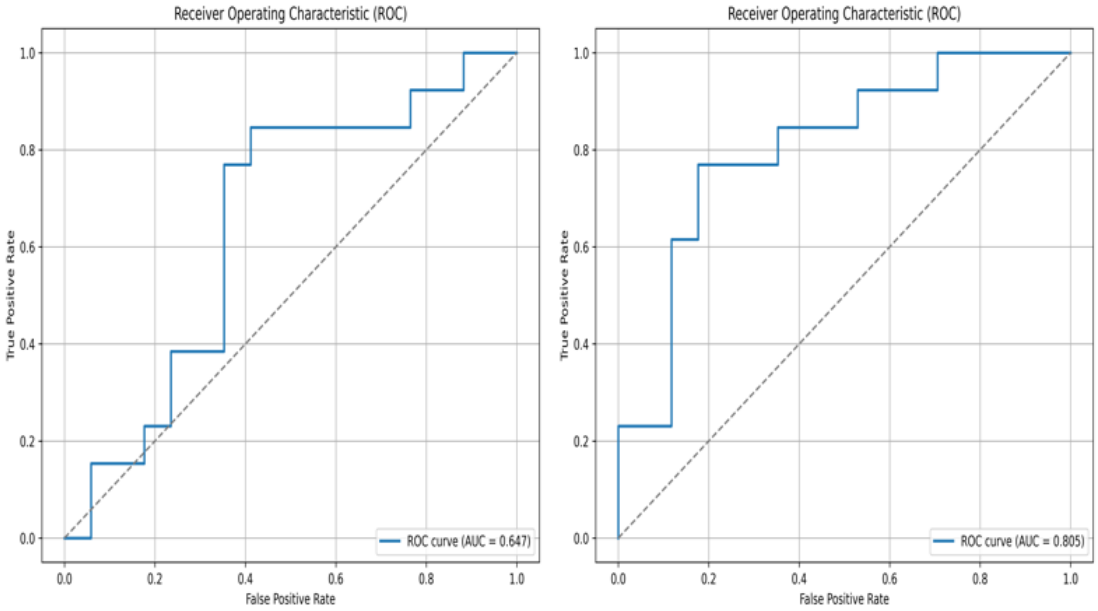

Figure S2.8. ROC curves and (left: all features; right: only informative) for the binary classifier trained with observations.

3. Informative Features *k*=10

Feature	Mutual Information
X1	0.3230
X2	0.3253
X3	0.3266
X4	0.3253
X5	0.3279
X6	0.3238
X7	0.3290
X8	0.3288
X9	0.3228
X10	0.3312
X11	-0.0014
X12	0.0050
X13	0.0000
X14	0.0000
X15	0.0000
X16	0.0000
X17	0.0072
X18	0.0000
X19	-0.0071
X20	0.0000

Figure S3.1. Mutual information for all features. Mutual information (MI) estimates between each of the continuous predictors and the binary label for the regime with oracle-informative predictors. The informative coordinates are constructed to achieve an empirical point-biserial correlation with the label of approximately 0.3, while remaining coordinates are oracle-irrelevant noise features.

