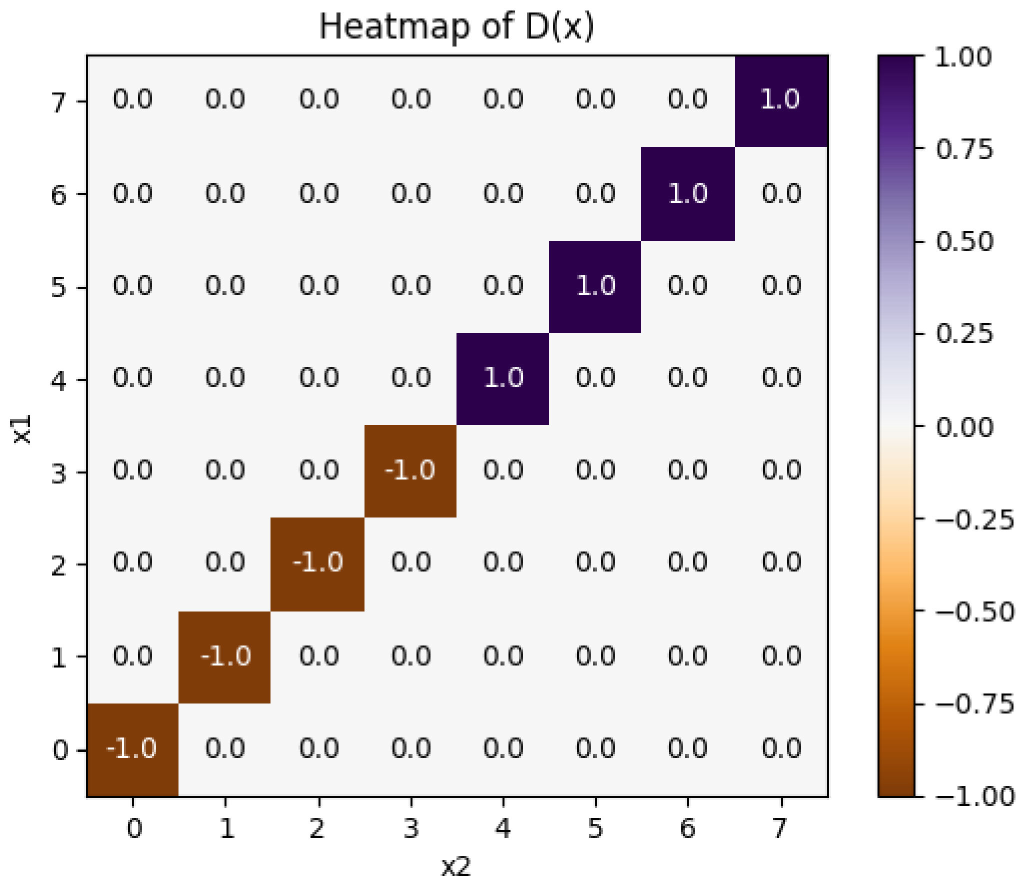
