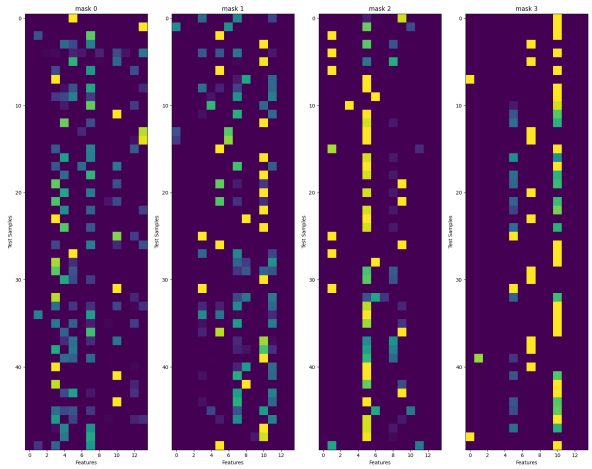
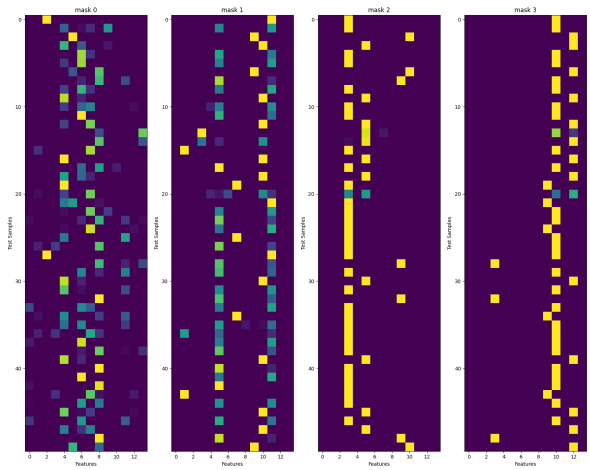
1. Ablation Study on TabNet's Sparsity Regularizer

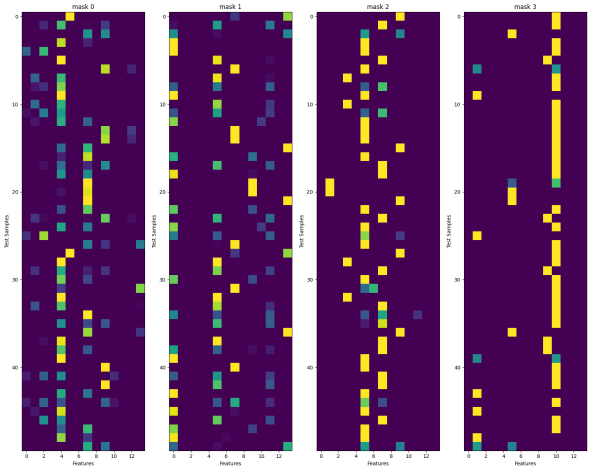
ADULT CENSUS INCOME MASK FIGURES



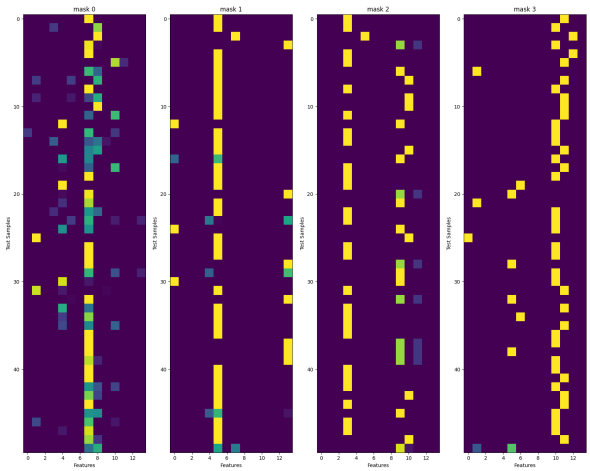
(a) TabNet ($\lambda_{sparse=0.0001}$, Test Accuracy: 85.46%)



(b) TabNet ($\lambda_{sparse=0.001}$, Test Accuracy: 86.07%)



(c) TabNet ($\lambda_{sparse=0.01}$, Test Accuracy: 86.04%)



(d) TabNet ($\lambda_{sparse=0.1}$, Test Accuracy: 83.41%)

Figure 1: As λ_{sparse} increases, TabNet's improvement in interpretability is negligible, coupled with a decrease in accuracy.

TabNet (Arik and Pfister, 2020) controls the sparsity of the selected features by employing a sparsity regularizer in the form of entropy (Grandvalet and Bengio, 2004), where $M[i]$ is a learnable mask. Quoted from TabNet,
$L_{sparse} = \sum_{i=1}^{N_{steps}} \sum_{b=1}^{B} \sum_{j=1}^{D} \frac{-M_{b,j}[i] \log(M_{b,j}[i]+\epsilon)}{N_{steps} \cdot B}$, where ϵ is a small number for numerical stability.

Per TabNet's recommendations, we tuned and produced feature masks in the range of $L_{sparse} = \{0.0001, 0.001, 0.01, 0.1\}$. As depicted in Figure 1, even though we increased TabNet's sparsity regularizer L_{sparse}, the improvement in interpretability is negligible as almost all the features are reasoned from at each decision step. At the highest L_{sparse} level of 0.1, interpretability might have improved slightly as we start to observe some salient features. However, there is a 2.63% decrease in test accuracy when compared to L_{sparse} of 0.01 (83.41% vs. 86.04%). This proves that naively adding regularization does not work in improving interpretability while maintaining a competitive accuracy.

On the other hand, InterpreTabNet achieves a higher test accuracy (87.42%) while attaining highly interpretable feature masks (refer to Figure 2 and Table 1 of our main text for InterpreTabNet's feature mask and accuracy).

References

S. O. Arik and T. Pfister. Tabnet: Attentive interpretable tabular learning, Feb 2020. URL https://arxiv.org/abs/1908.07442v4.

Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/96f2b50b5d3613adf9c27049b2a888c7-Paper.pdf.