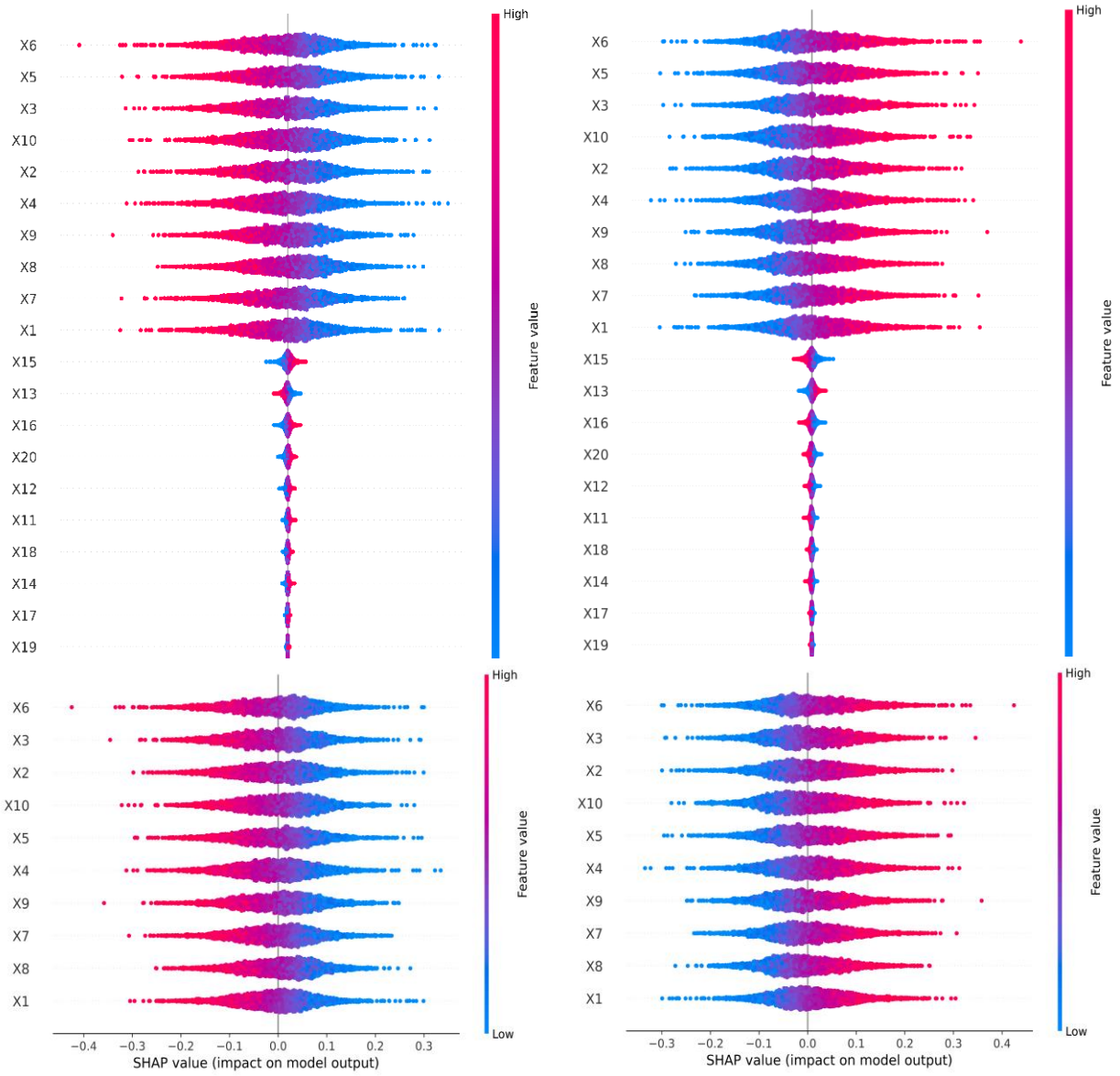

Figure S3.2. SHAP summary plots for informative features (all features vs only informative). SHAP summary plots shown separately for class 0 (left) and class 1 (right). Top row: model trained on all features. Bottom row: model trained only on the oracle-informative subset.

3.1 Full Dataset

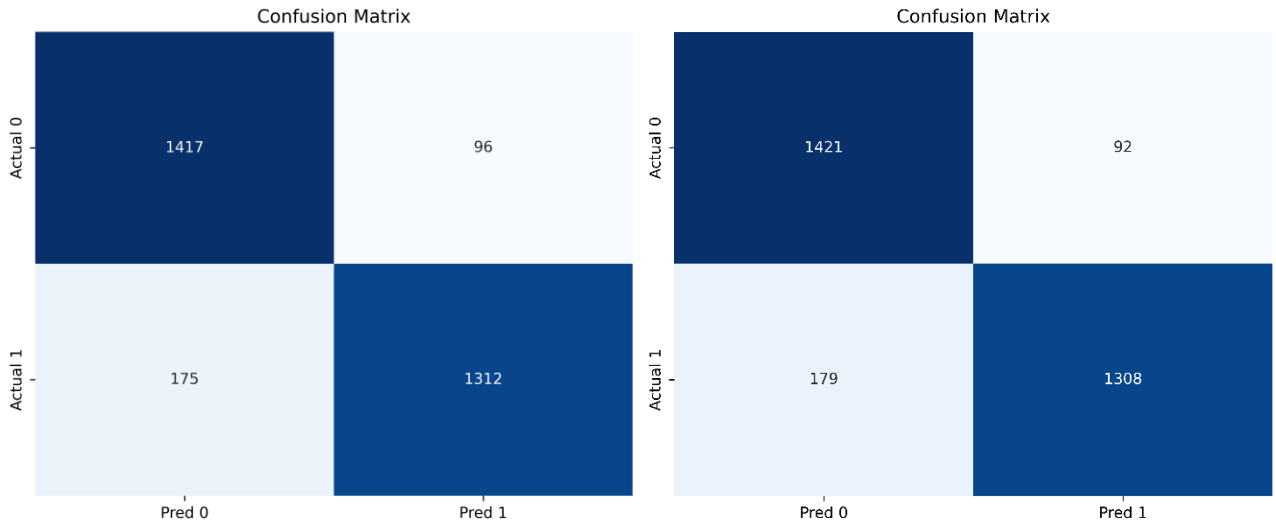

Figure S3.3. Confusion matrices (left: all features; right: only informative) for the binary classifier trained with observations.