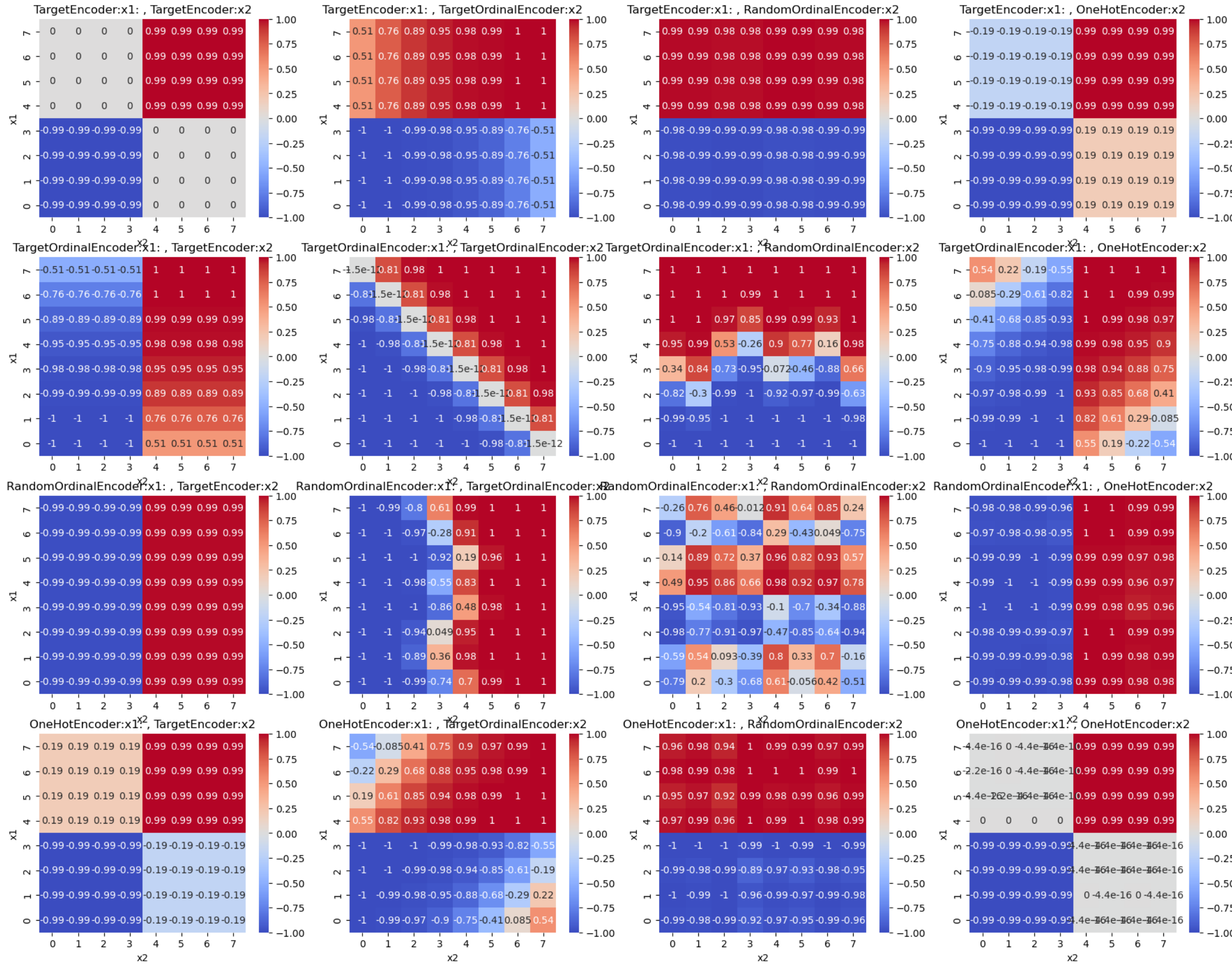
Symmetric Case