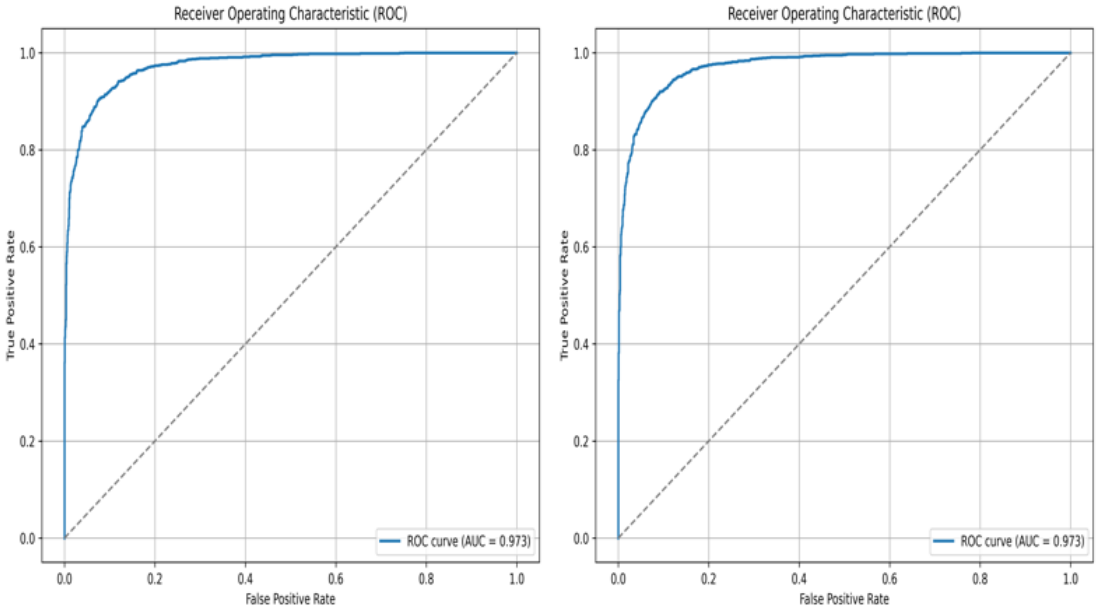

Figure S3.4. ROC curves and (left: all features; right: only informative) for the binary classifier trained with observations.

3.2 Reduced Data: 10% of Dataset

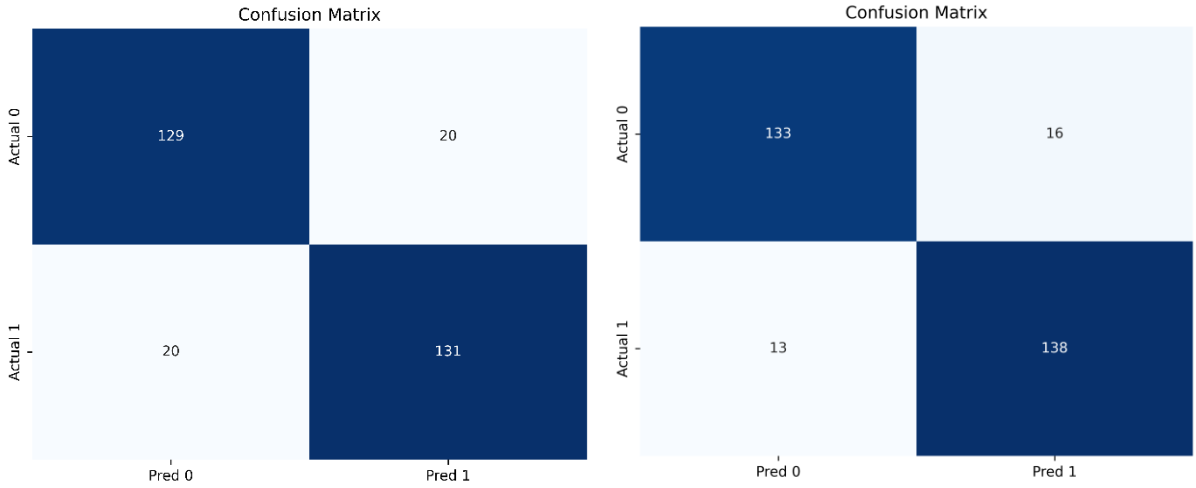

Figure S3.5. Confusion matrices (left: all features; right: only informative) for the binary classifier trained with observations.

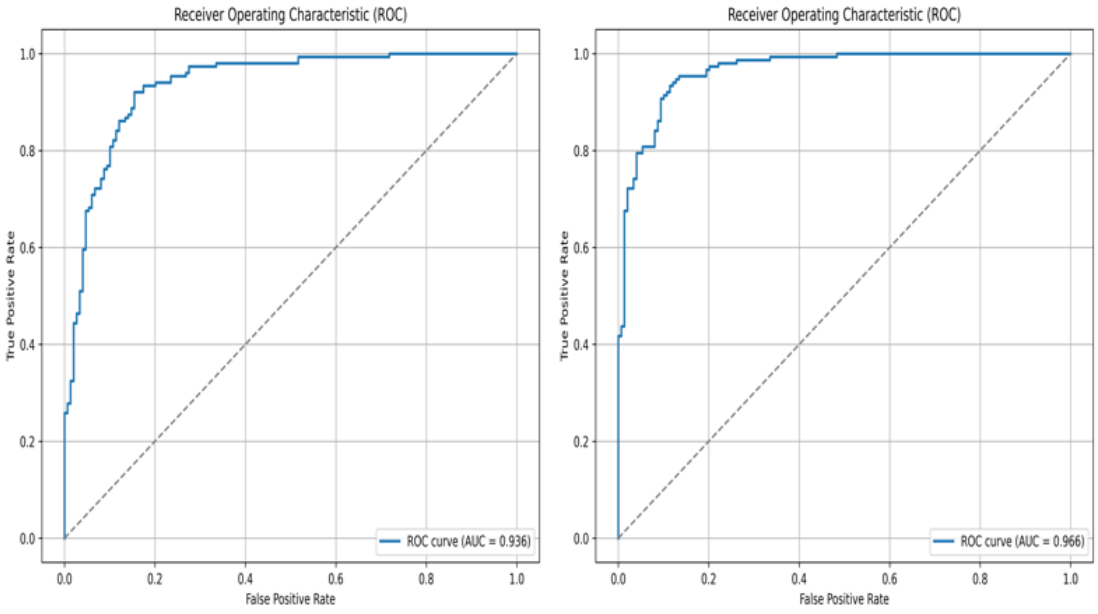

Figure S3.6. ROC curves and (left: all features; right: only informative) for the binary classifier trained with observations.

3.3 Sparse Data: 1% of Dataset

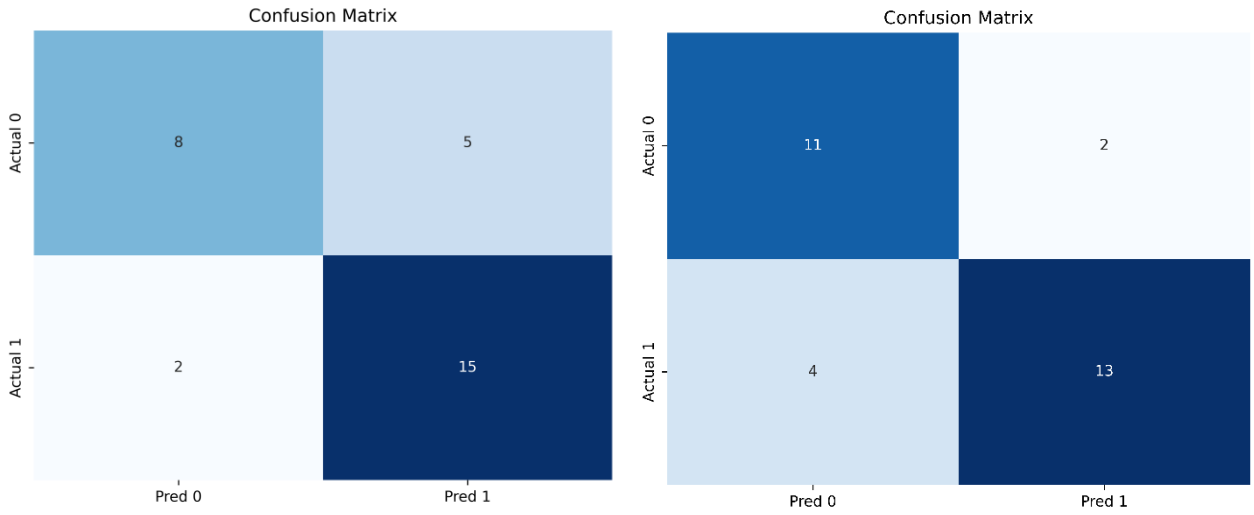

Figure S3.7. Confusion matrices (left: all features; right: only informative) for the binary classifier trained with observations.