Dataset



Shows the plot of
D(x) = (P(y=1|x) - P(y=0|x)) p(x) / p_{max} = (2 P(y=1|x) -1) p(x) / p_{max}. This dataset has identical x1 and x2 features such that the $x_1 = x_2 = y$. A fair classifier should not chose either feature independently in predicting y.
Each feature has 8 possible values, numbered 0 to 7. The training data distribution is concentrated on diagonal locations in input space: $y = 0 : \mathbf{x} \in \{(0,0),(1,1),(2,2),(3,3)\}$; $y = 1 : \mathbf{x} \in \{(4,4),(5,5),(6,6),(7,7)\}$. Hence, generalization outside of such input space locations are guided by encoding bias. The data generated for the following test has 1000 samples.

Logistic Regression, Default Parameters



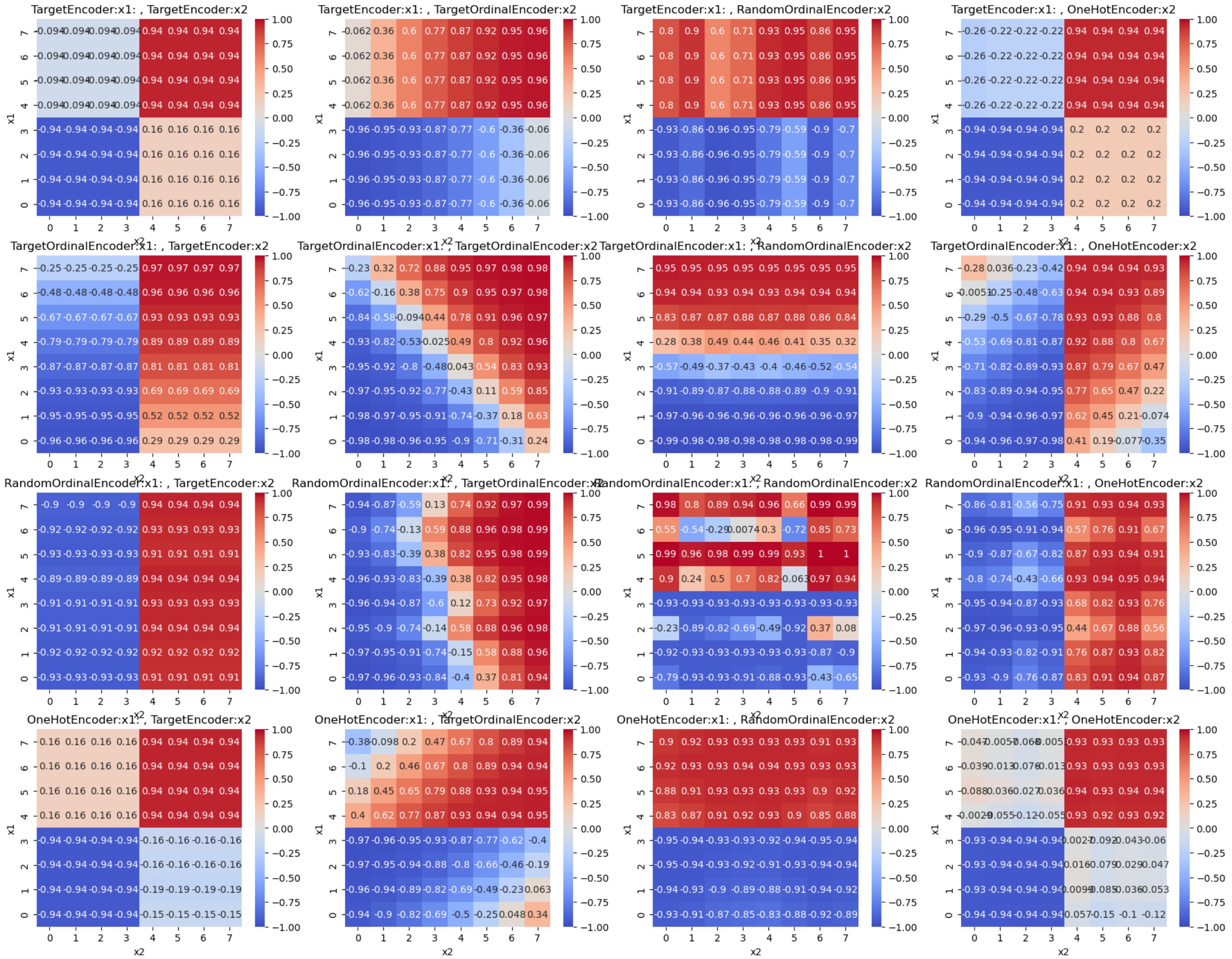
Classifier predictions $2P(y=1|\mathbf{x})-1$ are presented for *test data*. Each heatmap corresponds to a different combination of encoding of the two features x_1 and x_2. Heatmap (j) (one-hot encoding for both variables) is least biased, since the classifier avoids generalizing out-of-domain altogether: $P(y=1|\mathbf{x}) \simeq 0.5$, if $x_1 \neq x_2$. Heatmaps (a) (target encoding for both variables) presents an intuitive out-of-domain generalization (possibly acceptable, depending on the application domain): $P(y=1|\mathbf{x})=0$, if $x_1 \leq 3 \& x_2 \leq 3$, $P(y=1|\mathbf{x})=1$, if $x_1 > 3 \& x_2 > 3$, $P(y=1|\mathbf{x})=0.5$ otherwise. Case (h) results in arbitrary generalization, governed by a random ordering of variable values. The off-diagonal cases are revealing of the asymmetry introduced by encoding, which facilitates generalizing along one of the dimensions. In particular, the heatmaps of the last column (d), (j), and (i), clearly show that one-hot encoding facilitates generalization. Hence, if one variable uses one-hot encoding (here x_2) and the other uses another encoding, then generalization is driven by the one-hot encoded variable: $P(y=1|\mathbf{x}) < 0.5$, if $x_2 \leq 3$, $P(y=1|\mathbf{x}) > 0.5$, if $x_2 > 3$. While one-hot encoding "dominates" all other encodings, it does not strongly dominate target encoding. Furthermore target encoding dominates target-ordinal and random-ordinal and target-ordinal dominates random-ordinal. So, while encoding with on-hot encoding may seem preferable (because least biased), target encoding presents a good alternative, if one-hot encoding is not viable because of problems of curse-of-dimensionality.