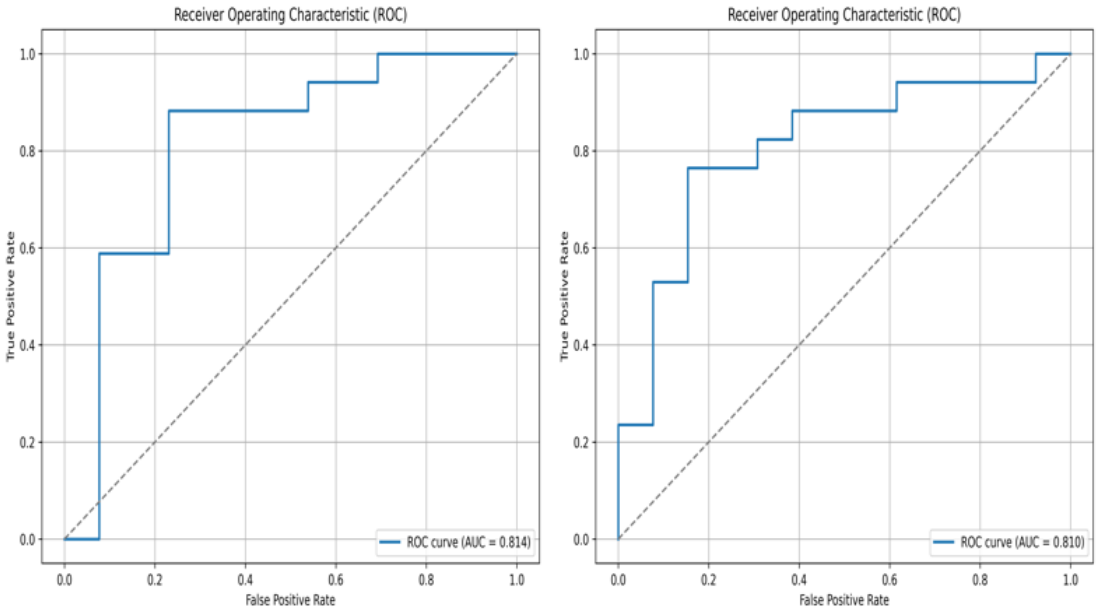

Figure S3.8. ROC curves and (left: all features; right: only informative) for the binary classifier trained with observations.

4. Informative Features *k*=18

Feature	Mutual Information
X1	0.3287
X2	0.3229
X3	0.3268
X4	0.3296
X5	0.3215
X6	0.3247
X7	0.3359
X8	0.3252
X9	0.3266
X10	0.3221
X11	0.3273
X12	0.3294
X13	0.3282
X14	0.3284
X15	0.3338
X16	0.3244
X17	0.3256
X18	0.3262
X19	-0.0036
X20	0.0020

Figure S4.1. Mutual information for all features. Mutual information (MI) estimates between each of the continuous predictors and the binary label for the regime with oracle-informative predictors. The informative coordinates are constructed to achieve an empirical point-biserial correlation with the label of approximately 0.3, while remaining coordinates are oracle-irrelevant noise features.

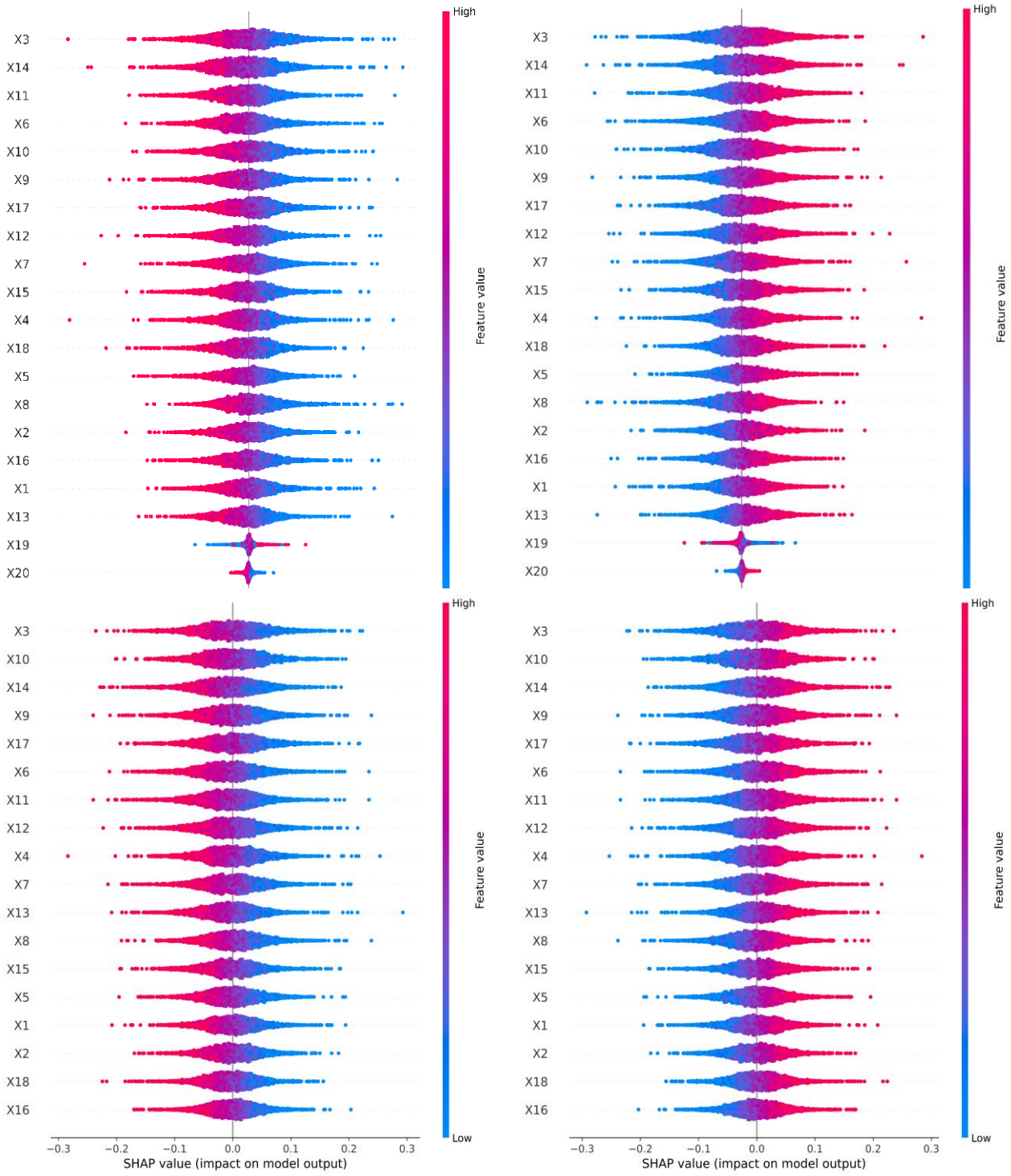

Figure S4.2. SHAP summary plots for informative features (all features vs only informative). SHAP summary plots shown separately for class 0 (left) and class 1 (right). Top row: model trained on all features. Bottom row: model trained only on the oracle-informative subset.

4.1 Full Dataset

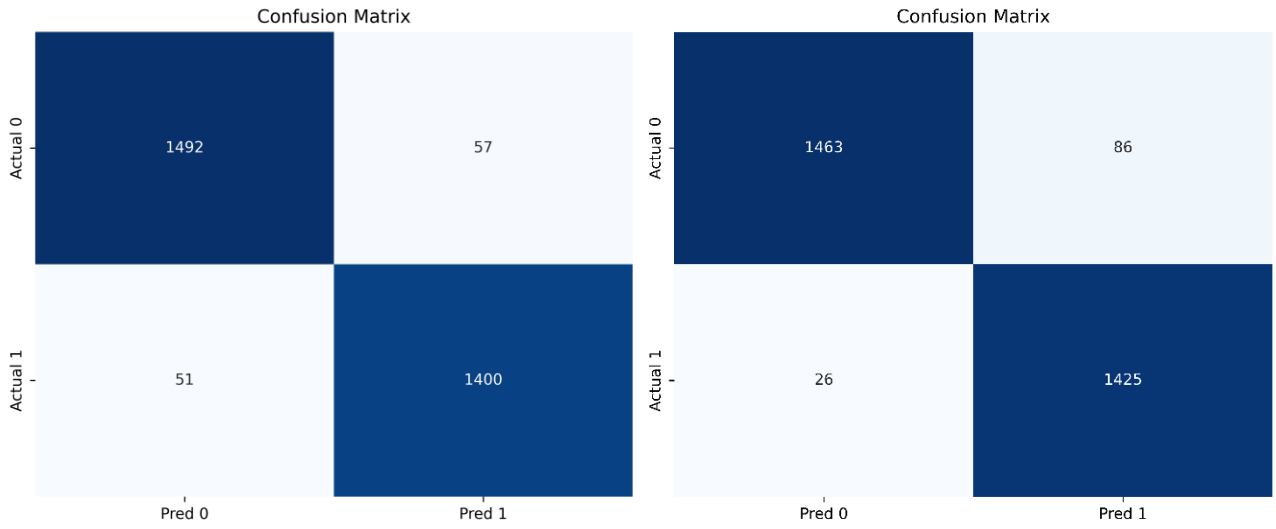

Figure S4.3. Confusion matrices (left: all features; right: only informative) for the binary classifier trained with observations.

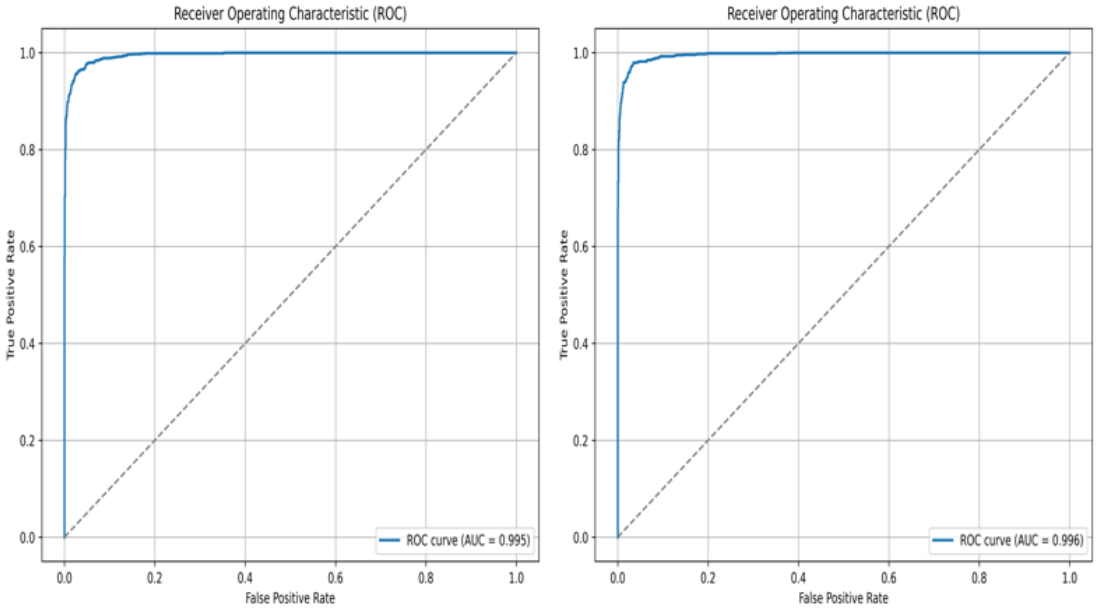

Figure S4.4. ROC curves and (left: all features; right: only informative) for the binary classifier trained with observations.