$R(e_1, e_2)$:

	TargetEncoder:x2	TargetOrdinalEncoder:x2	RandomOrdinalE
TargetEncoder:x1	-0.032181	-1.76737	-5.07958
TargetOrdinalEncoder:x1	1.74079	-0.00151107	-1.64611
RandomOrdinalEncoder:x1	6.02127	2.066	-0.561533
OneHotEncoder:x1	-0.427298	-0.907887	-4.04687

The quantitative results of $R(e_1, e_2)$ show align visually with which feature is "dominating" the other. We can see that when RandomOrdinalEncoding is used the other feature always dominates reflected by the negative values in the column for x_2 and the positive values in the row for x_1. We can also see that OneHot dominates all other encoding types reflected by its positive column for x_2 and negative row for x_1. We also notice that the diagonal is close to zero where the encoding type is not RandomOrdinalEncoding. A value of zero indicates no preference of variables. The non-zero value for $e_1, e_2 =$ RandomOrdinalEncoding results from the models inability to fit the training data resulting in a randomized preference.

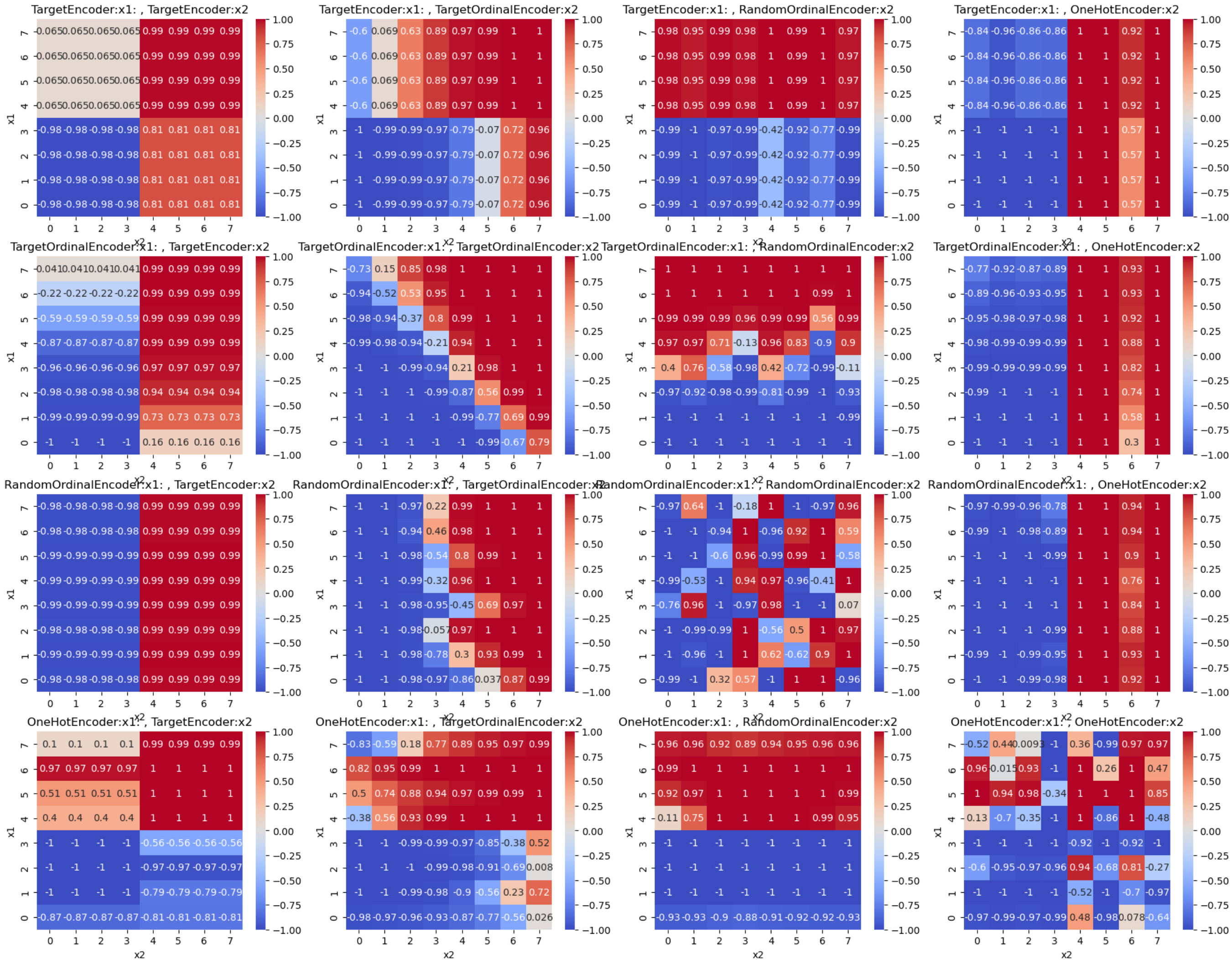
Neural Network, $\alpha = 10$



$R(e_1, e_2)$:

	TargetEncoder:x2	TargetOrdinalEncoder:x2	RandomOrdinalE
TargetEncoder:x1	0.24091	-0.779869	-3.80423
TargetOrdinalEncoder:x1	1.08786	0.195335	-0.484428
RandomOrdinalEncoder:x1	4.92722	2.30064	0.193524
OneHotEncoder:x1	-0.387975	-0.792507	-3.33748

Neural Network, $\alpha = 0$



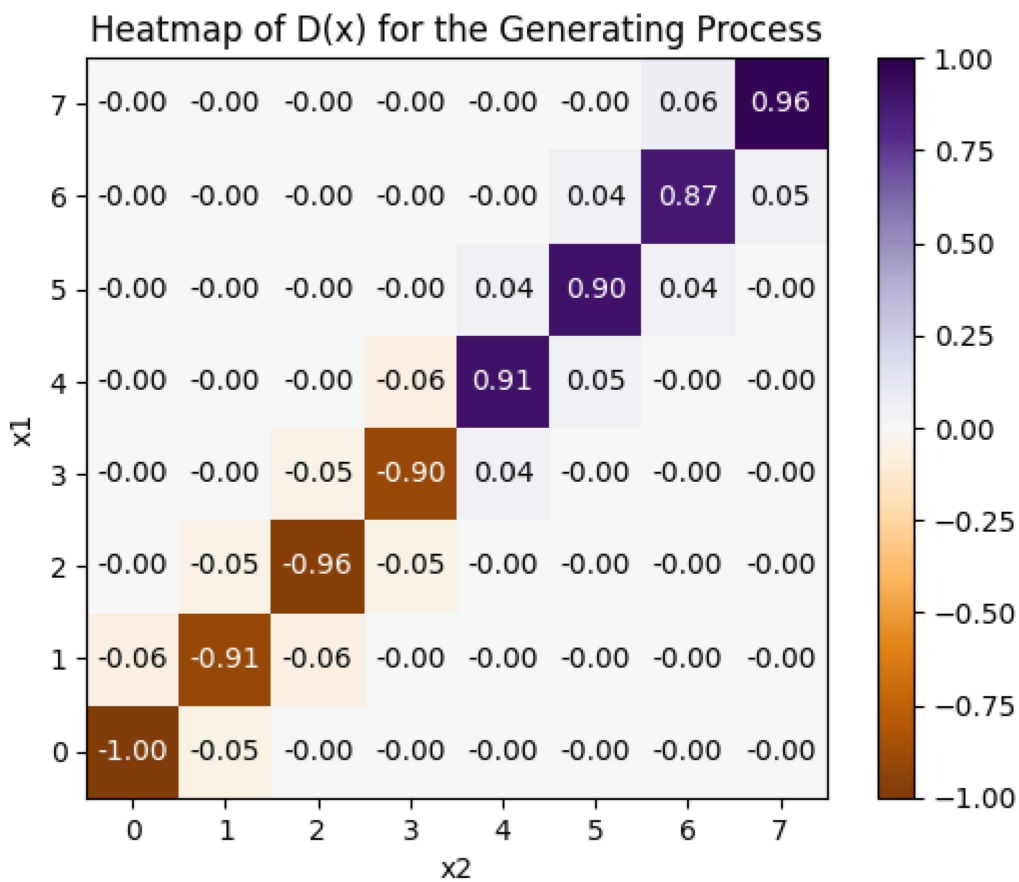
$R(e_1, e_2)$:

	TargetEncoder:x2	TargetOrdinalEncoder:x2	RandomOrdinalE
TargetEncoder:x1	0.764156	0.247328	-4.57222
TargetOrdinalEncoder:x1	0.922938	0.370938	-0.641183
RandomOrdinalEncoder:x1	6.08493	1.04696	-0.844642
OneHotEncoder:x1	-1.67252	-0.808272	-5.67367

\

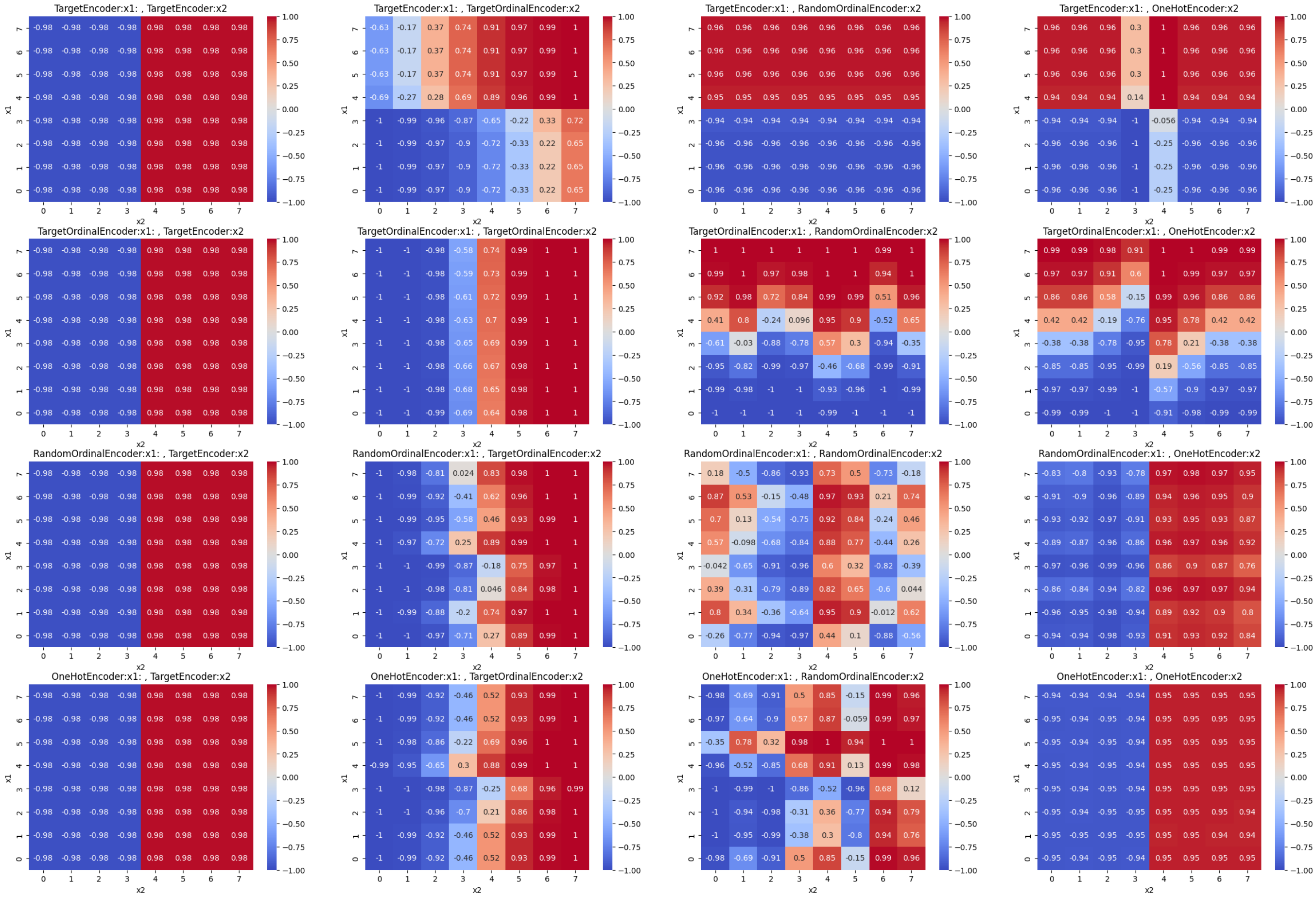
Asymmetric

Dataset



10000 Samples

Logistic Regression, High L1



Real Data

Adult Income Dataset

The Adult dataset, which contains features such as race and education, is used to study the impact of encoding choices in a real-world scenario. Features are preprocessed and encoded using Target Encoding, Target Ordinal Encoding, Ordinal Encoding, and One-Hot Encoding. The model is trained on the preprocessed data, and the performance of different encoding strategies is evaluated using mean squared differences and ratios, similar to the synthetic example.

Neural Network

Hidden Layers: [5,5], Alpha=1

$R(e_1, e_2)$:

	TargetEncoder:education	TargetOrdinalEncoder:education	Or
TargetEncoder:race	1.80761	0.378853	0.2
TargetOrdinalEncoder:race	1.94494	0.606847	0.5

	TargetEncoder:education	TargetOrdinalEncoder:education	Or
OrdinalEncoder:race	1.74858	0.575731	0.4
OneHotEncoder:race	1.64214	0.373556	0.3

Balanced Accuracy:

	TargetEncoder:education	TargetOrdinalEncoder:education	Or
TargetEncoder:race	0.584902	0.5	0.5
TargetOrdinalEncoder:race	0.584754	0.5	0.5
OrdinalEncoder:race	0.580292	0.5	0.5
OneHotEncoder:race	0.584902	0.5	0.5

Cook County

Logistic Regression, Default

$R(e_1, e_2)$:

	TargetEncoder:COMMITMENT_TYPE	TargetOrdinalEncoder:C
TargetEncoder:RACE	0.996685	0.996612
TargetOrdinalEncoder:RACE	0.996685	0.996612
OrdinalEncoder:RACE	0.996685	0.996612
OneHotEncoder:RACE	0.996685	0.996612

Balanced Accuracy:

	TargetEncoder:COMMITMENT_TYPE	TargetOrdinalEncoder:C
TargetEncoder:RACE	3.617903	2.144558
TargetOrdinalEncoder:RACE	4.496680	3.150707
OrdinalEncoder:RACE	3.818604	2.525171
OneHotEncoder:RACE	3.055226	1.509424

Neural Network

Hidden Layers: [5,5], Alpha=1

$R(e_1, e_2)$:

	TargetEncoder:COMMITMENT_TYPE	TargetOrdinalEncoder:C
TargetEncoder:RACE	0.996685	0.996612
TargetOrdinalEncoder:RACE	0.996685	0.996731
OrdinalEncoder:RACE	0.996685	0.996764
OneHotEncoder:RACE	0.996685	0.996688

Balanced Accuracy:

	TargetEncoder:COMMITMENT_TYPE	TargetOrdinalEncoder:C
TargetEncoder:RACE	3.985657	2.267989
TargetOrdinalEncoder:RACE	4.647600	3.303791
OrdinalEncoder:RACE	3.735236	2.547301
OneHotEncoder:RACE	3.283287	2.449550