4.2 Reduced Data: 10% of Dataset

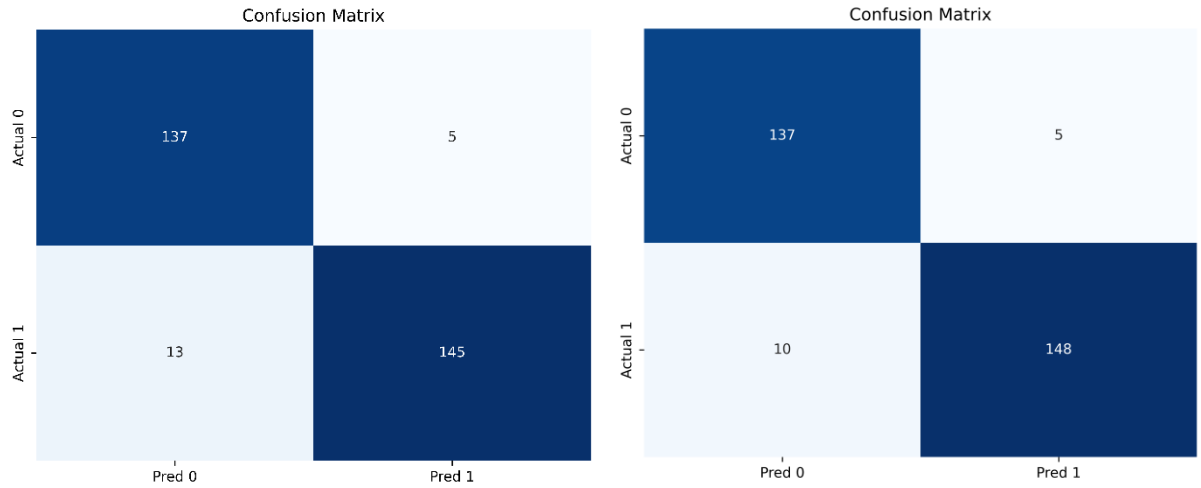

Figure S4.5. Confusion matrices (left: all features; right: only informative) for the binary classifier trained with observations.

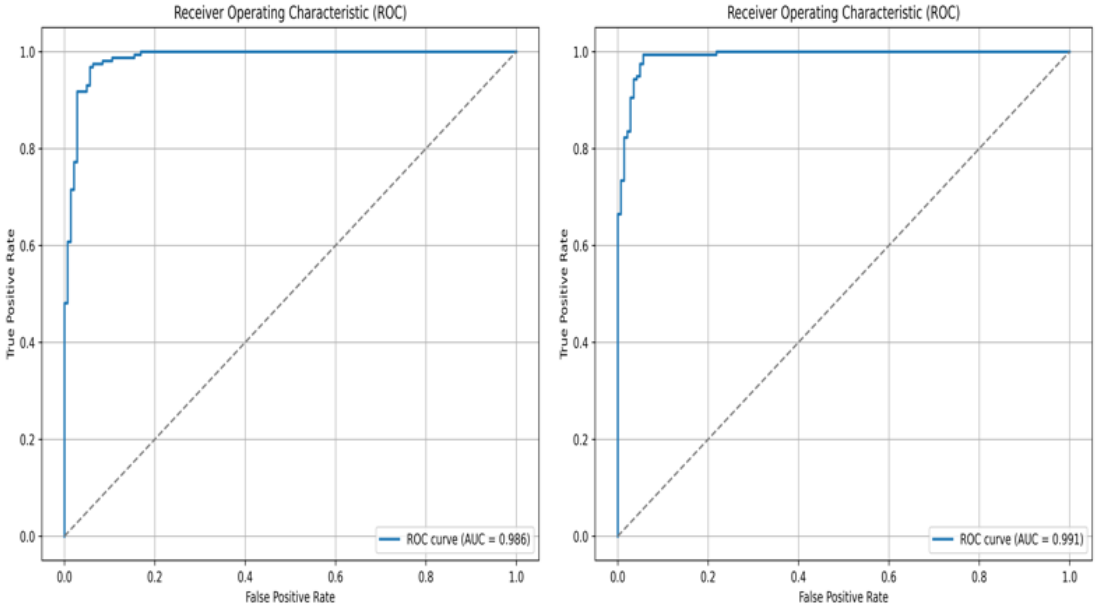

Figure S4.6. ROC curves and (left: all features; right: only informative) for the binary classifier trained with observations.

4.3 Sparse Data: 1% of Dataset

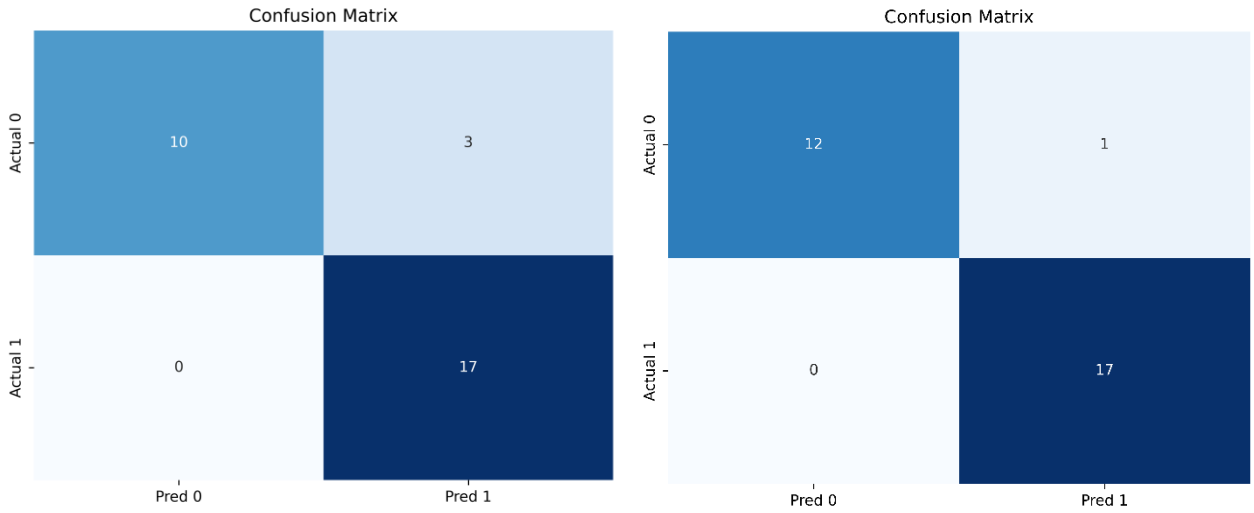

Figure S4.7. Confusion matrices (left: all features; right: only informative) for the binary classifier trained with observations.

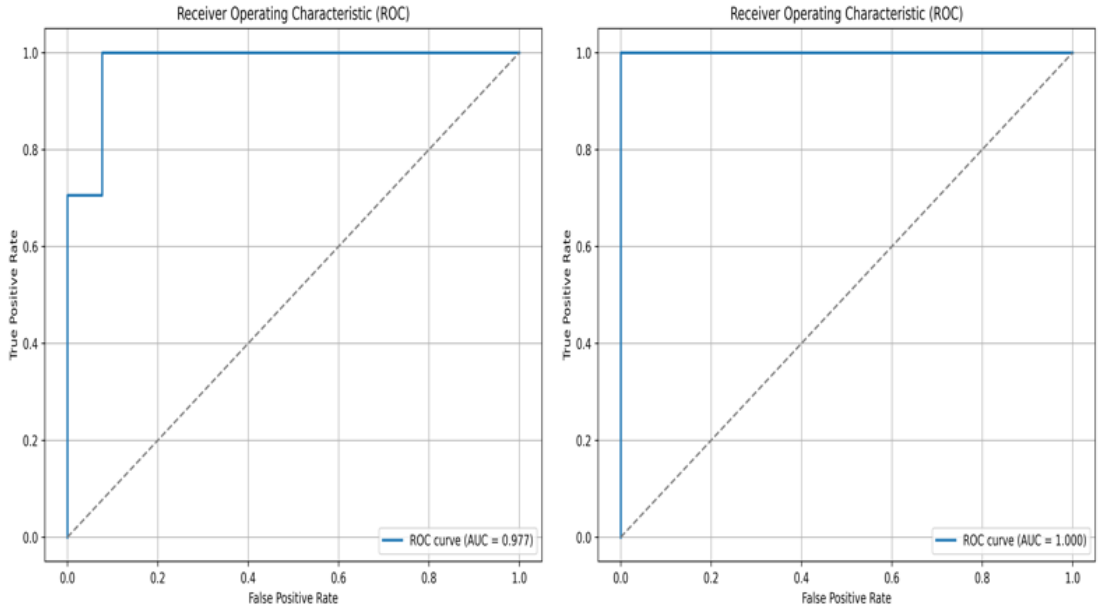

Figure S4.8. ROC curves and (left: all features; right: only informative) for the binary classifier trained with observations.

5. Conclusions

Across the full synthetic grid of configurations, the figures show that removing oracle-irrelevant predictors (training on only informative features) is most beneficial when data are scarce. In the smallest-sample regime, ROC curves and confusion matrices show the clearest operating-point differences between training on all features versus restricting to informative inputs, consistent with reduced variance and reduced susceptibility to fitting noise when redundant dimensions are removed. As sample size increases, the visual gap between the two training settings typically narrows, indicating that with sufficient data the classifier can often attenuate uninformative inputs through training and regularisation. The accompanying SHAP panels support this behaviour by showing that, after pruning, attribution is more concentrated on the intended informative coordinates. Overall, these synthetic results support the paper's main conclusion: feature pruning is effective and its practical value is strongest in limited-data settings where estimation error and instability are most pronounced.